LA CALIDAD TOTAL EN BASES DE DATOS ESPAÑOLAS: ESTUDIO DE LA TASA DE ERROR EN LAS BASES DEL CSIC

Víctor Herrero Solana*

Resumen. La calidad de productos y servicios es uno de los temas en boga de la presente década. Este trabajo tiene como objetivo la revisión del nivel de calidad de tres bases de datos españolas creadas y mantenidas por el Consejo Superior de Investigaciones Científicas: ICYT, ISOC e IME. Se utiliza como variable de análisis la tasa de error, uno de los indicadores más comúnmente utilizados para este tipo de estudios. Finalmente se comparan los resultados obtenidos con los estándares aceptados a nivel internacional.
Palabras clave: calidad total, tasa de error, bases de datos.

Abstract. The quality of products and services is the buzzword of the 90's, specially in the information retrieval field. This article analyzes de quality level of three Spanish databases developed by the Council for Scientific Research: ICYT, ISOC and IME. The variable studied is the error rate, a objective indicator for the quality measurement. Finally, the results are compared to international standards.
Keywords: Total quality management, TQM, error rate, database.

1 Introducción

La literatura en torno a la gestión de calidad de productos y servicios ha experimentado un sustancial aumento en los últimos años. Un estudio realizado sobre las bases de datos LISA e ISA indica que durante la década del 80 y principios del 90, nos encontramos con un promedio de 200 trabajos anuales que tratan dicha temática (1). En ese mismo estudio se puede apreciar un aumento abrupto, desde 1989, de la literatura que aborda el concepto de gestión de la calidad total o *total quality management* (TQM). Un tratamiento dirigido hacia el campo documental podrá encontrase en los trabajos de Daniel (2) y Brockman (1, 3).

Los principios y filosofía de la TQM se han comenzado a aplicar sistemáticamente a los diferentes productos y servicios pertenecientes al campo documental (4). Las bases de datos internacionales, por ejemplo, son actualmente evaluadas y controladas en función de una serie de indicadores que permiten medir su grado de calidad (5). Uno de los criterios más conocidos y ampliamente utilizado es el elaborado por el SCOUG (antiguo *Southern California Online Users Group,* actualmente denominado *Seven Continents Online Users Group*), a principios de los años 90 (6).

Dentro de cualquier proceso de transferencia y recuperación de información, la integridad de ésta es uno de sus factores críticos (7), y abarca cuatro aspectos principales. Toda información de calidad debe ser: *a*) exacta, *b*) completa, *c*) consistente, y *d*) actualizada. De estos cuatro aspectos nos centraremos en la exactitud, o lo que es lo mismo, en su complemento: el error. Desde un punto de vista gramatical, se pueden establecer

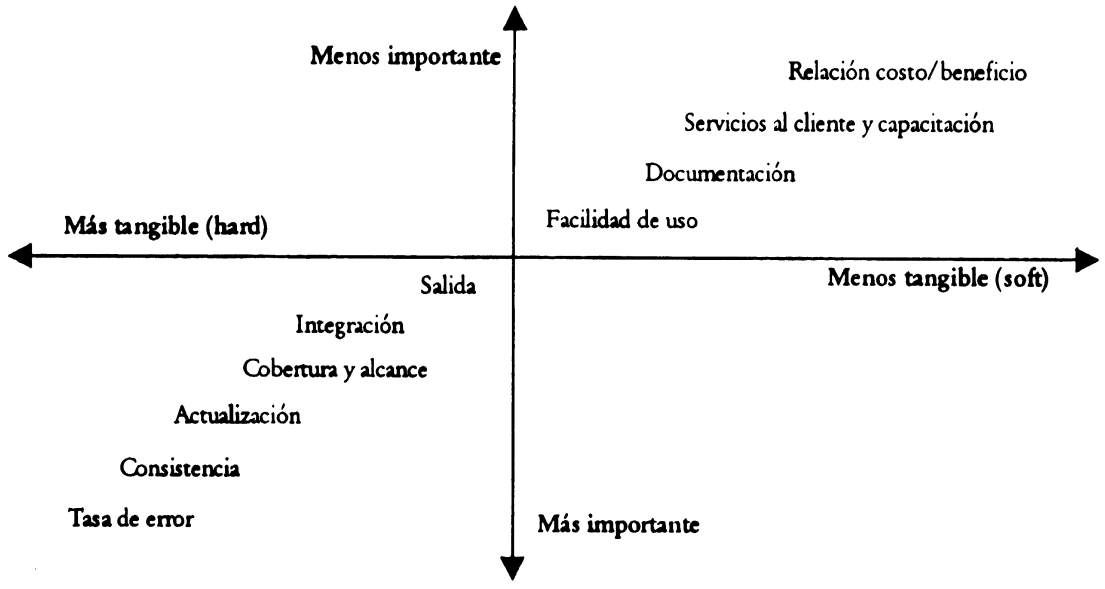
* Universidad de Nacional de Mar del Plata (Argentina) y Universidad de Granada (España).
Recibido: 10-8-97.

tres niveles de errores (8): *a*) ortográfico (palabras mal escritas), *b*) sintáctico (frases mal construidas, mala conjugación, número o género, etc.), y *c*) semántico (frases sintáctica y ortográficamente correctas pero carentes de todo sentido). Cuando se mide la calidad de una base de datos, solamente se consideran los errores a nivel ortográfico, desestimándose los dos restantes debido a su baja incidencia en la recuperación de la información. Los errores ortográficos se originan durante la entrada de datos a la base y pueden tener diversos orígenes: errores en la digitación del teclado, errores en el reconocimiento óptico de caracteres, error en la transferencia electrónica de datos, etc. Estos errores pueden ser clasificados en cuatro grandes grupos (9): 1) *inserción:* un carácter es insertado dentro de la palabra; 2) *omisión:* un carácter es omitido de la palabra; 3) *transposición:* dos caracteres adyacentes de una palabra son intercambiados; y 4) *sustitución:* un carácter de la palabra es reemplazado por otro diferente.

Al comienzo hicimos mención de unos indicadores propuestos por SCOUG para medir la calidad de una base de datos, entre los que se encontraba la tasa de error. Ahora bien, podríamos preguntarnos qué relación existe entre cada uno de los indicadores y cuáles son los más importantes a la hora de medir la calidad de un producto informativo. Lo cierto es que esta pregunta no tiene una respuesta concreta, ya que los indicadores no son comparables entre sí, incluso muchos de ellos no son cuantificables. En la figura 1 encontramos un gráfico cartesiano donde cada indicador es ubicado en función de su importancia y su facilidad de cuantificación (tangibilidad) (10). Podemos apreciar a simple vista que la tasa de error es considerada el indicador más tangible e importante de todos, lo que justifica su análisis.

Si bien existe unanimidad de criterios sobre la importancia de la tasa de error como el indicador más concreto para medir la calidad de una base de datos, no se ha definido cuáles son los valores máximos tolerables para esta variable. Revisando la bibliografía sobre el tema encontramos diferentes criterios. En uno de los primeros trabajos

Figura 1
Matriz de relevancia para los criterios de calidad de SCOUG



sobre el tema (11), el autor realizó un estudio sobre el porcentaje de error en 11 conocidas bases de datos disponibles en DIALOG, en el que halló tasas que oscilan entre el 0,01% (BIOSIS) y el 0,63% (Abstracted Business Information). En otro estudio realizado sobre bases de datos latinoamericanas (CEPAL, LILACS, REDUC, REPIDISCA, entre otras), Ernesto Spinak, de la Universidad de la República de Uruguay, afirma que la tasa de error promedio es del 0,52% (8). Este valor se calculó de forma automática teniendo como referencia el total de ocurrencias de cada término de las bases, y nos indica que en promedio encontraremos un error cada 200 palabras. Algunas bases latinoamericanas presentan mejores tasas de error que las detectadas por Spinak, tal es el caso del catálogo de la Biblioteca Daniel Cosío Villegas de El Colegio de México. Quienes lo llevan adelante prestan mucha atención al control de calidad y han detectado tasas de error, que afectan a la recuperación, cercanas al 0,3% (12). No obstante, el gran problema con que nos encontramos a la hora de establecer los estándares latinoamericanos es la falta de información sobre experiencias como las de El Colegio de México.

En otros estudios, J. J. Pollock y A. Zamora de Chemical Abstracts Service, afirman que el valor máximo tolerable no debe exceder el 0,2% (un error cada 500 palabras) (9). No obstante, cuando hablamos de calidad total de información, los niveles de exigencia son más elevados. Earl Beutler de Research Information Systems, establece que un producto informativo de alta calidad *(top quality)* no puede presentar una tasa de error superior al 0,02%, o lo que es lo mismo, un error de promedio por cada 5.000 palabras (7). Algunos autores van un poco más allá, e incluso discriminan el porcentaje de errores por cada uno de los cinco tipos establecidos (13).

2 Objeto de estudio

El Consejo Superior de Investigaciones Científicas (CSIC) es el organismo de investigación más grande de España. Entre sus diversas tareas se encuentra la de llevar adelante una serie de bases de datos: ICYT (información sobre ciencia y tecnología), ISOC (información sobre ciencias sociales y humanas), IME (información médica), ALAT (base de datos multidisciplinar sobre América Latina), CIRBIC (catálogo colectivo de libros y revistas existentes en las bibliotecas del CSIC), DATRI (transferencia de investigación), entre otras. Las tres primeras son quizás las más importantes debido a que recogen toda la información de cada especialidad publicada en España y son representativas de la ciencia española en su conjunto.

ICYT es una base de datos referencial y bibliográfica que recoge la literatura científica contenida en publicaciones españolas de ciencia y tecnología, cubriendo más de 500 revistas científicas nacionales, así como también diversos informes, actas de congresos, etc. Su actualización es mensual, y cubre desde 1979 hasta nuestros días. Actualmente la base cuenta con más de 100.000 registros.

ISOC está compuesta por un conjunto de bases de datos bibliográficas integradas en las que se recogen referencias bibliográficas especializadas en las distintas áreas de las ciencias humanas y sociales (Economía, Sociología, Política, Historia, Arqueología y Prehistoria, Bellas Artes, Documentación Científica, Derecho, Lingüística y Literatura, Psicología y Educación, Geografía y Urbanismo, entre otras). Las bases recogen los artículos publicados en más de 1.625 revistas científicas españolas y otros documentos,

como informes técnicos, comunicaciones a congresos, monografías, etc. En la actualidad reúnen más de 260.000 referencias bibliográficas que cubren desde 1975 hasta la fecha.

IME es una base de datos especializada en Biomedicina y disciplinas asociadas (Administración Sanitaria, Farmacia Clínica, Medicina Experimental, Microbiología, Psiquiatría, Salud Pública). Se recogen los trabajos publicados en 115 revistas españolas, desde 1971 hasta la fecha. Cuenta con un volumen de 174.763 registros, su actualización es mensual y su crecimiento anual ronda las 7.600 referencias.

Estas tres bases de datos serán nuestro objeto de estudio, ya que si bien han sido analizadas en diversos aspectos (14), no se ha publicado ninguna investigación en torno a su tasa de error. Para más información sobre las mismas y sobre el CSIC, consultar la página del servidor de información del organismo (http://www.csic.es/).

3 Metodología

Para el desarrollo del presente estudio se ha adoptado la metodología descrita por Charles Bourne (11), debido a su fácil implementación y a su amplia aceptación por varios autores posteriores (9, 15, 16). Esta consiste en definir, dentro del índice de la base, tres distancias o intervalos arbitrarios de términos: APPLE-AQUA, GRAPE-GREECE, y PLUM-PLUTO. Nosotros hemos adecuado las distancias propuestas por Bourne, definiendo las siguientes: AGUA-AJO, GRANO-GRIFO, y PLOMO-POLI. Se ha tenido especial cuidado en tomar intervalos lo más balanceados posible, con el fin de evitar los problemas con que se encontró Bourne al definir el tercer intervalo y luego comprobar que comprendía pocos términos.

Luego de esto se obtuvo una serie de listas detalladas con todos los términos del índice comprendidos en cada intervalo, junto con las ocurrencias *(postings)* asociadas a cada uno. La información del índice corresponde a los siguientes campos: título, título en otro idioma, título de revista, descriptores, resumen, topónimos, y lugar de trabajo.

El campo de autor no será utilizado ya que es muy difícil encontrar un error en un nombre propio de persona, a menos que se cuente con la fuente primaria. En estas listas se individualizan los términos erróneos: errores tipográficos (p .ej. *aguna* en lugar de *laguna*), palabras juntas (p. ej. *aguañcombustible*), o mal separadas (p. ej. *leng u aje*). Lo importante para destacar es que las listas de términos fueron analizadas en línea (y no impresas como en el caso de Bourne): esto ha permitido, en caso de duda, el acceso inmediato al registro completo y su análisis en el contexto. También es importante recalcar que, a diferencia del estudio de Bourne, aquí se tuvieron en cuenta las diferentes lenguas que aparecen en la base (castellano, catalán, francés e inglés, mayoritariamente), y cuyo análisis efectivo sólo pudo realizarse en función del contexto de cada registro (en línea). Finalmente, con esta información se construye una tabla de doble entrada donde se presentan los valores absolutos obtenidos, junto con sus respectivos cálculos porcentuales.

4 Resultados y discusión

Como hemos visto en el apartado anterior, la información fue extraída término a término de las tres bases de datos, para luego volcarla en una hoja de cálculo y elaborar la tabla I. De cada intervalo se indica: 1) el total de términos tenidos en cuenta

Tabla I
Tasas de error de cada base discriminadas por intervalo

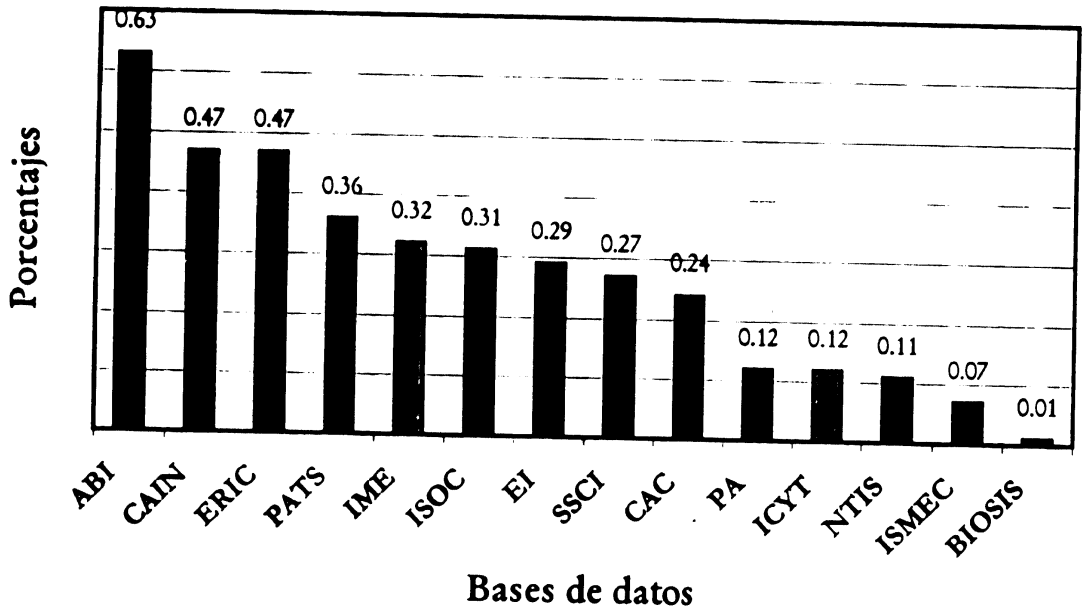
	ICYT	*ISOC*	*IME*
Agua-ajo			
Núm. de términos	259	342	145
Núm. de términos erróneos	9 (3,47%)	28 (8,19%)	23 (15,86%)
Núm. de ocurrencias	11.528	8.526	10.420
Núm. de ocurrencias erróneas	15 (0,13%)	42 (0,49%)	28 (0,27%)
Grano-grifo			
Núm. de términos	236	266	181
Núm. de términos erróneos	9 (3,81%)	21 (7,89%)	9 (4,97%)
Núm. de ocurrencias	3.876	5.141	3.188
Núm. de ocurrencias erróneas	9 (0,23%)	27 (0,53%)	9 (0,28%)
Plomo-poli			
Núm. de términos	286	485	205
Núm. de términos erróneos	8 (2,80%)	55 (11,34%)	17 (8,29%)
Núm. de ocurrencias	4.560	31.255	3.362
Núm. de ocurrencias erróneas	8 (0,18)	72 (0,23%)	17 (0,51)
Total			
Núm. de términos	781	1.093	531
Núm. de términos erróneos	26 (3,33%)	104 (9,52%)	49 (9,23%)
Núm. de ocurrencias	19.964	44.922	16.970
Núm. de ocurrencias erróneas	32 (0,16%)	141 (0,31%)	54 (0,32%)

(luego de eliminar nombres propios, etc.), 2) número y porcentaje de términos erróneos, 3) ocurrencias totales *(postings)*, y 4) número y porcentaje que representan las ocurrencias erróneas. Es muy importante este último dato, ya que nos indicará la tasa de error de la base.

De los valores obtenidos podemos claramente observar cómo la base ICYT arroja una tasa de error menor a la de las otras bases. Además, presenta una cierta homogeneidad en los tres intervalos, promediando una tasa del 0,16%. En cuanto a ISOC e IME, estas presentan un comportamiento homogéneo en los dos primeros intervalos, con una tasa mayor para la primera, pero ambas por encima de ICYT. En el tercer intervalo los papeles se cambian y cada base presenta los resultados a la inversa. Esto permite que ambas obtengan un promedio muy similar cercano al 0,3%. Sin embargo, si observamos la cantidad de errores, el tercer intervalo de ISOC arroja 55 (la mayor cantidad del estudio), pero logra bajar su tasa gracias a la enorme cantidad de ocurrencias (más de 30.000). Esta gran cantidad de ocurrencias se debe a unos pocos términos (variaciones de las palabras *poesía* y *poeta*), que llegaron a presentar más de 7.000 ocurrencias cada uno. Por otro lado, la subida en los valores de IME se debe a la baja cantidad de ocurrencias (el 10% de las del ISOC). No obstante, estas dispersiones son relativamente bajas si las comparamos con algunas del trabajo de Bourne, donde por ejemplo la base ERIC arroja los siguientes resultados: 0,09, 0,8, y 2.56%.

En la figura 2 podemos comparar los resultados obtenidos por Bourne con nuestros propios resultados. No se pretende hacer una comparación exhaustiva entre cada base de datos, entre otras razones porque los valores de Bourne son de hace 20 años, tiempo más

Figura 2
Tasa de error en base al porcentaje de ocurrencias con entradas erróneas



ABI (Abstracted Bussines Information); CAIN (Cataloging and Indexing from National Agricultural Library); ERIC (Research in Education and Current Index to Journals in Education files); PATS (Chemical Market Abstracts and Equipment Market Abstracts from Predicasts); IME (Base de datos en Medicina - CSIC); ISOC (Bases de datos en Ciencias Sociales y Humanas - CSIC); EI (Engineering Index); SSCI (Social Science Citation Index); PA (Psychological Abstracts); ICYT (Base de datos en Ciencia y Tecnología - CSIC); NTIS (Goverment Report Announcements); ISMEC (Information Service in MeChanical Engineering from INSPEC); BIOSIS (BIOSIS Previews).

que suficiente para que el contenido de dichas bases haya cambiado. Lo importante aquí es poder apreciar cómo en la década del 70, y cuando todavía no se hablaba de la TQM, ya existían bases como BIOSIS con un alto nivel de calidad.

5 Conclusiones

Podemos afirmar que las bases de datos del CSIC presentan tasas de error tolerables. Si comparamos los resultados con los valores de Spinak, nos encontramos que éstas se encuentran por debajo del promedio de las bases latinoamericanas. Si tomamos en cuenta la tasa Pollock-Zamora, obtendríamos dos grupos. Por un lado ICYT, que se encuentra por debajo del 0,2%, y por el otro ISOC e IME, que presentan valores mayores. De todos modos, ninguna de nuestras bases alcanza los niveles de propuestos por Beutler y que establecen el límite a partir del cual podríamos hablar de calidad total. Es más, si observamos nuevamente la figura 2, veremos que solo BIOSIS alcanza dicho nivel.

Por esto es importante que los productores de las bases del CSIC sigan trabajando sobre la calidad de las mismas. Es necesario bajar los niveles de error de ISOC e IME

por lo menos hasta igualarlos a ICYT, para luego llevarlos hacia niveles más bajos. Esta es quizás la primera tarea a acometer, ya que como hemos visto, la tasa de error es el indicador de calidad más inmediato y fácilmente cuantificable. Si no se cumple con este indicador, es en vano que se busque la calidad del producto a través de los otros.

Uno de los principales problemas detectados fue el uso ambiguo de la letra ü. Se encontraron varios términos escritos indistintamente (con y sin diéresis) y con una gran cantidad de ocurrencias para cada versión, lo que afecta claramente el cociente de error. Otro tema importante a ser solucionado es el de las mayúsculas/minúsculas. Mientras que ICYT las respeta, ISOC e IME se encuentran completamente en mayúsculas. Esto dificulta la lectura e induce a muchos errores porque los términos en mayúsculas prescinden totalmente de los acentos.

El presente trabajo ha arrojado algunos resultados relevantes sobre la tasa de error; sin embargo, sería importante que estos fueran complementados con otros estudios que aborden otras metodologías. El método de los trigramas expuesto por Spinak podría aportar información valiosa. El único problema que presenta es que necesita disponer del control del archivo índice para poder realizar el procesamiento automático que precisa este método. Lamentablemente, esto ha estado fuera de nuestro alcance, pero es una de las líneas por las que se puede continuar esta investigación.

6 Bibliografía

1. BROCKMAN, J. Just another management fad? The implications of TQM for library and information services. *Aslib Proceedings* 1992, vol, 44, n.º 7/8, p. 283-288.
2. DANIEL, E. Quality control of documents. *Library Trends* 1993, vol. 41, n.º 4, p. 644-664.
3. BROCKMAN, J. Information management and corporate total quality. *Journal of Information Science* 1993, vol. 19, n.º 4, p. 259-265.
4. MONNET, J . *The quality of electronic information products and services* 1995. Luxembourg: Information Market Observatory (IMO) (Working Paper 95/4).
5. DUFLOS, A. *Les criteres d'evaluation des banques de données: la demarche qualité chez les professionnels de l'information electronique* 1995. París: ADBS (Collection Sciences de l'information; Série Recherches et documents).
6. BASH, R. An overview of quality and value in information services. En: Bash, R., editor. *Electronic information delivery: ensuring quality and value* 1995. Hampshire: Gower, p. 1-10.
7. BEUTLER, E. Assuring data integrity and quality: a database producer's perspective. En: Bash, R., editor. *Electronic information delivery: ensuring quality and value* 1995. Hampshire: Gower, p. 59-68.
8. SPINAK, E. Errores ortográficos en el ingreso en bases de datos. *Revista Española de Documentación Científica* 1995, vol. 18, n.º 3, p. 307-319.
9. POLLOCK, J.; ZAMORA, A. Collection and characterization of spelling errors in scientific and scholarly text. *Journal of the American Society for Information Science* 1983, vol. 34, n.º 1, p. 51-58.
10. ARMSTRONG, C. J. The eye of the beholder. En: Bash R., editor. *Electronic information delivery: ensuring quality and value* 1995. Hampshire: Gower, p. 221-244.
11. BOURNE, C. Frequency and impact of spelling errors in bibliographic data bases. *Information Processing & Management* 1977, vol. 13, n.º 1, p. 1-12.

12. QUIJANO SOLIS, A.; ARRIOLA NAVARRETE, O. Medidas de calidad en la creación de catálogos de biblioteca. *Seminario-taller: «Modelación matemática de la actividad bibliotecaria»* 1997, mayo, 6-9, México DF (Universidad Nacional Autónoma de México, Centro Universitario de Investigaciones Bibliotecológicas).
13. PETERSON, J. A note on undetected typing errors. *Communications of the ACM* 1986, vol. 23, n.º 12, p. 676-687.
14. GIL LEIVA, Y.; RODRIGUEZ MUÑOZ, J. V. Análisis de los descriptores de diferentes áreas de conocimiento indizadas en bases de datos del CSIC. Aplicación a la indización automática. *Revista Española de Documentación Científica* 1997, vol. 20, n.º 2, p. 150-160.
15. O'NEILL, E.; VIZINE-GOETZ, D. Quality control in online databases. *Annual Review of Information Science and Technology (ARIST)* 1988, vol. 23, p. 125-155.
16. CAHN, P. Testing database quality. *Database* 1994, vol. 17, n.º 1, p. 23-30.