VI COLOQUIO SOBRE AUTOMATIZACION DE BIBLIOTECAS

UNIVERSIDAD DE COLIMA

COLIMA

1993

"Estudio de diagnóstico de errores en el catálogo automatizado de la Biblioteca de El Colegio de México"

Pilar María Moreno Jiménez
Biblioteca Daniel Cosío Villegas
El Colegio de México

Universidad de Colima
VI Coloquio sobre Automatización de Bibliotecas

ESTUDIO DE DIAGNOSTICO DE ERRORES EN EL CATALOGO AUTOMATIZADO DE LA BIBLIOTECA DE EL COLEGIO DE MEXICO

Pilar María Moreno
El Colegio de México

RESUMEN

Se presentan los procedimientos por los cuales se realizó un diagnóstico de errores de diversos tipos en el catálogo automatizado de la Biblioteca Daniel Cosío Villegas de El Colegio de México.

El análisis se realizó en: 1) registros completos; 2) índices por campos; 3) problemas específicos. Los errores se clasificaron en cuatro categorías: 1) espacios, puntuación, mayúsculas, errores tipográficos y ortográficos; 2) codificación y etiquetado MARC; 3) forma y uso de las entradas; 4) otros.

Se exponen los resultados obtenidos del análisis y se sugieren las medidas, tanto correctivas como preventivas, encaminadas a minimizar la existencia de errores y, por tanto, a elevar la calidad de la base de datos.

INTRODUCCION

La actual demanda en la calidad de las bases de datos viene dada por dos factores: el creciente uso de las bases de datos, ya sea en línea o en discos compactos, y el nuevo énfasis de la sociedad en general sobre la calidad.

La calidad de las bases de datos se mide, entre otras cosas, por la ausencia de errores de diversos tipos. Muchas de las investigaciones sobre la calidad de las bases de datos bibliográficos están encaminadas a encontrar procedimientos, en lo posible automáticos, para la detección y corrección de dichos errores.

Se puede decir que gran parte del éxito en la automatización de los catálogos depende de la calidad de los datos. Y, por oposición, nada puede mermar más la credibilidad de un sistema que una base de datos deficientemente preparada, ya que los datos capturados con escasa calidad afectan negativamente la productividad del sistema[1].

Además, los errores pueden resultar muy caros. Pensemos, por

---

[1]Desde luego, el éxito en la recuperación de los documentos depende también de las técnicas de recuperación, o sea, de la calidad del sistema mismo.

ejemplo, cómo el mayor tiempo de uso de algunas bases de datos bibliográficos en línea aumenta el costo de las búsquedas. De igual forma, un documento no recuperado puede generar una solicitud de compra innecesaria.

La investigación de procesos encaminados a prevenir y corregir diferentes tipos de errores en la información de las bases de datos bibliográficos debe comenzar con el diagnóstico preciso de los problemas por resolver.

ANTECEDENTES

El presente estudio de diagnóstico fue realizado dentro del marco del Proyecto de Control de Autoridad de la Biblioteca Daniel Cosío Villegas, por la que suscribe, coordinadora del proyecto, y Shirley Ainsworth, referencista de la biblioteca. El diagnóstico fue presentado en un documento interno en marzo de 1993 con el fin de proporcionar los elementos necesarios en la toma de decisiones conducentes a la limpieza del catálogo automatizado de la biblioteca para su difusión pública.

Los objetivos principales de esta depuración eran, por un lado, permitir la consulta del catálogo a través de un "OPAC" (Online Public Access Catalog) a partir de julio del mismo año y, por otro, preparar la información para la edición de un disco compacto con los registros de monografías de la biblioteca a finales del año[2].

PROCEDIMIENTO

El diagnóstico de errores se realizó en:

1. Registros completos
2. Indices por campos
3. Problemas específicos

Para cada uno se siguió un procedimiento diferente:

1. Procedimiento con registros completos

Supusimos que el universo de registros era relativamente homogéneo y extrajimos dos muestras al azar, de 80 y 70 registros respectivamente[3]. De todos ellos obtuvimos las correspondientes

---

[2]Para conocer los antecedentes y desarrollo del proyecto de automatización integral de la Biblioteca Daniel Cosío Villegas, véase: Quijano Solís, Alvaro y Clotilde Tejeda, "La organización del sistema integral de automatización de la Biblioteca Daniel Cosío Villegas", ponencia presentada en las XXII Jornadas Mexicanas de Biblioteconomía, 1991; y: Tejeda, Clotilde y Alvaro Quijano Solís, "Conversión retrospectiva: piedra angular de la automatización de la Biblioteca", ponencia presentada en el Seminario Anual ABIESI 1991.

[3]El tamaño del catálogo automatizado de la biblioteca es de aproximadamente 280,000 registros.

impresiones para su revisión. Estas impresiones contenían todos los campos codificados de cada registro tal como fueron capturados.

A continuación revisamos cada uno de los registros seleccionados y anotamos en ellos todos los errores que fuimos encontrando. Decidimos agrupar los errores por tipos, de acuerdo tanto a trabajos anteriores[4], como al tipo de acción que se debería tomar para abatirlos. Estas acciones serán contempladas más adelante en este mismo documento.

Así, consideramos errores del:

-Tipo 1: Espacios, puntuación, mayúsculas, errores tipográficos y ortográficos.
-Tipo 2: Codificación y etiquetado MARC
-Tipo 3: Forma y uso de las entradas.
-Tipo 4: Otros.

Sin embargo, observamos que este procedimiento presentaba las siguientes limitaciones:

a) Al considerar los registros aisladamente no se podían detectar totalmente las inconsistencias entre datos que resultan de la comparación con registros semejantes o cercanos. Por ejemplo, las diferentes formas de entrada para autores, materias, etc.

b) Por la misma razón no se podía diagnosticar la frecuencia total de errores específicos no encuadrados en las tres primeras categorías anteriormente definidas (errores del tipo 4).

Por otro lado, no todos los errores se consideran igualmente graves, dependiendo del campo en el que se presenten. Los errores que aparecen en los campos indizados revisten mayor importancia que aquellos encontrados en los campos no indizados. Así pues, vimos la necesidad de un análisis posterior con otros procedimientos.

2. Procedimiento con índices por campos

En este caso se sacaron impresiones de partes secuenciales, con longitudes arbitrarias, tomadas al azar de los índices de: autor personal (incluyendo todas las etiquetas MARCOLMEX[5] para autor personal), autor corporativo (etiqueta 110a), conferencias (en cualquier campo MARC para conferencias), editorial (260b), series (440a), temas generales principales (690ap)[6], temas geográficos principales (691ap), temas principales de autor personal (692ap), temas principales de autor corporativo (693ap), temas principales de conferencias (694ap) y temas principales de títulos uniformes (695ap). Estas impresiones incluyeron la entrada, seguida del campo correspondiente, el número total de registros con

---

[4]Véase: Edward T. O'Neill, y Diane Vizine-Goetz. "Computer generation of a subject authority file", Proceedings of the ASIS Annual Meeting, vol. 19, 1982, p. 220-223.

[5]El formato MARCOLMEX es una adaptación del formato USMARC, usado por la Biblioteca de El Colegio de México.

[6]Llamamos "tema principal" al primer encabezamiento de materia del registro. La notación 69Xap significa el subcampo "a" de un tema principal (indicado por la letra "p")

dicha entrada y el número de cada registro.

También aquí se agruparon los errores en los tipos ya descritos. Cuando una misma entrada presentó varias formas distintas se tomó una de ellas como buena y el resto de las variantes como errores.

Cabe señalar que en la mayoría de los campos sólo se tomó el primer segmento de información (subcampo a), puesto que, por el momento y por razones de espacio, así están definidos los índices en nuestra base de datos. Sin embargo, creemos que sería más útil en el futuro tomar partes de índices con campos completos, ya que la información adicional que contienen los otros subcampos puede constituir la diferencia entre una y otra entrada.

3. Problemas específicos

Hay problemas en el catálogo que escapan a la categorización antes descrita en los tres primeros rubros y que, aunque tal vez poco frecuentes, pueden revestir importancia por afectar significativamente la recuperación ya sea de la información o del material mismo.

Este tipo de problemas aparecen a diario en la interacción continua con el catálogo y nos fueron señalados por los bibliotecarios al ser encuestados informalmente. Sin embargo, resultan generalmente más difíciles de "imaginar" a priori y no siempre son fácilmente cuantificables.

En general, estos problemas pueden resolverse con relativa facilidad e implican básicamente tiempo de los bibliotecarios, sin necesidad de colaboración de personal de cómputo.

Algunos de estos problemas que consideramos importantes serán detallados más adelante.

RESULTADOS

Análisis de registros completos:

Como muestra el cuadro de la página siguiente, la distribución de errores de los tres primeros tipos fue muy similar, decreciendo ligeramente de los errores de tipo 1 (24.9 %) a los de tipo 2 (24.2 %) y los del tipo 3 (23.5 %). Los errores del tipo 4 presentaron una ocurrencia mucho menor (4.6 %). El total de los registros afectados por errores de los cuatro tipos representó el 77.1 % de la muestra. O, expresado de otra manera, la frecuencia promedio de errores por registro fue de 0.915.

- Resultados del análisis de registros completos -

	Muestra 1	Muestra 2	Muestra Total
Tamaño	73	80	153
ET1	15	36	51
RA	12	26	38
%	16.4	32.5	24.9
ET2	29	14	43
RA	24	13	37
%	32.9	16.2	24.2
ET3	20	19	39
RA	20	16	36
%	27.4	20.0	23.5
ET4	3	4	7
RA	3	4	7
%	4.1	5.0	4.6
ET	67	73	140
RA	59	59	118
%	80.8	73.8	77.1

FE= 140/153=0.915

FE= Frecuencia de error: número total de errores/número total de registros analizados
ET1= Núm. de errores del tipo 1; ET2= Núm. de errores del tipo 2; ET3= Núm. de errores del tipo 3; ET4= Núm. de errores del tipo 4; ET= Núm. total de errores de los cuatro tipos; RA= Núm. de registros afectados; %= Porcentaje de registros afectados por error sobre la muestra total.

Análisis de índices:

1) Búsqueda: autor personal

Tamaño de la muestra	:	726	entradas
Registros relacionados	:	1635	registros
Errores	:	66	entradas
Porcentaje de error	:	9.1%	entradas

Errores por tipo:

T1:	35	entradas	(4.8 %)
T2:	7	entradas	(1.0 %)
T3:	21	entradas	(2.9 %)
T4:	4	entradas	(0.6 %)

Nota: Hubo entradas con dos tipos de errores a la vez.

2) Búsqueda: autor corporativo (dos muestras)

Tamaño total de las muestras	:	223	entradas
Registros relacionados	:	1270	registros
Errores	:	78	entradas
Porcentaje de error	:	35 %	

Errores por tipo

T1	:	32	(14.3 %)
T2	:	20	(9.0 %)
T3	:	26	(11.6 %)

Nota: Hubo entradas con dos tipos de errores a la vez. No hubo entradas con errores del tipo 4.

3) Búsqueda: conferencias

Tamaño de la muestra	:	360	entradas
Registros relacionados	:	1902	registros
Errores	:	49	entradas
Porcentaje de error	:	13.6%	entradas

Errores por tipo

T1:	28	entradas	(7.8 %)
T2:	15	entradas	(4.2 %)
T3:	7	entradas	(1.9 %)

Nota: Hubo entradas con dos tipos de errores a la vez. No hubo entradas con errores del tipo 4.

4) Búsqueda: editorial (dos muestras)

Tamaño total de las muestras	:	272	entradas
Registros relacionados	:	1597	registros
Errores	:	57	entradas
Porcentaje de error	:	20.9%	entradas

Errores por tipo

T1	:	29	(10.7 %)
T2	:	7	(2.6 %)
T3	:	22	(8.1 %)

Nota: Aquí es notorio cómo las diferencias formales nos dan una enorme variedad de entradas distintas para una misma editorial (ej.: Maissoneuve, 19; Putnam's, 15; Allen and Unwin, 13). Además estas diferencias hacen que se alejen unas de otras, siendo más difíciles de detectar. Por lo mismo, el conteo de errores lo estimamos aproximado. Tómese en cuenta que

este campo, en principio, no está sujeto a reglas estrictas de uniformidad en las entradas. Sin embargo, las dificultades en la recuperación podrían hacer recomendar algún tipo de control.

5) Búsqueda: serie

Tamaño de la muestra	315	entradas
Registros relacionados:	1071	registros
Errores	82	entradas
Porcentaje de error	26.0%	entradas

Errores por tipo

T1:	46	(14.6 %)
T2:	26	(8.2 %)
T3:	10	(3.2 %)

Nota: El mayor porcentaje de errores tipográficos se debe en parte a la existencia de series en alemán dentro de la muestra.

6) Búsqueda: tema general principal (dos muestras)

Tamaño total de las muestras:	514	entradas
Registros relacionados	4118	registros
Errores	154	entradas
Porcentaje de error	30.0%	entradas

Errores por tipo

T1:	46	(8.9 %)
T2:	103	(20.0 %)
T3:	24	(4.7 %)

Nota: Hubo entradas afectadas por más de un error a la vez.

7) Búsqueda: tema geográfico principal (dos muestras)

Tamaño total de las muestras:	509	entradas
Registros relacionados	1337	registros
Errores	151	entradas
Porcentaje de error	29.7%	entradas

Errores por tipo

T1:	34	(6.7 %)
T2:	36	(7.1 %)
T3:	88	(17.3 %)
T4:	3	(0.6 %)

Nota: Hubo entradas afectadas por más de un error a la vez.

8) Búsqueda: tema principal autor personal

Tamaño de la muestra	102	entradas
Registros relacionados:	256	registros
Errores	26	entradas
Porcentaje de error	25.5%	entradas

Errores por tipo

T1:	10	(9.8 %)
T2:	9	(8.8 %)
T3:	4	(3.9 %)
T4:	4	(3.9 %)

Nota: Hubo entradas afectadas por más de un error a la vez.

9) Búsqueda: tema principal autor corporativo

Tamaño de la muestra	:	130 entradas
Registros relacionados:		711 registros
Errores	:	51 entradas
Porcentaje de error	:	39.2% entradas

Errores por tipo

T1:	6 (4.6 %)
T2:	26 (20.0 %)
T3:	19 (14.6 %)

10) Búsqueda: tema principal conferencia

Tamaño de la muestra	:	20 entradas
Registros relacionados:		24 registros
Errores	:	10 entradas
Porcentaje de error	:	50.0% entradas

Errores por tipo

T2:	9 (45 %)
T3:	1 (5 %)

Nota: El tamaño de la muestra fue menor porque el tamaño del índice también lo es. No hubo errores del tipo 1 ni del tipo 4.

11) Búsqueda: tema principal título uniforme

Tamaño de la muestra	:	23 entradas
Registros relacionados:		70 registros
Errores	:	9 entradas
Porcentaje de error	:	39.1% entradas

Errores por tipo

T1:	1 (4.3 %)
T2:	8 (34.8 %)

Nota: El tamaño de la muestra fue menor porque el tamaño del índice también lo es. No hubo errores del tipo 3 ni del tipo 4.

Análisis de problemas específicos:

Los problemas que aparecieron como atípicos en los anteriores procedimientos, junto con los señalados por los bibliotecarios en la encuesta informal que realizamos fueron:

- Registros sin título: 19
- Registros sin número de adquisición: 37
- Signos de arroba (@) inadecuados (por problemas de transferencia de información capturada previamente en diskettes): 1983 registros.
- Temas "pegados" y revueltos (por problemas de transferencia de información capturada previamente en diskettes): aproximadamente mil registros.
- Analíticas de revistas del COLMEX con distinta signatura topográfica: detectado, pero no cuantificado.
- Registros en los que la información codificada de los campos fijos no coincidía con la información correspondiente de los campos variables: detectado, pero sólo parcialmente cuantificado.

- Duplicados (registros idénticos capturados dos veces): detectado, pero no cuantificado.

RECOMENDACIONES

Aunque no hay un estándar específicamente definido sobre el nivel de calidad aceptable para los datos incluidos en las bases de datos bibliográficos, parece claro que éste debería marcarse por debajo de 1. Es decir, una frecuencia promedio mayor de un error por registro indicaría un nivel deficiente de calidad de la base, y cuanto más se acercara a cero dicha frecuencia la calidad de la base aumentaría.

Cada biblioteca debe estimar, de acuerdo a criterios de costo-beneficio, qué nivel de calidad puede alcanzar con los recursos materiales y humanos de que dispone.

El estudio de Reeb que mencionamos en la bibliografía fija un nivel máximo deseable de error de 0.8 para los registros recién catalogados. Este nivel presumiblemente disminuye al incorporarse los registros a la base, ya que se supone que pasaron una revisión y se corrigieron los errores detectados.

En el caso de la Biblioteca Daniel Cosío Villegas cabe señalar que a los errores de proceso, que se infiltraron durante el programa de conversión retrospectiva del catálogo, se unieron los de transferencia de la información que había sido anteriormente capturada con otro sistema. De cualquier manera, consideramos que antes de ofrecer al público nuestra base de datos debíamos disminuir el nivel de error de los registros y, a partir de que presentamos nuestro informe de diagnóstico, se dedicaron recursos a este fin.

Las medidas propuestas para abatir los errores que encontramos fueron las siguientes:

Para los errores del tipo 1:
-Acciones preventivas: Establecimiento de mecanismos de validación de datos a través de catálogos de autoridad.
-Acciones correctivas: Algoritmo de corrección de errores tipográficos[7].

Para los errores del tipo 2:
-Acciones preventivas: Establecimiento de mecanismos de catalogación asistida, por medio de pantallas de ayuda que especifiquen códigos MARC y uso de los mismos.

---

[7]Para una explicación del funcionamiento de este tipo de algoritmos véase: Moreno, Pilar M. y Alvaro Quijano Solís "Los catálogos de autoridad de materia en el contexto automatizado", ponencia presentada en el V Coloquio sobre Automatización de Bibliotecas, Colima 1991, pp. 9-11

Mejoramiento del sistema de revisión ("analiza") de registros catalogados antes de su ingreso definitivo al catálogo.

-Acciones correctivas: Combinación de procedimientos automáticos (como la emisión de informes ad hoc resultado de búsquedas específicas, cambios globales automáticos, etc.) y manuales (revisión uno a uno de registros para seleccionarlos según tipo de problema, corrección uno a uno de registros, etc.)

Para los errores del tipo 3:

-Acciones preventivas: Creación y mantenimiento de catálogos de autoridad.

-Acciones correctivas: Combinación de procedimientos automáticos y manuales.

Para los errores del tipo 4:

-Acciones preventivas: Todas las ya mencionadas.

-Acciones correctivas: Para los problemas de duplicados, actualizar resultado del inventario (mediante etiquetado de códigos de barras) con el sistema.
Combinación de procedimientos automáticos y manuales.

General:

Repetir estas mismas pruebas después de haber puesto en marcha algunas de las acciones propuestas, para así medir el grado de calidad alcanzado gracias a éstas.

BIBLIOGRAFIA

Bourne, Charles P. "Frecuency and impact of spelling errors in bibliographic data bases", Information Processing and Management, v.13, 1977, pp. 1-12

Carpenter, Ray L. y Ellen Storey Vasu. Métodos estadísticos para bibliotecarios. México, Dirección General de Bibliotecas de la UNAM, 1980.

Johnson, Judith J. y Clair S. Josel. "Quality control and the OCLC data base: a report on error reporting", Library Resources and Technical Services, Jan.-Mar. 1981, pp. 40-47

Moreno Jiménez, Pilar María y Shirley Ainsworth, Informe de diagnóstico de errores en el catálogo automatizado de la BDCV. Documento interno de trabajo. México, 1993.

O'Neill, Edward T. y Diane Vizine-Goetz. "Computer generation of a subject authority file", Proceedings of the ASIS Annual Meeting, v.19, 1982, pp. 220-223

O'Neill, Edward T. y Diane Vizine-Goetz. "Quality control in online databases", Annual Review of Information Science and Technology, v.23, 1988, pp. 125-156

Reeb, Richard. "A quantitative method for evaluating the quality of cataloging", Cataloging and Classification Quartely, v.5 (2), Winter 1984, pp. 21-26