LINEAMIENTOS PARA LA RECOPILACIÓN DE

BASE de DATOS

F. W. Lancaster

cub

Lineamientos para la recopilación de bases de datos

CENTRO UNIVERSITARIO DE INVESTIGACIONES BIBLIOTECOLÓGICAS
SERIE:
FOLLETOS DE APOYO PROFESIONAL 3

Lineamientos para la recopilación de bases de datos

Frederick W. Lancaster

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

QA76.9
D3L34 Lancaster, Frederick W.

Lineamientos para la recopilación de bases de datos / Frederick W. Lancaster ; rev. tec. Surya Peniche de Sánchez Macgrégor. — México : UNAM, Centro Universitario de Investigaciones Bibliotecológicas, 1996.
12 p. — (Folletos de apoyo profesional ; 3)

1. Administración de Bases de Datos I.t.

Revisión técnica: **Surya Peniche de Sánchez Macgrégor**

Diseño de portada: **D.G. Ignacio Rodríguez**
D.G. Mario Ocampo

Primera Edición 1996

Ciudad Universitaria, 04510, México, D.F.
Impreso y hecho en México

CONTENIDO

Introducción

Durante los últimos 30 años, el costo de la distribución de información en forma electrónica, comparado con el de la impresión en papel, se ha venido reduciendo de manera constante. Esto ha llevado a una explosión en la producción de bases de datos en forma electrónica, tanto de las diseñadas para el uso interno de una sola institución, como de aquellas diseñadas para una difusión más amplia. Esta explosión se ha incrementado notablemente debido a la tecnología de CD-ROM, que ofrece la manera de distribuir grandes cantidades de datos en una forma práctica y económica, así como por la facilidad de instalar bases de datos en Internet.

Pero una base de datos no sólo se puede justificar sobre la base de que es relativamente fácil y económica de producir y de fácil disponibilidad. Si bien la producción de bases de datos tiene como finalidad mejorar el acceso a la información, la proliferación continua de bases de datos —particularmente de aquellas dedicadas a áreas muy reducidas de conocimientos o de aquellas que se traslapan sustancialmente con otras bases de datos— de hecho tiene el efecto contrario, ya que aumenta la fragmentación de la información del mismo modo en que la proliferación de publicaciones académicas dedicadas a áreas cada vez más especializadas de investigación ha fragmentado los registros académicos.

Aun cuando la producción de una nueva base de datos se justifica en términos de su contenido (por ejemplo, por el hecho de que abarca recursos no cubiertos en ninguna otra) esto no es suficiente para ello, a menos que satisfaga otros criterios relativos a su calidad y utilidad.

Estos criterios ayudan a las instituciones a decidir si deben o no crear una nueva base de datos, o bien ampliar la existente, tanto en términos de su contenido como de su utilidad en general. Aunque tratan específicamente de la situación en México, los criterios serían también aplicables en cualquier otro lugar.

Antes de ocuparse de los criterios mismos, es necesario examinar los diversos tipos de bases de datos producidas en México, porque hasta cierto punto por lo menos, a los diferentes tipos de bases de datos corresponden criterios diferentes.

Tipos de bases de datos

Las principales bases de datos que se producen en México, según se desprende del Directorio de Bases de Datos de América Latina y el Caribe (Facultad de Filosofía y Letras, Universidad Nacional Autónoma de México) y del Catálogo 1994-1995 del Centro Nacional Editor de Discos Compactos (Universidad de Colima), pueden dividirse en seis categorías principales: 1) bases de datos bibliográficos, 2) bases de datos de textos completos, 3) catálogos de materiales especiales, 4) bases de datos de imágenes, 5) datos estadísticos, numéricos y otros relacionados con ellos, 6) recopilaciones descriptivas, y 7) otras fuentes de consulta. Cada una de estas categorías se verá con más detalle a continuación.

Bases de datos bibliográficas

Éstas incluyen los catálogos de los acervos de una sola biblioteca o de grupos de bibliotecas, así como bases de datos que contienen referencias bibliográficas de las publicaciones de un tipo en particular (v.g., tesis) o de aquellas dedicadas a una área temática en particular. También se incluirían aquí los catálogos de las publicaciones de una institución o de un grupo de instituciones.

Bases de datos de textos completos

En esta categoría aparecen las bases de datos que contienen el texto completo de: 1) leyes, reglamentos y material semejante, 2) de una sola publicación periódica o grupo de publicaciones periódicas, y 3) de otros tipos de materiales (v.g., ponencias presentadas en una conferencia).

Catálogos de materiales especiales

Esta categoría incluye catálogos de materiales audiovisuales, de exposiciones, de manuales de operación y así sucesivamente.

Bases de datos de imágenes

Éstas son bases de datos que se relacionan principalmente con imágenes (v.g., de pinturas, piezas de museo, fenómenos biomédicos, fotografías).

Compendios de datos

Se incluyen aquí todas las bases de datos consistentes en recopilaciones de datos estadísticos, numéricos y otros que guardan relación con ellos.

Recopilaciones descriptivas

Son bases de datos que presentan información descriptiva referente a una región o a una institución en particular.

Otras fuentes de consulta

Las bases de datos de este tipo incluyen directorios (v.g., de investigadores, proyectos de investigación, oportunidades educativas), diccionarios biográficos, y equivalentes electrónicos similares a los "libros de consulta" convencionales.

No todas las bases de datos caen nítidamente dentro de una sola categoría; algunas presentan características que caben en más de una (v.g., aquellas que contienen texto completo además de referencias bibliográficas).

Criterios concernientes al contenido

Los principales criterios para justificar la creación de una nueva base de datos son aquellos que se refieren a sus contenidos. Estos criterios son los siguientes:

1.- Campo de cobertura.

2.- Accesibilidad del contenido.

3.- Pronóstico de la cobertura.

4.- Continuidad de la cobertura.

5.- Vigencia de la cobertura.

6.- Masa crítica.

Campo de cobertura

La principal justificación para la creación de una nueva base de datos es el hecho de que el material que habrá de incluirse no está cubierto, o bien cubierto en otra parte. "No estar bien cubierto" incluiría el caso en que el material que interesa está cubierto en otro sitio, pero de manera fragmentada (v.g., disperso en muchas bases de datos diferentes).

Obviamente, que no tendría sentido que una institución de México crease una base de datos que duplicara muy de cerca alguna de las bases de datos internacionales ya existentes. Por lo tanto, es obvio que la prioridad más alta debería darse a la cobertura de publicaciones que se generan en México, ya que no es probable que éstas sean cubiertas satisfactoriamente por las principales bases de datos producidas en los Estados Unidos, en el Reino Unido y en otros sitios. La "cobertura mexicana" podría extenderse a la "cobertura de América Latina", a la "cobertura Hispano Americana", a la "cobertura en lengua española", o bien a la "cobertura de América Latina y el Caribe", donde el tema de la base de datos es de interés para los mexicanos y se requiere un ámbito geográfico o lingüístico más amplio para complementar la cobertura de las bases de datos internacionales.

También se justifican plenamente las bases de datos que incluyen publicaciones mexicanas, así como publicaciones producidas en otros sitios y referentes al comercio mexicano. La justificación aquí es de conveniencia. Aunque los puntos de referencia no mexicanos concernientes a México pueden estar bien cubiertos por las bases de datos internacionales, conviene reunirlos con la información recabada en México, sobre todo si dicha información se encuentra dispersa en varias de las bases de datos internacionales.

El hecho de que una base de datos de dimensiones razonables pudiera crearse sólo con publicaciones mexicanas no justifica, por sí misma, su creación. Por ejemplo, raras veces se justificaría construir una base de datos que fuera una subserie de otra base de datos mexicana ya existente (v.g., una base de datos sobre contaminación del aire, si ya existe una base de datos que se ocupa de los problemas ambientales en general; o una base de datos de tesis mexicanas sobre un tema en particular, si ya existe una base de datos completa sobre todas las tesis mexicanas, y así sucesivamente). Dicha duplicación estimula la proliferación innecesaria de bases de datos, y con ello el incremento en la fragmentación del conocimiento.

En la mayoría de los casos no se justificaría que una institución mexicana crease una base de datos internacional, que intentara incluir toda la literatura mundial referente a un cierto tópico. Sin embargo, podrían existir excepciones a dicha regla general. Se justificaría la construcción de una base de datos amplia (v.g., extraída de la literatura mundial)

sobre alguna área temática especializada de particular importancia para México si no existiera una base de datos especializada sobre este mismo tema. Quizá el ejemplo más claro sería el de una base de datos sobre alguna cosecha, o sobre un grupo de cosechas en particular, de especial importancia para la economía de la agricultura de una nación, como podría suceder en países latinomaericanos (una base de datos sobre café en Costa Rica, y otra sobre azúcar en Cuba). Algo semejante podría ser aplicable en el área de salud pública; esto es, una extensa base de datos internacional podría justificarse si se refiriese a una enfermedad o grupo de enfermedades que fuesen particularmente predominantes en México.

Accesibilidad del contenido

El criterio de "campo de cobertura" se aplica más a las bases de datos bibliográficas y a las que se dedican a las estadísticas de datos numéricos; para otros tipos puede ser más apropiado algún criterio diferente. Por ejemplo, las bases de datos que presentan imágenes y descripciones de objetos (v.g., los de un museo, de las pinturas de una galería de arte), o de todos los objetos de un cierto tipo (v.g., máscaras o trajes tradicionales) son fáciles de justificar sobre la base de que coleccionan imágenes que pueden estar diseminados por áreas geográficamente distantes, o que ponen a disposición de estudiantes y estudiosos las imágenes de colecciones que quizá no puedan ellos mismos visitar con facilidad. Mejorar la accesibilidad a los objetos y obras de arte mediante fotografías es una justificación importante de la existencia de bases de datos de imágenes, especialmente cuando las colecciones son exclusivas de México.

La accesibilidad del contenido tiene dimensiones que sobrepasan a las de la geografía. Por ejemplo, una base de datos de texto completo puede justificarse por el hecho de que la habilidad de localizarlo por computadora facilita en gran medida el acceso intelectual a los contenidos mismos de dicho texto. Así, el texto completo de un periódico en CD-ROM, desde su inicio, puede ser mucho más valioso que tener el periódico impreso en papel y encuadernado, no sólo por la facilidad para su manejo y fines de conservación, sino también por la accesibilidad de su contenido, ya que el texto puede ser buscado mediante diversas combinaciones de palabras.

Pronóstico de la cobertura

Los lineamientos que conforman el "pronóstico de cobertura" se refieren primordialmente a la producción de bases de datos para su distribución (por ejemplo, a través de las ventas), más que al uso interno exclusivo de una sola institución. Para que sea útil a otros, la cobertura de una base de datos debe ser completamente predecible. El contenido temático del que se trata debe estar definido con precisión. Una base de datos, que cubre un grupo heterogéneo de temas que están conectados de manera imprecisa, es de escaso valor porque los usuarios potenciales no sabrán lo que está o no incluido.

Dentro del campo tratado (área temática y/o tipo de material), la cobertura deberá de ser completa, o lo más completa que sea posible. Una base de datos que pretende una cobertura "selectiva" de algún tema o tipo de material tendrá un valor muy limitado, porque no es probable que los usuarios potenciales sepan cuáles han sido los criterios de selección. En el caso de una base de datos que indiza la literatura periódica, el editor debe aclarar cuáles títulos de publicaciones periódicas se cubren, y debe ser consistente al indizar todos los números de dichas publicaciones.

En forma similar, una base de datos que pretende incluir el texto completo de una revista, o grupo de revistas correspondientes a un periodo en particular, deberá incluir todos los números y todos los artículos de este periodo. Los usuarios tendrán poca confianza en una base de datos si encuentran vacíos impredecibles e inexplicables en su cobertura. El factor de "predictibilidad" es un elemento importante para establecer la confianza de los usuarios y de los usuarios potenciales.

Continuidad de la cobertura

Ninguna institución debe producir una base de datos que requiera actualización, a menos que esté dispuesta y tenga la capacidad, en especial, la económica, para comprometerse a mantenerla actualizada. Una base de datos que cubre solamente algunos años —de una revista en particular, de publicaciones de cierto tipo, o de publicaciones que se ocupan de una área temática específica— y que no se complementa o actualiza de alguna otra forma, es de escasa utilidad. Si bien una institución puede tener buenas razones para descontinuar una base de datos (v.g., ya no se necesita porque ha sido sustituida por otra más amplia o de mayores alcances), perderá credibilidad si se descontinúa por razones que no son obvias a los suscriptores y usuarios.

Vigencia de la cobertura

Claro está que algunas bases de datos nunca tendrán que actualizarse (v.g., aquella que reproduzca los trabajos de un artista que ha fallecido), y otras, que existen principalmente con fines históricos o de archivo, requieren una actualización poco frecuente.

Por otra parte, las bases de datos creadas principalmente con fines de alerta informativa deben actualizarse con frecuencia para que conserven su valor. Obviamente, una base de datos de investigaciones en proceso debe incluir los proyectos de investigación que aún se están realizando, más que aquellos que se han concluido y deberá incluir, de preferencia, proyectos en su momento inicial. Una base de datos que contiene información acerca de conferencias y otras reuniones deberá incluir las listas de dichas acciones mucho antes de que tengan lugar. Asimismo, una base de datos bibliográfica dedicada a una área temática en particular deberá actualizarse con frecuencia, no menor que trimestral, si es en forma de CD-ROM, y mensual si es accesible en línea.

La actualización deberá ser predecible, ocurriendo a intervalos regulares programados. Una actualización "irregular" es insatisfactoria desde el punto de vista del consumidor.

Masa crítica

Éste es probablemente el más controvertido de los criterios relativos al contenido. **Masa crítica** se refiere al hecho de que una base de datos debe ser suficientemente grande para ser de interés y valor para un número importante de usuarios potenciales. "Suficientemente grande" no es un concepto preciso, no se puede cuantificar con exactitud. Sin embargo, sería difícil justificar una base de datos bibliográfica en CD-ROM acerca de un tema tan especializado, del que solamente se publica un puñado de artículos cada año; tampoco resultaría económico actualizarla, incluso anualmente.

Otro caso es el de los catálogos de bibliotecas. Tratándose de una biblioteca de gran importancia nacional, puede valer la pena publicar el catálogo en forma de CD-ROM para obtener una distribución más amplia. Sin embargo, en el caso de bibliotecas de menor importancia, no es probable que el catálogo de una sola biblioteca amerite mayor distribución. Los catálogos colectivos que representan las colecciones de varias bibliotecas serán mucho más valiosos. En todo caso, un catálogo colectivo deberá tener alguna lógica subyacente para ser de utilidad. Esto es, deberá incluir bibliotecas que estén lógicamente relacionadas de alguna forma —por el contenido temático, o por ser de una ciudad o región en particular— para facilitar que se compartan los recursos.

Una base de datos que contenga el texto completo de una sola publicación periódica, o de un mismo diario puede justificarse si la publicación es de especial importancia, particularmente si la cobertura temporal está completa (v.g., si se remonta hasta los inicios de su publicación) debido a que este tipo de base de datos puede tener significación histórica o archivística. Sin embargo, una base de datos que cubre varios diarios, o la que incluye el texto de varias publicaciones periódicas (si se refieren a la misma área temática) es probable que resulte mucho más valiosa.

Una base de datos que contenga el texto completo de las ponencias presentadas en varias conferencias puede valer la pena, siempre que la recopilación sea lógica (v.g., todos los documentos de las conferencias celebradas por una agrupación durante un periodo de varios años, o todas las ponencias de una conferencia sobre algún tema amplio), pero una base de datos que contenga el texto de las ponencias presentadas en una o dos conferencias aisladas no es probable que tenga gran valor.

Criterios semejantes se aplican a otros tipos de bases de datos; las de imágenes de objetos en varios museos es probable que sean de mayor valor que las que se restringen a un solo museo (a menos que sean de dimensiones e importancia poco comunes).

Lo mismo se aplicaría a las galerías de arte: una base de datos dedicada al trabajo de un solo artista se puede justificar,

pero una que se dedique al de varios artistas puede ser más valiosa, siempre que los artistas estén lógicamente conectados de alguna manera (v.g., que todos sean de una misma región, o que compartan un estilo o forma en particular).

La masa crítica no es solamente cuestión de dimensiones útiles. A la larga, se refiere también a cuestiones de fragmentación y a la economía de acceso. Esto se aplica más a bases de datos distribuidas en forma de CD-ROM. El costo de producir una base de datos grande en CD-ROM puede ser menor que el de producir una más pequeña. Debido a que los presupuestos institucionales son limitados, la proliferación de bases de datos más pequeñas tiende a reducir tanto la accesibilidad como el tamaño del mercado potencial para los productos de CD-ROM. Para dar un ejemplo: una institución que esté dispuesta y tenga posibilidades de adquirir un CD-ROM en el que se ilustre el trabajo de cinco artistas, puede no estar en disposición o en posibilidades de adquirir cinco discos semejantes, cada uno referente a un solo artista. El factor de fragmentación, que afecta a las bases de datos asequibles a través de las redes, así como aquéllas en CD-ROM, también es significativo. En general, es más eficiente y económico buscar en una sola base de datos más amplia, que en varias más pequeñas, todas referentes a áreas temáticas muy próximas, especialmente cuando cada base de datos puede emplear enfoques completamente diferentes para indizar y puede requerir el uso de estrategias y hasta de lenguajes diferentes para la búsqueda.

Criterios concernientes a calidad y utilidad

Refiriéndose específicamente a las bases de datos en CD-ROM, Jacsó (1992) hace una distinción entre los términos **software**, **hardware** y **dataware**. Dataware se refiere a los contenidos de la base de datos, mientras que software se refiere a las capacidades que se proporcionan para explotar el dataware, como son, interface de usuarios, capacidad de búsqueda, capacidad de salida, y así sucesivamente. Los aspectos referentes al hardware están fuera del ámbito de estos lineamientos. A continuación se tratan los requerimientos principales del software y del dataware.

Requerimientos del software

Las responsabilidades del productor de la base de datos se extienden más allá de su elaboración. Es de igual importancia el hecho de que el productor deba proporcionar de alguna manera el software necesario para permitir a los usuarios la explotación efectiva de la base de datos. Sólo para las bases de datos bibliográficas ya se han escrito miles de páginas respecto de los requerimientos de software para interfaces de los usuarios, capacidad de búsqueda, capacidad de salida y así sucesivamente. Estos lineamientos solamente podrán cubrir requerimientos muy generales.

El software relacionado con la base de datos debe proporcionar a los usuarios medios efectivos para la búsqueda, obtención de información, y producción de diversos tipos de salidas o productos. Dependiendo del tipo de base de datos

de que se trate, las búsquedas pueden requerir que se utilicen elementos de datos precisos y nada ambiguos (v.g., año de publicación, ISBN o ISSN), relativamente poco ambiguos (v.g., el nombre de una persona), o muy ambiguos (v.g., términos que describen el contenido temático), por lo que se debe proporcionar a los usuarios capacidades de búsqueda para los tres tipos de elementos de datos.

Los usuarios deben tener la habilidad de buscar todos los tipos de términos que pudiesen indicar el contenido temático (términos de indización y/o palabras clave), en cualquier combinación lógica (booleana), y quizá la que se requiere para combinar términos temáticos con términos no temáticos (v.g., nombres de autores, fechas de publicación). Para las bases de datos bibliográficas y de texto completo, la habilidad para truncar los términos (búsqueda mediante fragmentos de palabras) puede ser importante, y la búsqueda por proximidad de palabras (especificando la cercanía de las palabras para que se les pueda considerar relacionadas) será una característica esencial en la recuperación del texto completo.

Se deberá proporcionar a los usuarios la posibilidad de visualizar todos los términos que se pueden buscar en la base de datos (palabras que ocurren en texto libre, así como términos tomados del tesauro, o de otros vocabularios controlados), en el contexto de los términos alfabéticamente próximos.

Además de las capacidades efectivas de búsqueda, los usuarios deben tener algunas opciones de salida —por ejemplo, la posibilidad de imprimir o desplegar registros de diversas extensiones y/o diferentes formatos— y quizá la posibilidad de organizar la salida en diversas formas (v.g., los últimos registros mostrados, o los primeros impresos).

La interface —a través de la cual los usuarios buscan en la base de datos y especifican las opciones de salida— deberá ser simple y autoexplicativa para facilitar y estimular su uso. Esto puede lograrse de varias formas: menús, comandos para el usuario o un curso tutorial completo. El libro de Jacsó (1992) proporciona mucha información útil sobre diseño apropiado de interfaces.

Requerimientos del dataware

Estos lineamientos se ocupan con mayor detalle de los requerimientos del dataware que de los del software, por dos razones: 1) los del dataware están menos cubiertos en otras fuentes, y 2) los del dataware se pueden generalizar mejor que los del software para las bases de datos de diferentes tipos.

Extensión y detalle de los registros

Sea cual fuere el tipo de bases de datos, los registros incluidos deben ser suficientemente completos para proporcionar toda la información que sea necesaria para el usuario típico. Los registros bibliográficos serán más útiles si están completos, que abreviados. Deberán incluirse resúmenes siempre que sea posible. Un resumen detallado puede evitar que un usuario desperdicie esfuerzos tratando de adquirir un

elemento de información que puede resultarle de poco interés. En algunos casos un resumen informativo puede servir como sustituto del elemento mismo de información (v.g., el usuario obtiene lo que necesitaba sin tener que consultar el original). En forma semejante, las descripciones de objetos (v.g., pinturas o piezas de museo) deberán ser de lo más completas para compensar el hecho de que el usuario solamente tenga acceso a una imagen y no al objeto mismo.

Número de puntos de acceso

Entre más extenso es el registro, más puntos de acceso debe proporcionar. **Punto de acceso** se refiere a un elemento de dato que es "buscable" y que permite así que el registro sea recuperado. Entre más puntos de acceso se proporcionen más fácilmente recuperable o accesible será un registro. Por ejemplo, el registro bibliográfico de un artículo de revista podrá ser accesible por medio de las palabras del título, de las palabras del resumen, y mediante los términos de indización que se les haya asignado (los cuales pueden ser indicativos de su contenido temático), así como por los nombres de autores, título de la publicación periódica y/o ISSN, y quizá la fecha de la publicación. Esto implica, naturalmente, que todos estos campos del registro (título, resumen, término de indización, autor, título de la publicación periódica, fecha) deben ser campos de búsqueda.

Entre más grande es la base de datos, mayor número de puntos de acceso se requerirán para permitir que las búsquedas sean más selectivas y así impedir que los usuarios recuperen un número excesivo de elementos de información. Por ejemplo, los registros que suelen encontrarse en el catálogo de una biblioteca proporcionan puntos de acceso temático muy limitados: dos o tres encabezamientos de materia, palabras clave del título y quizá, el número de clasificación. Puesto que las palabras del título, los encabezamientos de materia y los números de clasificación con frecuencia se duplican, el número de puntos de acceso temático únicos pueden ser del rango de 3-4 en promedio. Esto no es necesariamente un problema si el catálogo cubre algunos miles de piezas, pero no es adecuado para una colección que alcanza millones (el catálogo de una biblioteca muy grande o de un consorcio de bibliotecas), porque cada búsqueda tenderá a recuperar un número excesivo de registros y habrá muy pocos puntos de acceso temático para permitir que el usuario restrinja la búsqueda posteriormente.

Normalización y consistencia

Los usuarios de una base de datos pueden confundirse si los contenidos de los registros se presentan en aparente desorden. Los datos deben organizarse de manera consistente, siguiendo alguna norma. Por ejemplo, las referencias bibliográficas deben presentarse en un formato consistente, adhiriéndose a una norma aceptada, como la de la American Psychological Association.

Para las bases de datos en que los registros son indizados empleando descriptores temáticos, es muy conveniente que sean normalizados mediante el uso de algún vocabulario controlado, tal como un tesauro o lista de encabezamientos de materia. El valor de una base de datos puede incremen-

tarse de manera considerable a través de una indización temática cuidadosamente controlada. Si bien la aptitud para buscar en texto libre (palabras/frases en títulos y resúmenes) reduce la necesidad de que una persona asigne los términos de indización, tales términos tienen el valor de reducir las ambigüedades del texto libre y de facilitar la conducción de búsquedas relativamente más amplias.

Especificidad de los puntos de acceso

Los términos empleados para el acceso temático deben ser suficientemente específicos para permitir que las búsquedas se realicen con el detalle apropiado a los intereses de los usuarios de la base de datos. Por ejemplo, si un usuario busca información sobre el cultivo de naranjas, puede no querer conocer todo lo relativo al cultivo de cítricos, y tampoco todo lo referente al cultivo de la fruta en general. Para este usuario, el vocabulario deberá ser más específico que "fruta", o aun que "frutos cítricos"; se requiere el término preciso "naranjas". Está claro que el productor de una base de datos deberá estar bien enterado acerca de las necesidades e intereses de quienes probablemente la empleen.

El requisito de especificidad se aplica a las bases de datos de todo tipo. Los términos empleados para indizar recopilaciones estadísticas deberán tener la especificidad apropiada a las necesidades de búsqueda de los usuarios en estas fuentes, e igualmente para las bases de datos de imágenes. En el caso de las bases de datos bibliográficas, el requisito de especificidad se aplica más a los términos temáticos controlados, puesto que las palabras del texto (en títulos y resúmenes), usualmente estarán al nivel de especificidad de los contenidos de la publicación misma. Para una base de datos que se puede consultar tanto por palabras del texto, como por términos controlados, el requisito de especificidad de los últimos es menos importante, en la medida en que los usuarios puedan combinar los términos controlados más generales con las palabras del texto, para lograr la precisión deseada.

Factores de Calidad

Los productores de una base de datos son los responsables de asegurar que los datos incluidos en ella sean lo más exactos posible. Desafortunadamente, el control de la calidad de los datos rara vez recibe alta prioridad por parte de los productores de las bases de datos. Por consiguiente, se presentan altos porcentajes de error en muchas de las bases de datos, aun en las principales bases de datos internacionales; por ejemplo, imprecisiones en las referencias bibliográficas (números incorrectos del volumen, o del número que corresponde a la publicación periódica, o referencia incorrecta a las páginas de los artículos) pueden provocar que los usuarios desperdicien su tiempo, se frustren, o se les causen gastos innecesarios. Los errores tipográficos no son cuestiones triviales. Un usuario puede desconfiar de una base de datos en la que ocurren errores de tipografía. Además, semejantes errores pueden tener efectos significativos sobre los resultados de la búsqueda, haciendo que algunos registros no sean recuperados cuando debieran serlo, y que

otros lo sean, cuando no debieran serlo. Es notable que Cahn (1994) haya reportado hasta 46 errores diferentes de ortografía de un mismo término en 14 de las principales bases de datos que examinó; obviamente, es muy reducida la oportunidad de que quienes hicieron dicha búsqueda pudieran encontrar todos los registros relacionados con este término.

Deben hacerse todos los esfuerzos posibles para lograr la calidad de los resúmenes u otras narrativas descriptivas, en particular para asegurar que reflejen con exactitud el contenido, o el carácter de las piezas a las que se refieren. Tenopir y Jacsó (1993) ofrecen una discusión útil y concisa sobre la calidad de los resúmenes.

La exactitud de los datos no es una cuestión que deba tomarse a la ligera. Bajo ciertas circunstancias, los editores pueden ser demandados ante los tribunales si la información incorrecta contenida en una base de datos da por resultado alguna pérdida para un individuo o institución (O'Neill y Vizine-Goetz, 1988). Los productores de bases de datos deberán animar a los usuarios a reportar los errores que descubran y estar dispuestos a comprometerse a corregir cualesquier error identificado por esta vía, en las actualizaciones que se vayan realizando de las respectivas bases de datos.

Documentación

Otra responsabilidad de quien produce la base de datos es la preparación de un manual impreso y/o accesible en terminal, que sea adecuado para los usuarios. El manual debe dar una descripción completa y exacta de los contenidos de la base de datos en su totalidad, así como del contenido y formato de los registros individuales y debe contener explicaciones claras de cómo se puede usar la base de datos, incluyendo muestras de las formas de búsqueda. Deberá también proporcionar información sobre otras cuestiones que puedan interesar a los usuarios, tales como la frecuencia con que se actualiza la base de datos.

Conclusiones

Para los fines de estos lineamientos las bases de datos se pueden clasificar de la siguiente manera:

1. Las que se desarrollan para el uso interno de una sola organización y las que se destinan a una distribución o empleo más amplio.
2. Las que se destinan a una distribución en CD-ROM o en otro formato electrónico y las que se cargan en línea para su acceso en red.

Los criterios de calidad y utilidad se aplican por igual a todos los tipos de bases de datos. Los criterios relacionados con los contenidos se aplican más a las bases de datos para distribución general que a las que se destinan a uso interno. Sin embargo, algunos de los criterios relacionados con el contenido (los que se refieren al pronóstico, vigencia y continuidad de cobertura) son tan relevantes para una base de datos de uso interno como para cualquiera otra.

Si bien, la mayoría de los criterios contenidos en estos lineamientos se aplica tanto a las bases de datos accesibles en red, como a aquellas que físicamente se distribuyen (v.g., en CD-ROM), algunos son menos importantes en el ambiente de redes. Por ejemplo, los aspectos de economía de masa crítica son poco aplicados, aunque la proliferación constante de pequeñas bases de datos especializadas dentro de Internet en las que es más problable que decrezca la accesibilidad a la información a que se incremente.

La tecnología de CD-ROM y la de Internet han contribuido de manera importante a la facilidad con que una organización puede crear una base de datos para obtener una distribución más amplia. Pero la relativa facilidad y economía de producción por sí mismas no justifican la construcción de una base de datos. Cuando una organización decide construir una base de datos, debe estar dispuesta a asumir nuevas responsabilidades; evidentemente que deberá aprender muchísimo sobre los usuarios potenciales de la base de datos y sus necesidades. Este conocimiento es esencial para el desarrollo de una base de datos "utilizable", tal como se define en estos lineamientos.

El productor de una base de datos también asume otras responsabilidades adicionales —señaladas en estos lineamientos— para la continuidad y vigencia de la base de datos, para el control de calidad y para proveer de opciones adecuadas de acceso, búsqueda y salida.

El requerimiento de la masa crítica implica que las organizaciones que se propongan producir una nueva base de datos deberán considerar la realización de algún convenio para compartir recursos con instituciones similares, a fin de crear una base de datos de mayor valor potencial para más personas, y que sea una herramienta más para el acceso a la información, en función de su costo-beneficio.

Referencias

Cahn, P. "Testing database quality". *Database*, 17(1), 1994, 23-30

Jacsó, P. *CD-ROM Software, Dataware and Hardware: Evaluation, selection, and Installation*. Englewood, CO, Libraries Unlimited, 1992.

O'Neill, E. and Vizine-Goez, D. "Quality control in outline databases". *Annual Review of Information Science and Technology*, 23, 1988, 125-156.

Tenopir, C. and Jacsó, P. "Quality of abstracts". *Online*. 17(3), 1993. 44-55.

Lineamientos para la recopilación de bases de datos La edición consta de 500 ejemplares y estuvo a cargo de Carlos Ceballos Sosa. Corrección de estilo y revisión de pruebas, Blanca Furber Chicas / Centro Universitario de Investigaciones Bibliotecológicas / UNAM. Fue impreso en papel couché mate de 100 gr., en los talleres de MAR Impresiones, ubicados en Calle Fresnos M-6, L-10, Col. Bosques del Pedregal, México, D.F. Se terminó de imprimir en el mes de marzo de 1996.

fa
p■

FOLLETOS DE APOYO
PROFESIONAL