Revista de Historia de la Psicología
1994, vol. 15, nº 3-4, pp.369-372

PROBLEMAS METODOLÓGICOS DEL DESARROLLO DE BASES DOCUMENTALES BIBLIOGRÁFICAS DIRIGIDAS AL TRATAMIENTO BIBLIOMÉTRICO

Jaime Sanmartín
Dpto. de Metodología Psicobiología y Psicología Social
Francisco Alonso
Dpto. de Psicología Básica
Facultad de Psicología. Universidad de Valencia.

En este trabajo se pretende discutir parte de la problemática metodológica que surge al plantear, como tarea de I+D (investigación y desarrollo), el desarrollo de un Sistema de Gestión de Base de Datos (SGBD; Bishop, 1992) para la manipulación de referencias bibliográficas orientada al tratamiento bibliométrico. En este sentido, se pueden considerar tres grandes aspectos:

1) La estructura de los datos que subyace a la información que se quiere manipular y las relaciones que mantienen estos datos entre sí (estructura).

2) El tipo de operaciones y manipulaciones que deben realizarse con los datos; el tipo de resultados que se pretenden obtener con el SGBD (output).

3) Los procedimientos por los que se va a alimentar de información a la base de datos y las características de esta (input).

Aunque todos ellos son igualmente importantes, el objetivo de esta exposición se centrará solamente sobre el primero de ellos. Además, es precisamente la estructura de datos uno de los aspectos que más condiciona en la mayoría de los casos el tipo de base de datos que se necesita.

Una de las fuentes fundamentales de problemas, sino la que más, es que la información de referencias no sigue un patrón suficientemente estandarizado que permita una estructura consistente de campos. Los problemas fundamentales surgen de dos aspectos:

1) Muchos de los campos no son simples, sino múltiples, cual es el caso de los autores o los descriptores, que pueden ser de uno a varios para cada una de las referencias.

2) La estructura de campos cambia según el tipo de documento a que se hace referencia. Mientras que el campo que se refiere al autor o autores principales es común a casi todos los documentos, en cierto tipo de documentos pueden aparecer otros autores secundarios, cual es el caso del editor del texto en un capítulo de libro.

Para discutir esta problemática vamos a considerar diversa estrategias implementadas en dos tipos de productos informáticos:

A) BASES DE DATOS DE MANIPULACIÓN DE REFERENCIAS BIBLIOGRÁFICAS.

En el mercado ya existen desde hace tiempo diversos productos de este tipo, que no hay que confundir con los sistemas de gestión de bibliotecas, orientadas hacia otro tipo de problemas. Un caso interesante de producto es el programa EndNote Plus (Niles & Associates, Inc. 1990). Esta base de datos sigue un modelo no relacional, de gestión de una sola base de datos, con una estructura de registros simple, en la que los campos no son de longitud fija, aunque con un límite de tamaño.

Aunque su orientación se dirige a la facilitación de bases de referencia personales que permiten la recopilación automática de las referencias citadas en procesadores de texto, la simplicidad de su estructura de datos llama la atención. Si bien no solventa el problema de las informaciones múltiples en un mismo campo (p.e. el caso de los autores), para resolver el caso de la distinta estructura de campos según el tipo de referencia, utiliza un número determinado de campos genéricos que adquieren distinto significado según el tipo de referencias. Naturalmente, utiliza otro campo más para registrar cuál es su tipo. Esto lleva a compartir, por ejemplo, un mismo campo para registrar la información del título de la revista en un artículo y el título de un libro en un capítulo de libro. La denominación genérica de dicho campo sería la de *título secundario*. Otro caso sería el de *autor secundario*, que registraría el editor en un capítulo de libro o el director de una colección en un libro. Esta estrategia resulta enormemente simple y operativa para el almacenamiento no complejo de información y búsquedas, pero en el fondo está mezclando informaciones de significado muy distinto que si se requieren operaciones de resumen de información más allá de las simples recuperaciones puntual resultarían confundentes. De hecho, el propio programa no está concebido para mucho más.

B) BASES DE DATOS DOCUMENTALES DE REFERENCIAS.

Con esta nomenclatura nos estamos refiriendo a las bases de distribución y recuperación de información de referencias que proporcionan servicios documentales como DIALOG o empresas como Silver Platter. Estas empresas ofrecen información documental que es recuperable a través de programas realizados *ad hoc* para tal fin.

Estos programas se basan en estrategias típicas de bases documentales, pero que no hay que confundir con lo que actualmente se entiende como Sistemas de Gestión de Bases Documentales. En un sistema de este estilo, se pretende dar cabida a informaciones altamente heterogéneas en cuanto a su estructura (p.e. pueden almacenar igualmente imágenes como radiografías digitalizadas, junto con textos digitalizados o en códigos ASCII, información multimedia,....). En cualquier caso, toda aquella información codificada en ASCII suele ser accesible debido a que se indexa en su totalidad o permite procesos más lentos de búsqueda sin indexación. Aspecto este especialmente interesante puesto que permite buscar cualquier tipo de información aunque sea una palabra suelta o parte de ésta dentro del texto de un campo.

En este caso se han analizado varias bases de las que da servicio DIALOG, como son el caso de MEDLINE, ERIC, PsycINFO y PsycLit (como PsycINFO pero en CD-ROM y servida por Sylver Platter). Todas ellas comparten una estructura muy similar de campos delimitados por una serie de identificadores (códigos alfabéticos) más o menos comunes entre las distintas bases.

La estrategia utilizada en estas bases es distinta a la anterior, puesto que se utilizan delimitadores distintos para cada campo. En este caso cada campo recoge la información que le es propia, pero con una doble dificultad. Por un lado las informaciones múltiples aparecen en un mismo campo y seguidas, utilizando, en todo caso, una separación por comas u otros delimitadores. este es el caso de, por ejemplo, los autores y los descriptores. Por otro lado, en muchas ocasiones, combinan en un

mismo campo informaciones que convendría disponer por separado. tal es el caso del campo donde se recoge el título de la revista, en el que se incluye el volumen, el número e incluso el año. Otro caso lo tenemos con los capítulos en libros editados, en los que en el campo de libro se recoge conjuntamente el título del libro, el editor e incluso la editorial y las páginas.

Por otro lado también se utilizan distintas estrategias, según la base concreta, en función del tipo de referencia. Por ejemplo, PsycLit diferencia dos bases de información distinta para artículos y para libros. En esta última recoge además capítulos de libros y libros editados. Otras bases, como es el caso de ERIC, incluye además otro tipo de referencias que incluyen informes, trabajos no publicados, comunicaciones a congresos, descripción de programas de ordenador, y toda una serie de materiales de difícil clasificación. Parece que queda claro que, si bien no plantean problemas desde el punto de vista de las búsquedas puntuales, tal desestructuración no hace sino dificultar en gran manera posibles operaciones de recuento y clasificación.

UN PROTOTIPO OPERATIVO

En muchos casos la resolución de problemas de manipulación de información requieren un proceso de desarrollo paralelo al de investigación. En este caso se ha partido de un prototipo anterior (Valero, Molina y Sanmartín, 1992) desarrollado sobre FileMaker Pro v.2.0 (Claris™, 1992), del cual existen versiones compatibles para ordenadores MS-DOS y Macintosh. Esta sistema de base de datos es de tipo pseudorelacional que permite facilidades documentales y de *scripting* (posibilidad de realizar *macros* o agrupaciones de instrucciones automatizadas denominadas *guiones*).

Este prototipo, denominado BibRef, comparte las mismas características respecto a las facilidades de búsqueda de información que el prototipo anterior. Consta de un archivo principal, al que se le han implementado *scripts* para el recuento de referencias según años, revistas e instituciones, además de las combinaciones de años por instituciones y años por revistas. Además se ha incorporado otro archivo subsidiario, denominado BibAuthor que se utiliza para recoger la información de los autores y establecer recuentos en función de éstos (autores por revistas y por años)

Aunque el prototipo es operativo y viene utilizándose en diversos trabajos, su objetivo fundamental es servir de sistema de prueba para contrastar la viabilidad de las distintas alternativas de diseño informático.

CONCLUSIONES

Desde el análisis anterior y después de la experiencia de desarrollo y utilización del prototipo BibRef, parece oportuno extraer las siguientes conclusiones respecto a los requisitos que debe cumplir un SGBD apropiado para las bases de referencias desde el punto de vista de su estructura:

1) La necesidad de una estructura clara de información apunta hacia una un SGBD de tipo relacional, donde cada campo es definido con claridad en cuanto a su contenido y en cuanto a su estructura simple o múltiple. Además, la posibilidad de establecer distintas bases de información relacionada permitiría recoger la casuística de dependencia entre distintos campos. Este es el caso de las relaciones que se establecen entre, por ejemplo, referencias y revistas, en las que por cada revista se encuentran en relación una serie de referencias.

De todas formas, otra tipo de información, cual es el caso del título del trabajo o el abstract requiere un tratamiento de tipo documental. En este caso lo que se requiere

son facilidades amplias de búsqueda de información en base a palabras simples o incluso partes de éstas. Esto conduce a tener que concluir que tal SGBD debería ser de tipo mixto, combinando una estructura relacional, por una parte, con facilidades de tipo documental, por otra.

2) Si bien simplificaría mucho los procesos de desarrollo y estructuración la consideración de bases de datos distintas según el tipo de referencia (para artículo, para libros, para congresos,...), se complicaría mucho el trabajo si se quisiera hacer estudios bibliométricos considerando varios tipos de referencias. Por el contrario, los procesos de búsquedas de información se verían multiplicados al tener que repetir las búsquedas para cada base según el tipo de referencia. El tener una base única, además permitiría, en todo caso, realizar análisis para un solo tipo de referencia, sin problemas añadidos, cuando sea necesario.

3) El solapamiento de informaciones en los campos debería tratarse con cuidado y pormenorizadamente. En algunos casos, posiblemente sería oportuno separarlo. Por ejemplo, sería discutible el juntar en un mismo campo de *título secundario* el título de la revista del título de un libro editado para el caso del capítulo de libro. Por otro lado, si que parece razonable recoger en el mismo campo de autor, al autor o autores principales, ya sean éstos referidos a artículos, libros u otro tipo de referencias. En otros casos también estaría bastante clara ésta solución (p.e. abstract o descriptores).

BIBLIOGRAFÍA

Bishop, P. (1992). *Fundamentos de informática*. Anaya Multimedia, Madrid.

Claris™. (1992). *FileMaker Pro™, v.2.0*. Computer program.

Niles & Associates, Inc. (1990). *EndNote Plus™, v 1.0*. Computer program.

Valero, P. M.; Molina, J.G. y Sanmartín, J. (1992). Un grupo de herramientas informáticas para el análisis de conjuntos de referencias bibliográficas. *Revista de Historia de la Psicología, 13*, pp.93-103.