SEGUNDA PARTE

INDICES DE DOCUMENTOS FÍSICOS Y ELECTRÓNICOS

1. Introducción

En el apartado anterior se ha visto como un documento de cualquier tipo necesita ser conservado en lugar apropiado, de lo contrario corre el riesgo de perderse o de no encontrarse la información requerida en el momento adecuado. También se discutió la importancia del acceso a la información por parte del grupo más amplio posible de investigadores. En este apartado se reflexiona sobre las ventajas de la digitalización desde el punto de vista de la elaboración de índices personalizados a las necesidades del investigador. Así mismo se ponderan las ventajas de la rapidez que brinda el documento digitalizado.

2. ¿Qué es un índice?

El índice es un almacenamiento de información clave que facilita el acceso a los documentos[1]. En el caso de los libros el índice proporciona cierta palabras claves relacionadas con el número de página que posiblemente contenga más datos sobre el

1 Hernán Pérez de Inestrosa Sánchez; Modelos Avanzados de Bases de Datos Disponible: http://alarcos.inf-cr.uclm.es/doc/bbddavanzadas/documentales.pdf

tema investigado. Existen varios tipos de índice y el que por regla general no falta, es el índice de contenido (en inglés *table of contents* o *contents*), el cual, muestra la división y subdivisión de los capítulos en los que el libro está organizado. Otros tipos de índices son: analítico, onomástico, temático, geográfico, de fotografías y de figuras o gráficas. Estos índices proveen al investigador con formas alternativas y más detalladas de los datos que se encuentran en el índice de contenido ya que se presenta información adicional sobre personas temas o países que aparecen dentro del libro y que son datos relevantes o datos clave.

El objetivo de cualquiera de ellos es facilitar la búsqueda de la información que se desea y encontrarla de manera rápida y efectiva. En el caso de los índices de fotografías o material iconográfico también se le facilita al autor proveer información sobre las fuentes utilizadas y los derechos de autor de cada una de las ilustraciones utilizadas. En el campo de la historia del arte esto es muy útil porque al lector se le facilita la consulta rápida del material iconográfico.

3. El uso y tipos de índices

Desde hace varios años la impresión de índice de contenido que articulan la división del libro en capítulos se ha vuelto reglamentaria. Con frecuencia estos índices también incluyen subdivisiones de capítulos, a menos que el autor por razones de principio deseé que la lectura de su libro no se administre de esta manera. Este es el caso de las obras de Theodor Adorno.

Los demás tipos de indexación son relativamente más recientes y han tenido un gran auge dentro de los últimos 20 años. Podemos observar que la edición en español de 1987 de *Dialéctica del iluminismo* (1944) de T. W. Adorno no presenta más que el

índice de contenido en la penúltima página (Editorial Sudamericana).[2] En cambio, la impresión de 1960 de editorial Península de *Mahler: una fisiognómica musical* (1960) del mismo autor, muestra al final sus índices de conceptos, de nombres propios, y de las composiciones citadas además del <<sumario>> o índice de contenido.[3] Es probable que la casa editora haya sido la encargada de anexar los índices, pues esta segunda edición es de 1999.

De igual manera podemos observar índice alfabético de datos biográficos de músicos, musicólogos, intérpretes y obras en la *Enciclopedia de la Música*[4] traducida por Otto Mayer-Serra, editorial Atlante. En el área de música podemos apreciar otros tipos de indexación: obras, autor, términos (que hay veces que no se muestran como índices, sino como glosario).

La aparición de los índices depende de varios factores: la casa editorial, la calidad, público a quien va dirigido el texto, etc. El índice de contenido no era suficiente para los libros de consulta, donde el usuario busca información específica y no necesariamente va a leer el total del libro. De esta manera surgen formas que complementan y empiezan a formar un sistema de búsqueda. Pero en el caso de los libros que no son técnicos, sino que abordan las facetas de la literatura como el ensayo, el cuento y la novela, la indexación no es necesaria —o incluso deseable— pues su propósito es que la obra se lea de principio a fin[5].

El factor económico también entra en juego y es más complicado pues una vez que se ha establecido el público a para quienes se hizo la obra, se sigue con el diseño de

2 T. W. Adorno Dialéctica del iluminismo, Editorial Sudamericana, Buenos Aires, 1960.
3 T. W. Adorno Mahler: una fisiognómica musical, Editorial Península, Barcelona, 1999.
4 Otto Mayer- Serra tr., Enciclopedia de la Música. Editorial Atlante, México, año?
5 Entrevista a Mtro. Arturo Arrieta. 06/08/04

la edición: es en este momento cuando se agregan los índices, los cuales suben los costos –junto con las imágenes, partituras, diagramas–. En este paso se determina el tipo de papel, pasta y otros detalles. En el caso de hacer un estudio de sondeo de mercado –con el que también el producto sube su precio– para observar si el público meta tiene intención de comprarlo con la mejor calidad o con la conveniencia económica, se observa a la vez si el libro será vendido y se recuperará la inversión. Por lo tanto, el libro que ofrece buenos índices y mantiene el uso de calidad óptima de papel, será un mejor producto pero también tendrá un costo elevado[6]. Antes de que los procesos de impresión se realizaran en la computadora la elaboración de índices añadía tiempo al investigador, o costos por los salarios de ayudantes de investigación que debían checar en cada página la aparición de ciertos términos. El margen en estos casos es mucho más amplio y es frecuente encontrar que ciertas apariciones de un término no aparecen registradas en el índice.

4. Estrategias para realizar índices

En la realización de los índices se deben seleccionar palabras clave de acuerdo a un público determinado que implica uno o varios tipos de lector. Un profesor, un estudiante o un lector aficionado no siempre buscan lo mismo en un libro, por ejemplo, de historia de la música. El estudiante puede estar simplemente interesado en una fecha o una obra mientras que el profesor podría buscar si el libro en cuestión trabaja con ciertos sistemas analíticos. Por ejemplo, la aparición de términos como “mixtura” o “prolongación” vienen de un sistema especifico de análisis Schenkeriano. En un libro de Schoenberg se podría tratar de ver si se ha realizado un análisis tonal, o un

6 Ibid.

determinado tipo de análisis dodecafónico o atonal y esto estaría indicado por la presencia de nombres como Babbitt, Morris, o de términos como "serie", o "combinatorialidad".

Los índices digitales no se realizan manualmente como en los libros de hace 50 años sino que hoy en día se realizan por medio de estadísticas y de un análisis léxico donde se extraen las palabras de un documento sin guiones, signos, barras u operadores matemáticos. No se utilizan textos largos. Las palabras clave son las que representan el contenido de un documento y deben tener expresión satisfactoria o palabra cercana sintáctica y semánticamente. La lista de palabras que no son buenos términos de indexación se utilizan para filtrar el análisis léxico –"stoplist" – . Posteriormente se deben convertir todas las palabras parecidas a una forma común para ganar eficiencia y eficacia. Se debe hacer una agrupación de las palabras con la misma raíz –"stemming". De todo esto se obtiene el "vocabulario" para realizar el índice[7]. Todo esto se aplica para almacenar un documento o una colección de documentos. El cuadro 1 muestra el almacenamiento de datos de un solo documento mientras que el cuadro 2 muestra el almacenamiento de datos de una colección de documentos

Tabla 2.1: almacenamiento de datos de un solo documento

7 Hernán Pérez de Inestrosa Sánchez; Modelos Avanzados de Bases de Datos Disponible: http://alarcos.inf-cr.uclm.es/doc/bbddavanzadas/documentales.pdf

Término 1	---------->	Docto. 1	2, 56, 78...
Término 2	---------->	Docto. 5	46, 58, 87...
Termino 3	---------->	Docto. 4	5, 15...
Término 4	---------->	Docto. 8	4, 25, 32...
Término 5	---------->	Docto. 2	32, 55...
Término 6	---------->	Docto. 4	65, 76...
...	---------->	...	...

Tabla 2.2: almacenamiento de datos de una colección de documentos

Término 1	---------->	Docto. 1	2, 56, 78...	Docto. 4	5, 26, 27...	...
Término 2	---------->	Docto. 5	46, 58, 87...	Docto. 6	2, 4, 10...	...
Termino 3	---------->	Docto. 4	5, 15...	Docto. 6	48, 59...	...
Término 4	---------->	Docto. 8	4, 25, 32...	Docto. 9	60, 71...	...
Término 5	---------->	Docto. 2	32, 55...	Docto. 3	46, 51, 53...	...
Término 6	---------->	Docto. 4	65, 76...	Docto. 5	2, 3...	...
...	---------->	...	...	...	...	...

De acuerdo con el lector los índices también contienen palabras clave que no son frecuentes pero sí importantes. Por ejemplo, en un libro de historia de la música general el nombre de Richard Strauss podría aparecer una sola vez pero este nombre debe de estar en el índice de compositores o de palabras importantes. En este tipo de libros se recomienda incluir los nombres de todos los compositores que aparecen sin importar el dato estadístico de a quién se menciona más. También es importante diferenciar los nombres de compositores o personas parecidas y no solo incluir los apellidos (Richard Strauss, Johann Strauss padre, Johann Strauss hijo) o Silvestre Revueltas compositor y Silvestre Revueltas sobrino (el hijo de Fermín Revueltas). Este caso específico se

encuentra en el libro de Naranjo en Flor de José Angel Leyva cuya publicación por Juan Pablos Editor no tiene índice y la simple búsqueda electrónica arrojaría una estadística incorrecta puesto que se trata de dos personas distintas.[8]

En el caso de un documento digitalizado el editor puede incluir diversas formas de administración del material que llevan al usuario a lugares específicos del texto en forma expedita. También se puede dar una lista de términos cuya relevancia es obvia al editor. Lo más importante del documento digitalizado es que el estudioso puede realizar su propia búsqueda de manera rápida y exhaustiva. Esto puede facilitar el proceso de investigación como se verá en el siguiente capítulo donde se sugieren ciertas alternativas para la consulta de los textos digitalizados que se presentan en esta tesis. Antes de ello es necesario aclarar algunos puntos relativos al almacenamiento de datos.

El conjunto de información estructurada en registros y almacenada en un soporte electrónico legible desde una computadora es una base de datos. Nos referimos a una base de datos documental cuando a cada registro corresponde un documento del tipo que sea: una publicación impresa, documental audiovisual, documento gráfico o sonoro, documento de archivo o documento electrónico.[8] Específicamente vamos a tratar documento de texto. Esto no implica la inexistencia de fotografías e imágenes, pero en su gran mayoría, el documento es de texto. En el caso de archivo fotográfico o de video, sería documento gráfico o audiovisual.

La base de datos de texto completo debe contener la información del título, autor, fecha, etc., además del texto completo. Los archivos electrónicos de imágenes serán las referencias digitalizadas en el correspondiente formato para imagen. La base de datos

8 José Angel Leyva, El Naranjoen flor. Homenaje a los Revueltas¨Juan Pablos Editor, Instituto Municipal de Arte y la Cultura , 1999.

referencial nos da la información fundamental para describir y permitir la localización del documento original (la referencia bibliográfica).

Para satisfacer las necesidades de información del usuario, la colección de textos –base de datos documental– contará con indexación para la facilidad y acceso y deberá tener un sistema de recuperación y almacenamiento. El proceso de la recuperación y almacenamiento de la información lo estudia una parte de la informática, ésta centra su atención en una colección de documentos escritos, no de datos. Este proceso, expresado en el lenguaje natural, puede satisfacer una necesidad de información al usuario.[9]

En la década de 1960 se utilizó la máquina Memex de Vannervar Bus, el cual es un dispositivo en el que se pueden guardar libros, comunicaciones, registros, etc., y pueden ser consultados con rapidez y flexibilidad por medio de procesos mecánicos. Durante los últimos 20 años, como ya se había mencionado, hubo gran desarrollo por el auge en la indexación de textos y la búsqueda efectiva de documentos. Esto fue a todos los niveles: en la planeación de un libro y en un documento electrónico. Hacia la década de 1990 surge el World Wide Web, repositorio universal de conocimiento donde las búsquedas son mucho más complejas. El consorcio World Wide Web se encarga de desarrollar las tecnologías interoperables –especificaciones, pautas, software y herramientas– para conducir la tela a su capacidad máxima. Este es un foro para la información, el comercio, la comunicación y la comprensión colectiva[10] –dicho sea de paso, el dominio www. etc., surgió de sus siglas. Actualmente se pueden ver varios registros con www que corresponden más a la tendencia de moda que posteriormente

9 Ibid.
10 http://www.w3.org/
Adquirido: 20/08/04

surgió.[11] Al presente, se trata de abarcar, modelar, clasificar y categorizar documentos, arquitectura de sistemas, interfaz de usuario, visualización de datos, filtrado de lenguajes, entre otros.[12]

Una de las herramientas más rápidas para la aplicación, el uso correcto de la indexación y elaborar consultas rápidas es el sevidor de bases de datos MySQL. Este término significa "Lenguaje Estructurado de Consultas,"[13] aunque no está bien definida la razón del prefijo "My".[14] La "Base de Datos" es una colección de información (datos variados) bien estructurada y almacenada. La base de datos de MySQL tiene un sistema relacional, es decir, no almacena los datos en un solo lugar, sino que son almacenados en tablas separadas y enlazadas para definir relaciones y combinaciones dando mayor velocidad y flexibilidad.[15]

A pesar de que esta es una de las herramientas más rápidas, y por lo tanto, de las más importantes, los nuevos desarrolladores no la tienen en uso por desconocimiento o por falta de comprensión. La mejor forma de optimización de una Base de Datos consiste en un buen diseño inicial, tanto lógico como físico. El objetivo es minimizar el tiempo de respuesta para cada petición y tener al máximo el rendimiento de todo el sistema disminuyendo el tráfico de red, el acceso a disco y el tiempo de gasto de energía del ordenador. La propuesta de los planteamientos de optimización deben ser propuestos durante el ciclo del desarrollo, y aún implementado ya el sistema, pues las mejoras en el rendimiento pueden ser previstas desde el desarrollo. Durante dicho desarrollo de la aplicación de la base de datos, se debe apreciar la eficiencia además del

11 Luis Marín Fdz. Entrevista.
12 Ibid.
13 Disponible: http://www.trucostecnicos.com/webmasters/listar.php?op=ver&id_co=51
14 Disponible: http://www.mysql-hispano.org/index.php?m=read&id=259
15 Disponible: http://www.trucostecnicos.com/webmasters/listar.php?op=ver&id_co=51

funcionamientos de las consultas con el fin de ofrecer un servicio rápido. Al utilizar la indexación de los documentos, hay que evitar debilidades en la base de datos tales como: superar el tamaño máximo del fichero –en caso de tener muchos índices– y observar que no decaiga la velocidad de las operaciones de insertar, borrar, así como las actualizaciones de valores en las columnas indexadas para obtener índices ágiles. Es importante hacer resaltar el uso preferentemente de índices cortos pero específicos, ahorran espacio y probablemente ayude a la rapidez del servicio. Tampoco se debe abusar del uso de los índices.[16]

Un documento o una colección de documentos necesita un enlace, el cual, al seleccionar y pulsarlo, podrá abrirse para su lectura e incluso este proceso servirá para buscar dentro de una biblioteca. Dentro de este último caso, el formulario que se abre es de búsqueda Avanzada. Sólo aparecerán los documentos ya indexados, los que no cumplan con éste proceso no se incluirá en ninguna búsqueda. Nos referimos ya al índice de búsqueda, que es una copia comprimida de un documentos o conjunto de documentos. Evitando la revisión de "longitud completa" del documento, la búsqueda con índices simplifica el proceso. El servidor de documentación es un sistema que contiene un conjuntos de documentos y posee instalado y configurado el servicio de biblioteca de documentación, el cual puede ser implementado utilizando un sistema local o un sistema remoto en la red.[17]

Dentro de la biblioteca de documentación no se muestran todos los archivos:

> "La biblioteca de documentación no muestra todos los documentos que se encuentran en el servidor de bibliotecas. Sólo

16 Andrés Javier Pulido Benal. Optimización de consultas en MySQL
Disponible: http://www.mmlabx.ua.es/mysql-postgres.html
17 http://publibn.boulder.ibm.com/doc_link/es_ES/docsearch/docsearch_help.html

> muestra los documentos que están registrados con la biblioteca. Esto permite que el administrador de bibliotecas restrinja los documentos que están visibles en la biblioteca. Además de registrar un documento o un conjunto de documentos, el administrador de bibliotecas debe crear un índice del documento o del conjunto de documentos, de forma que se pueda buscar en ellos dentro de la biblioteca."[18]

Para evaluar la calidad del proceso, se debe plantear el objetivo y motivación del trabajo, su alcance y su contribución. Por último, los resultados principales y su aplicación. De está forma obtendremos una evaluación objetiva.

18 Ibid.