Sistemas de gestión documental y bases de datos documentales

Ernest Abadal
Lluís Codina

PID_00201456

UOC
Universitat Oberta de Catalunya

Índice

Introducción

Las bases de datos son la mejor tecnología de que disponemos en la actualidad para gestionar información, ya que es el único sistema que permite procesar la información de una forma, a la vez, segura, rápida y eficaz. De hecho, existen otras tecnologías basadas en ordenadores para gestionar información, como editores de texto, programas de hojas de cálculo, gestores de ficheros, navegadores de Internet, etc. Pero solamente las bases de datos permiten acceder a la información selectivamente, mostrarla de forma diferente a diferentes grupos de usuarios, explotarla de forma diferente si cambian los objetivos, etc. Todo ello, en el marco de una relativa seguridad y confidencialidad, tanto frente a accesos maliciosos como frente a errores involuntarios.

En este contexto, las bases de datos documentales cumplen una función de particular importancia. Se trata de un tipo de producto pensados para tratar, no tanto con datos, sino con información cognitiva o conocimiento. Para explicar esta función es necesario tener en cuenta que, cuando hablamos de información podemos estar pensando en los datos de una factura (quién la emite, su importe, quién debe abonarla, etc.) o de una tesis doctoral, para mencionar tan solo dos extremos.

En el apartado "Algunos conceptos básicos" del módulo "Sistemas de bases de datos" hay una discusión sobre los conceptos *datos*, *información*, *conocimiento*.

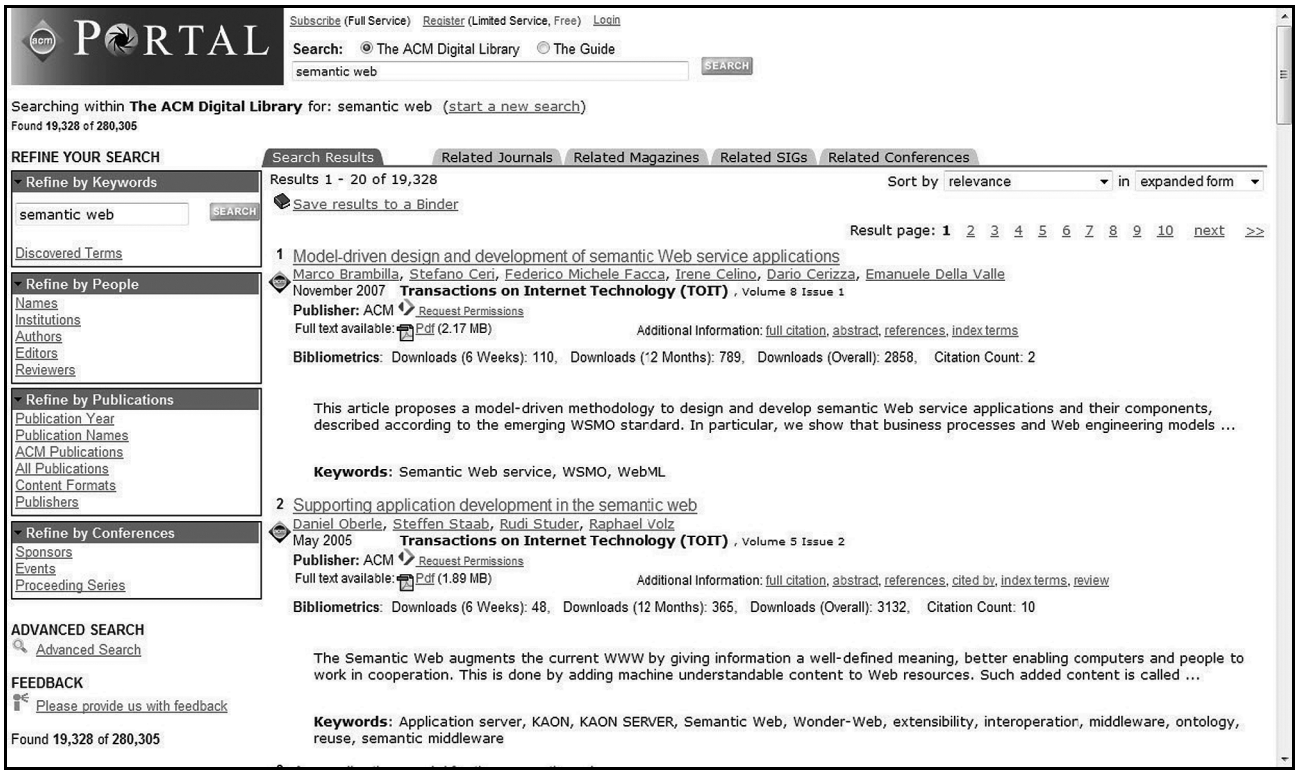
En el primer ejemplo estamos hablando de información administrativa, mientras que en el segundo nos referimos a información cognitiva, de conocimiento expresado y registrado en un documento (en este caso una tesis). En el primer documento (la factura) hay datos númericos que son fáciles de representar, por ejemplo, en forma de tabla con valores atómicos (cada celda un valor único). También hay texto, pero en forma de datos factuales breves y compactos (un nombre propio, una dirección, un nombre de producto, etc). En el segundo documento, pueden encontrarse datos factuales, pero sobre todo hay texto en forma de discurso razonado, exposición de teorías, razonamientos inductivos o deductivos, etc. El contenido de este segundo tipo de documentos no puede ser reducido a una tabla con valores atómicos.

Este es uno de los motivos por los que los sistemas de gestión de bases de datos relacionales, basados en tablas con valores atómicos, no puedan gestionar bien documentos cognitivos como los mencionados. Otros ejemplos de tales documentos son los artículos de revista, los informes técnicos o científicos de cualquier tipo, las informaciones periodísticas, la documentación de mantenimiento de equipos, las patentes, etc.

De hecho, la tecnología que fundamenta las bases de datos documentales es la única que puede dar soporte a aplicaciones tan importantes como las bases de datos científicas o académicas (*Web of Knowledge*, *Scopus*, etc.), los buscado-

res de Internet, las hemerotecas y repositorios digitales de la web, los buscadores internos de sitios e Intranets, los catálogos de bibliotecas, los portales de revistas, las bases de datos de patentes, de tesis doctorales, etc.

Figura 1. Página de resultados de una típica base de datos documental de tipo académico (*ACM*, en este caso)



Si echamos la vista atrás, podemos ver que, históricamente, las primeras bases de datos documentales surgen a finales de los años sesenta en los EUA, y aparecen vinculadas tanto al mundo de la información periodística como de la información científico-técnica. Desde entonces no han dejado de extenderse a otros terrenos y actividades sociales como hemos intentado explicar anteriormente.

Con objeto de contribuir a fijar los conceptos básicos que vamos a manejar en los siguientes apartados, introduciremos la dicotomía base de datos versus sistema de gestión de bases de datos, que no siempre se diferencia con claridad aunque se trata de una distinción de gran importancia. En primer lugar, hace falta recordar que una *base de datos* es un conjunto o colección de datos estructurados almacenados digitalmente, mientras que un *sistema de gestión de bases de datos* (SGBD) es el programa que permite la creación, el mantenimiento y la explotación de la base de datos.

Los conceptos *base de datos* y *SGBD* se definen en el módulo "Sistemas de base de datos".

A partir de aquí podemos comentar las principales características de las bases de datos.

1) Los datos están interrelacionados y estructurados siguiendo un modelo

Los modelos en la fase de diseño se tratan en el subapartado "La fase de diseño" del apartado 3.

Los datos han de poseer alguna estructuración interna, es decir, no se puede tratar de un mero depósito o almacén de información. Para ello hay que recu-

rrir a diversos modelos que ayudan a estructurar e interrelacionar los datos para facilitar la recuperación de la información.

2) Los datos están almacenados en un soporte informático

Este es otro aspecto fundamental: el contenido de una base de datos debe estar grabado en un soporte digital. De otra forma nos encontramos, por ejemplo, frente a un listado impreso.

3) Existe un programa que se ocupa de la gestión y manipulación de los datos

Los sistemas de gestión de bases de datos (SGBD) son los programas que permiten la creación, el acceso y la manipulación de las bases de datos. Sin su concurso no podría darse salida a lo que constituye el principal objetivo de una base de datos: la selección, recuperación y explotación de la información que contiene.

4) Los datos serán usados o bien por otros programas informáticos o bien por personas

En la concepción característica de la informática de gestión, las bases de datos con frecuencia no son para usuarios finales (personas), sino para dar soporte a procesos informáticos que llevan a cabo programas de ordenador. Por ejemplo, los contenidos de una base de datos de recursos humanos servirán, principalmente, para confeccionar de modo automático la nómina de cada mes. En cambio, las bases de datos de tipo documental, casi siempre están orientadas a dar servicio a usuarios finales: por ejemplo, a los usuarios de un centro de documentación o de una biblioteca.

Después de esta presentación básica, vamos a profundizar en las bases de datos documentales a partir de las diferentes operaciones que se pueden realizar con ellas:

- **Producción y administración** (apartado 1), donde nos ocuparemos fundamentalmente de la estructura y las características de los programas informáticos para crear bases de datos documentales.

- **Distribución** (apartado 2), donde introduciremos el concepto de interfaz de consulta de una base de datos documental y daremos indicaciones para evaluarlo y diseñarlo.

- **Diseño** (apartado 3), donde nos centraremos en la metodología para la creación de bases de datos documentales.

- **Evaluación** (apartado 4), donde presentaremos los indicadores fundamentales para la evaluación de una base de datos documental.

Objetivos

El estudio de este módulo permitirá conocer a fondo los contenidos siguientes:

1. Saber cuáles son las características de los sistemas de gestión de bases de datos documentales.

2. Saber qué clases de bases documentales hay.

3. Conocer cómo se puede crear una base de datos documental.

4. Saber cuál es el diseño más adecuado de una base de datos documental.

1. Producción y administración de bases de datos

El elemento fundamental para la producción y administración de bases de datos son los los sistemas de gestión de bases de datos (SGBD), programas informáticos que permiten la creación y explotación de bases de datos.

En este apartado vamos a analizar las características generales de los SGBD, su tipología (SGBDR y SGD) y nos centraremos fundamentalmente en describir la estructura, funcionamiento y mercado de los distintos tipos de SGD (SGBDD y sistemas de indexación).

1.1. Los sistemas de gestión de bases de datos (SGBD)

Para profundizar en la comprensión de los SGBD podemos tomar como referencia la siguiente definición:

> "[Un SGBD es un] conjunto coordinado de programas, procedimientos, lenguajes, etc. que suministra a los diferentes tipos de usuario los medios necesarios para describir y manipular los datos almacenados en la base de datos, garantizando su seguridad." (Miguel, 1997, pág. 38)

Como hemos apuntado en la introducción, no todos los SGBD son iguales en funciones y objetivos, y podemos distinguir entre los SGBD *relacionales* y los SGBD *documentales* (o textuales). Veamos sus diferencias:

1) Sistemas de gestión de bases de datos relacionales (SGBDR)

Suelen utilizar un modelo lógico de datos denominado *relacional*. Son programas especialmente adecuados para la gestión de información muy estructurada (datos propiamente dichos). En la concepción informática clásica, de hecho es el único tipo de sistema de gestión de bases de datos que se considera. Están muy implantados en el ámbito de la empresa para gestionar y automatizar procesos, de modo que muchas bases de datos gestionadas con SGBDR no están pensadas para ser consultadas por personas (usuarios), sino para ser usadas como parte de procesos informáticos (generar la facturación mensual, por ejemplo).

Ejemplo

Se utilizan para la gestión de volumen de ventas, sueldos o existencias de almacén, etc.

2) Sistemas de gestión documentales (SGD)

Suelen utilizar un modelo lógico denominado *textual*. Su característica común es que están concebidos para gestionar la clase de información con gran cantidad de texto de tipo discursivo y poco estructurado (desde el punto de vista informático) que es típica de los documentos cognitivos. Mientras que

Ejemplo

Se utilizan para la gestión de artículos de revistas, páginas web o reportajes fotográficos, para mencionar tres ejemplos muy dispares.

uno de los elementos fundamentales del modelo relacional son las tablas homogéneas (filas y columnas iguales), en el caso del modelo textual lo son el registro irrestricto (sin limitaciones) y los índices analíticos.

La siguiente tabla ofrece un resumen de los rasgos diferenciales fundamentales de los dos grandes tipos de SGBD considerados.

Tabla 1. Sistema relacional contra sistema documental

Tipo de sistema	Contexto	Tipo de datos	Finalidad
Relacional	Gestión administrativa, contable, etc., típica de cualquier organización pública o privada.	Estructurado y muy regular (p.e., cifras de ventas, o direcciones postales).	Gestión, administración, supervisión, planificación, etc., de empresas y todo tipo de organizaciones.
Documental	Adquisición de conocimiento y satisfacción de necesidades de información más o menos complejas.	Texto de tipo discursivo, propio de artículos de revistas, noticias de prensa, etc., o texto descriptivo para *describir* objetos multimedia: imágenes, vídeo, sonido, etc.	Estudio, investigación y adquisición de conocimientos al servicio de proyectos, procesos de enseñanza-aprendizaje, investigación, soporte a la I+D, etc.

¿Qué impide usar un sistema relacional (SGBDR) para gestión documental? En principio, nada. Es decir, nada lo impide si el volumen de información a tratar es pequeño, si no necesitamos prestaciones de control terminológico y si no necesitamos salidas en formatos bibliográficos específicos, por mencionar solamente tres tipos de prestaciones funcionales.

Veamos los dos primeros aspectos por separado que, probablemente, son los fundamentales. Los SGBDR no indizan todo el contenido de los campos de texto. Por defecto, los campos con mucha información textual o bien no se indizan o bien indizan únicamente la primera palabra de cada campo. En este contexto, si la base de datos contiene poca información y se utiliza un ordenador suficientemente rápido, una búsqueda secuencial puede imitar el uso de un índice de tipo documental. Sin embargo, en cuanto crezca la base de datos, las prestaciones del sistema se degradarán. La experiencia indica que, a partir de un determinado número de documentos, un sistema relacional difícilmente podrá gestionar con eficacia un contenido de tipo cognitivo. En cambio, el mismo sistema relacional podrá gestionar con gran eficacia millones de registros de tipo tabular (es decir, datos del estilo de direcciones, contabilidad, datos de ventas, etc.).

Indexar (elaborar un índice)

Es la acción de registrar datos ordenadamente con el fin de obtener resultados relevantes de forma más rápida en una busca de información.

En segundo lugar, por **prestaciones de control terminológico** nos referimos a la posibilidad de definir y utilizar diccionarios de palabras vacías, diccionarios de sinónimos, tesauros, etc., que controlan el resultado de la indización y facilitan la realización de búsquedas. Otras deficiencias de los sistemas relacionales con relación a la gestión documental se refieren a dificultades técnicas para definir el número óptimo de caracteres de cada campo (que es preciso prefijar de antemano), el número óptimo de campos que se utilizarán para

contener descriptores, la ausencia de herramientas para gestionar y producir bibliografías, etc.

Así, podemos retomar la pregunta anterior "¿qué impide utilizar un sistema relacional para gestión documental?" y responder ahora: todo. Todo lo impide si lo que necesitamos es gestionar el contenido de grandes volúmenes de documentos de tipo cognitivo o si necesitamos utilizar algún tipo de control terminológico para optimizar los resultados.

1.2. Sistemas de Gestión Documental (SGD)

Los sistemas de gestión documental, que en inglés reciben denominaciones como *information retrieval system, text retrieval systems*, *document retrieval system* (o *digital asset management* cuando se utilizan para documentos icónicos), son el tipo de programa especialmente adecuado para la gestión de información textual y de documentos cognitivos.

Como ya hemos indicado, en general, los SGD están concebidos para gestionar documentos de tipo científico (artículos, ponencias, tesis, etc.), técnico (informes, patentes, etc.) o cultural (artículos de prensa, fotografías, etc.). Permiten, por tanto, la gestión de fondos documentales de cualquier naturaleza, y esto incluye la gestión de cualquier colección de textos, imágenes y objetos multimedia (sonido, música, video, etc). Al modelo aproximadamente similar (pero no idéntico) que siguen la mayoría de los SGD se le suele denominar, a falta de un mejor nombre, *modelo textual*.

1.2.1. Características del modelo textual

El modelo textual o documental presenta, al menos, cinco importantes características:

1) Un modelo de registro irrestricto

En los SGD no hay restricciones previas al tipo de registro que pueden manejar. En este sentido, los modelos de registro pueden ir desde esquemas totalmente abiertos, como si se tratase de documentos de un editor de texto (por ejemplo, *askSam*), hasta modelos perfectamente articulados en campos y tipos de datos (ejemplos, *Inmagic*, *CDS/ISIS*), pasando por tipos intermedios que aportan una buena flexibilidad para trabajar con campos articulados, pero sin excesivas complicaciones (*FileMaker* o *Inmagic*). El modelo irrestricto se refiere también a la posibilidad de que en una misma base de datos puedan convivir modelos de registros distintos (*CDS/ISIS*, *Inmagic* o *Refworks*).

Figura 2. Ejemplo de un típico registro documental (en este caso, del sistema *RefWorks*)

No. de Identificación:	560
Tipo de Referencia:	Artículo de Revista Acacémica (Journal)
Tipo de fuente:	Impreso
Idioma de salida:	Inglés
Autores:	Diaz Noci,Javier
Título:	Multimedia and Reading Ways: a State of the Art
Publicación Completa:	COMUNICAR
Año de Publicación:	2009
Fecha de Publicación - Formato Libre:	OCT
Ejemplar:	33, Sp. Iss. SI
Página Inicial:	213
Otras Páginas:	219
Descriptores:	Multimedia; reading; online journalism; communication; hypertext; interactivity}
Resumen:	Multimedia is one of the less studied characteristics, probably because of the less-developed level of the digital language. Along with hypertext and interactivity, it is one of the characteristics that defines the digital edition. Those characteristics have been always studied from the point of view of production, although not so much from the point of view of reception. How do users read a digital text? The reader's participation, reading depth, different trailblazing, the relation user-interface and the conception of multimedia text as a module of a database introduce major changes in the reception of the text, which can and must be studied.
Notas:	Article}
Editorial:	GRUPO COMUNICAR
Lugar de Publicacion:	APDO CORREOS 527, HUELVA, 21080, SPAIN}
ISSN/ISBN:	1134-3478
Dirección/Afiliación	Noci, JD (Reprint Author), Univ Pompeu Fabra Barcelona, Fac Comunicac, Dept Comunicac, Barcelona, Spain. Univ Pompeu Fabra Barcelona, Fac Comunicac, Dept

2) Capacidad monobase o multibase indistintamente

Es característico de algunos SGD que únicamente puedan abrir y utilizar una sola base de datos cada vez (*askSam*). Sin embargo, cada vez son más los SGD con capacidad multibase, es decir, que pueden abrir y utilizar con más de una base de datos a la vez (*CDS/ISIS*, *Inmagic*, o *FileMaker*).

3) Índice analítico (fichero invertido)

El fichero inverso es un índice (o un conjunto de índices) compuesto por todas y cada una de las palabras que aparecen en todos y cada uno de los registros de la base de datos. Desde el momento que estas palabras representan temas, ideas y conceptos, el índice de un base de datos documental es una representación de los todos los asuntos presentes en todos los documentos que forman parte de la base de datos. Los índices analíticos suelen basarse en una estructura denominada *fichero inverso* o *fichero invertido*.

El fichero invertido

El fichero invertido recibe este nombre por oposición al fichero directo, que es el conjunto de todos los registros introducidos en la base de datos y que están dispuestos de manera secuencial (siguiendo el orden cronológico en el que fueron dados de alta).

La estructura de los índices analíticos está optimizada para permitir la existencia de valores repetidos (por ejemplo, documentos indizados con el mismo descriptor), para realizar búsquedas en documentos de texto completo con gran rapidez y para realizar tareas de control terminológico.

Ejemplo

Si en una base de datos documental aparece cien veces el término *economía*, en cambio hay una sola entrada en el fichero invertido (en el índice de un sistema relacional debería haber cien entradas). Los ficheros invertidos relacionan además datos de contexto con cada término de la entrada, por ejemplo, su frecuencia, su posición exacta en cada registro (número de orden), los posibles sinónimos, etc.

En la clase de índices analíticos que permiten los ficheros invertidos, cada término o entrada del índice es único, lo que facilita tiempos de respuesta muy bajos.

Las dos tablas siguientes ilustran el concepto de un índice analítico mediante el sistema del fichero invertido.

Tabla 2. Composición típica de un índice invertido

Elemento	Explicación
Término	Todas y cada una de las palabras que forman parte de los registros o de los documentos de la base de datos (y que no constan en el fichero de palabras vacías). Son siempre términos únicos, es decir, hay una sola entrada para cada término aunque aparezca muchas veces en uno o en muchos registros de la base de datos.
Frecuencia	Número de registros (por tanto, número de documentos) en los que aparece el término. En algunos ficheros invertidos se consigna como número de veces (frecuencia) con la que aparece en total el término.
Localización	Indicación de los parámetros de localización, imprescindible para la recuperación. La información necesaria consta, al menos, de los siguientes elementos: número documento - número de campo (si es que hay campos) - número de palabra. El motivo es que hay que conocer la posición absoluta de la palabra en el documento para poder aplicar correctamente algunos operadores como el de proximidad.

Tabla 3. Ejemplo de índice invertido

Término	Frecuencia	Localización
…	…	…
Barcelona	2	(00017, 03, 01) (03401, 01, 04)
…	…	…
Madrid	2	(00017, 03, 03) (17200, 02, 01)
…	….	…
Zaragoza	3	(00017, 03, 04) (03401, 01, 02) (17001, 04, 01)
…	…	…

El ejemplo de índice de la tabla 3 incluye, para simplificar, tan sólo tres entradas del índice: las correspondientes a las palabras *Barcelona*, *Madrid*, *Zaragoza*. Si miramos la entrada *Barcelona*, por ejemplo, vemos que hay en total dos registros en la base de datos que contienen la palabra *Barcelona* (ver la columna "Frecuencia"). Como hay dos registros con la palabra *Barcelona*, vemos en la columna "Localización" dos vectores, o sea dos conjuntos de datos: "(00017, 03, 01)" y "(03401, 01, 04)". Según estos vectores, los dos registros que contienen la palabra "Madrid" son el "00017" y el "03401", o sea el primero de cada uno de los tres números que forman cada vector "(**00017**, 03, 01)" y "(**03401**, 01, 04)".

Decimos que los conjuntos "(00017, 03, 01)" y "(03401, 01, 04)" son vectores porque en tales conjuntos la posición de cada elemento es significativa. De este modo, el primer número siempre es el identificador del registro, el segundo número es el identificador del campo y el tercer número identifica el número de orden de la palabra en cuestión dentro del campo considerado. Lo anterior significa que el índice invertido de nuestro ejemplo corresponde a una base de datos con un modelo de registro como este:

01	Título
02	Autor
03	Fuente
04	Descriptores
…	….

En concreto, vemos que *Barcelona* aparece en el campo número 3 del primer registro "(00017, **03**, 01)" pero aparece en cambio en el campo número 1 del segundo de los registros (03401, **01**, 04). Por tanto, lo anterior significa que en el registro número 0017 la palabra "Barcelona" aparece en la **Fuente** (campo 03) y en cambio en el registro número 03401 aparece en el campo **Título** (campo 01).

Vemos así mismo que la palabra *Barcelona* aparece en primera posición en el primero de los dos registros (00017, 03, **01**); pero aparece en cuarta posición en el segundo de los dos registros (03401, 01, **04**), etc.

Para acabar de entender cómo genera (e interpreta en el momento de la búsqueda) un SGD el índice anterior, vamos a representar como podría ser el registro que correspondería al segundo vector [(03401, 01, 04)] de la palabra *Barcelona*:

ID Campo		**03401**
01	Título	Historia ilustrada de **Barcelona**
02	Autor	F. Pujol
03	Fuente	Vic: Editorial ZYX, 2010
04	Descriptores	Barcelona, Historia

Si comparamos el registro anterior con el vector correspondiente [(03401, 01, 04)] podemos ver la correspondencia de una forma clara: el primer número del vector es el número (**03401**) del registro, el segundo número es el identificador del campo (**01**, por tanto, el **título**) y el tercero es el orden de la palabra en cuestión en la frase (la cuarta palabra en el título del documento).

4) Herramientas de control terminológico o lingüístico

Diccionario de palabras vacías

Diccionario que contiene palabras vacías de significado, que tienen únicamente valor gramatical (artículos, preposiciones, pronombres, etc.). En inglés se denomina *stoplist* o *stopword list*.

Aunque hay grandes diferencias entre ellos, casi todos los SGD suelen disponer de diversas herramientas de control terminológico. La más simple es la posibilidad de utilizar diccionarios de palabras vacías, es decir, de términos que no se usarán para indizar los documentos. La más sofisticada es la posibilidad de usar uno o más tesauros, es decir, un lenguaje documental que permite establecer relaciones lógicas entre los términos y los descriptores de una base de datos. En medio, hay diversas posibilidades: uso de sinónimos, listas de descriptores, etc.

Tesauro

Es un diccionario estructurado de conceptos (con jerarquías y relaciones).

5) Lenguaje e interfaces de consultas orientados al usuario

Los SGD están orientados al usuario, y no tanto a otros programas informáticos. Por eso su lenguaje de interrogación dispone de herramientas que facilitan la conversión de una necesidad de información del usuario en una estrategia de consulta, así como facilidades para el mantenimiento y la gestión de operaciones de búsqueda complejas, que pueden requerir consultas reiteradas. Las posibilidades y prestaciones en este sentido son mucho mayores y más versátiles que las que nos ofrecen los SGBDR y esto se explica, fundamental-

mente, por dos razones: en primer lugar, porque el tipo de información que contienen es distinta y, en segundo lugar, porque, como ya hemos visto anteriormente, las necesidades de los usuarios de este tipo de sistemas son muy diferentes a los usuarios de sistemas administrativos.

1.2.2. Síntesis: sistema relacional contra sistema documental

Las diferencias comentadas entre sistemas relacionales y documentales las sintetizamos y sistematizamos en la tabla 4. La comparación se ha llevado a cabo partiendo de dos modelos puros a los que seguramente no todos los programas tienen por qué ajustarse. Por ejemplo, algunos SGD están incorporando herramientas que permiten relacionar bases de datos como si fueran relacionales (*Inmagic*, *FileMaker*). Además, algunos programas relacionales integran bajo una misma interfaz o capa de programación un sistema relacional y un sistema documental (*Oracle* o *BRS*).

De esta forma, la tendencia que se sigue va hacia la paulatina integración de las prestaciones de uno y otro modelo en un solo programa. Así pues, en el mercado podemos encontrar programas que, a pesar de pertenecer a una de las tipologías, dispone de algunas características de la otra. Aún así, es útil comparar los modelos "puros" de cada categoría.

Tabla 4. Principales diferencias entre SGBDR y SGD

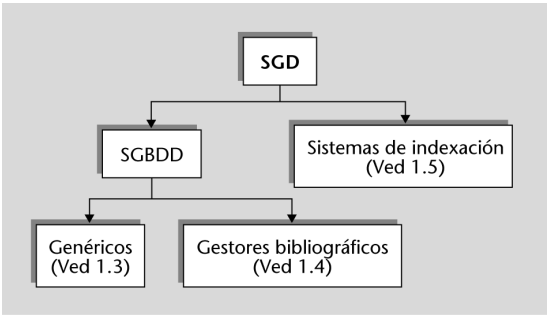
SGBDR	SGD
Estructura	
• Tabular (tablas para representar datos). • Campos con longitud fija. • No pueden haber registros ni campos repetidos. • Tablas homogéneas	• Textual (modelo irrestricto). • Campos y registro de longitud variable • Campos repetibles • Se pueden combinar estructuras diferentes dentro de la misma base de datos (para representar diversos tipos de documento, por ejemplo).
• Conjunto de diversas tablas, con la posibilidad de crear tablas nuevas mediante operaciones de álgebra relacional integradas en el lenguaje de consulta del sistema de gestión de la base de datos.	• Monobase (fichero plano) o bien con un solo tipo de registros por cada base de datos (Knosys) o bien con diversos tipos de registros en la misma base de datos (askSam), pero pudiendo abrir y operar una sola base de datos cada vez. • Multibase, o bien en la forma poder abrir y consultar datos de varias bases de datos a la vez (Inmagic), o bien con la posibilidad de relacionar y operar con diversas bases de datos a la vez en un estilo similar al relacional (CDS/ISIS o Inmagic).
• No utilizan índices analíticos (fichero inverso).	• Usan índices analíticos (fichero inverso).
• Instrumentos de recuperación (recuperación *determinista*) limitados.	• Amplios instrumentos de consulta y recuperación, con muchas ayudas para las búsquedas: búsqueda global en cualquier campo, operadores booleanos, de proximidad, combinación de conjuntos de búsqueda, consulta de índices, etc. (recuperación probabilista).

SGBDR	SGD
Estructura	
• No disponen de controles terminológicos.	• Disponen de instrumentos de control terminológico para la indexación, para la entrada de datos y para la consulta (palabras vacías, listado de autoridades, sinónimos, etc.).
Objeto	
• Información muy estructurada (información de gestión, administrativa, etc.).	• Información poco o nada estructurada (documentos científicos, técnicos o culturales; o bien documentos icónicos).
• Información muy volátil: los datos acostumbran a cambiar con frecuencia	• Información acumulada: los datos suelen ser permanentes y acumulativos
Ámbito	
• Gestión administrativa (ofimática). P.e. matriculaciones, nóminas, etc.	• Servicios de información y documentación (centros de documentación, bibliotecas, museos, editoriales, bancos de imágenes, etc.).
Ejemplos	
• Access, Oracle, Informix, dBase, DB2, SQL server, PostgreSQL, Sybase, MySQL	• Inmagic, CDS/ISIS, FileMaker, Knosys, askSam, RefWorks

1.2.3. Tipología de SGD

En lo que se refiere a la tipología de programas de gestión documental, podríamos resumir en dos los principales modelos presentes actualmente en el mercado:

Figura 3. Tipología de SGD



1) Sistemas de gestión de bases de datos documentales (SGBDD)

Los primeros programas de gestión documental estaban pensados para gestionar solamente referencias de documentos, y no el texto del documento completo. Algunos ejemplos son *CDS/ISIS* o *Inmagic*.

SGBDD genéricos

En el apartado 1.3 se tratan con detalle.

También se tienen que considerar los sistemas de gestión bibliográfica, como una variante dirigida a usuarios personales. Los ejemplos más conocidos son *RefWorks*, *EndNote*, *Zotero* o *ProCite*.

Los sistemas de gestión bibliográfica

Se van a tratar con más detalle en el apartado 1.4.

2) Sistemas de indexación

Estos sistemas están especialmente orientados al tratamiento del texto completo de los documentos. Son programas que no necesitan definir modelos de registro, aunque algunos de ellos lo pueden hacer de modo opcional. Su especificidad radica en la capacidad de generar índices analíticos (ficheros invertidos) del contenido de los documentos y guardarlos en un disco duro o en una red de discos duros. Los documentos siguen en su formato original y el índice es únicamente un puntero a cada documento concreto. Para visualizar el documento, el sistema activa al programa con el cual fue creado el documento en cada caso. Se denominan sistemas de indexación o motores de búsqueda. Algunos ejemplos de sistemas de indexación o motores de búsqueda son *Autonomy* o *Google Search*.

Sistemas de indexación o motores de búsqueda

Se tratan con mayor profundidad en el apartado 1.5.

Para diferenciar entre ambos modelos de SGD hay que observar cuáles son las funciones priorizadas. En el caso de los SGBDD destaca especialmente el apartado de definición de la base de datos, que facilita la aplicación de un diccionario de datos complejo, mientras que los sistemas de indexación tienen muy desarrollado el módulo de indexación, que permite generar los índices invertidos de documentos extensos.

1.3. Sistemas de gestión de bases de datos documentales (SGBDD)

Bajo esta denominación se incluiría a los primeros SGD, a los más tradicionales, aquellos que facilitan básicamente la gestión de referencias de documentos de todo tipo. Estos programas comparten una serie de elementos estructurales que permiten la creación y explotación de bases de datos.

1.3.1. Estructura

Dispone de cinco módulos básicos:

1) Definición de los registros de la base de datos

Este grupo funcional está relacionado con el diseño y la creación de las bases de datos. En particular, permite definir campos, especificar el comportamiento de los mismos y definir modelos de registros mediante agrupaciones de campos. En el proceso de creación de bases de datos, las especificaciones detalladas de cada modelo de registro y del comportamiento de cada campo suelen detallarse previamente en un documento escrito que recibe el nombre de *diccionario de datos* (ved 3.3.1.). En el diccionario de datos y en módulo funcional correspondiente del SGBDD se detallan también aspectos relacionados con el tipo de dato asignado a cada campo (textual, numérico, lógico, etc.) y al control terminológico (campo con descriptores, campo indizado, etc.).

2) Mantenimiento

El módulo de mantenimiento controla las operaciones de altas, bajas y modificaciones de registros.

3) Indexación y recuperación

Se trata de uno de los módulos fundamentales y, sin duda, el más específico de este tipo de programas. Aquí se incluyen las funciones relacionadas con el proceso de generación de los índices, las prestaciones de recuperación (operadores booleanos, de proximidad, etc.) y las formas de mostrar y ofrecer los resultados a los usuarios.

4) Salida e intercambio

Estas opciones comprenden tanto los aspectos relacionados con la salida de los registros (exportaciones) como las operaciones que se ocupan de la incorporación y adaptación de ficheros externos de registros (importaciones). Ambos procesos son fundamentales en un SGBDD ya que aseguran la difusión y la posibilidad de intercambio de sus datos con el exterior del sistema.

5) Administración de la base de datos

Este módulo agrupa todas las funciones y procesos relacionados con el control y la gestión de la base de datos, es decir, el sistema de seguridad (poder crear grupos de usuarios y la adscripción de distintos privilegios a cada grupo de usuarios, así como la administración de nombres de usuario y de contraseñas) y la programación y modificaciones en la interfaz.

1.3.2. Mercado

A continuación se presenta una breve ficha descriptiva individualizada de los principales SGBDD que pueden encontrarse actualmente en el mercado español. Hemos reducido la lista a cuatro programas que consideramos que son los que están más implantados y que pueden responder a un tipo de necesidad de más alto nivel (sería el caso de *Inmagic DB/Text* y de *CDS/ISIS*) o de un nivel medio (en esta situación están *FileMaker* y *Knosys*).

Nombre	CDS/ISIS
Productor	Unesco <http://www.unesco.org/webworld/isis/isis.htm>
Distribuidor	Cindoc <http://www.cindoc.csic.es>
Comentarios	• Especialmente adecuado para el tratamiento de la información bibliográfica. • Utilización de subcampos. • Indexación con diversas técnicas (palabra a palabra, grupos de palabras, campo entero, subcampos).
Ejemplos	• Icomos - http://databases.unesco.org/icomos/ • Biblioteca del IDAIC: http://www.dba.it/idaic/ricerca.html • Bases de datos de Bireme: http://bases.bvs.br

Nombre	FileMaker
Productor	Claris <http://www.filemaker.fr/spain>
Distribuidor	Claris <http://www.filemaker.es>
Características	• De muy fácil utilización. • Muy versátil, de hecho es el programa con capacidad documental más integrado a la vez en el mundo ofimático. • Capacidad relacional. • Escasas herramientas de control terminológico.
Ejemplo	Bibliografia sobre Bibliotecas Públicas: http://www.bibliotecaspublicas.info/

Nombre	Inmagic DB/Text
Productor	Inmagic <http://www.inmagic.com>
Distribuidor	Doc 6 <http://www.doc6.es>
Características	• Muy versátil: adecuado para el tratamiento referencias bibliográficas y para gestionar cualquier tipo de objeto o entidad. • Capacidad relacional. • Amplias posibilidades de adaptación y personalización de las interfaces de usuario. • Amplias posibilidades de control terminológico. • Gestión integrada de tesauros. • Dos formas distintas de indexación (palabra a palabra, por frases).
Ejemplos	• Foundation center - http://lnps.fdncenter.org/search.html? • Directorio de bases de dades - http://www.andornot.com/WebPubLinks

Nombre	Knosys
Productor	Micronet <http://www.micronet.es>
Distribuidor	Micronet
Comentarios	• Fácil utilización. • Posibilidades limitadas de control terminológico. • No diferencia entre el fichero de definición de campos, entrada de datos y visualización.
Ejemplo	Dialogyca BDDH (Universidad Complutense).

1.4. Sistemas de gestión bibliográfica o gestores bibliográficos

Se trata de una clase de sistemas de gestión documental centrada en una necesidad muy característica (y posiblemente exclusiva) del colectivo académico e investigador, a saber: el almacenamiento de referencias bibliográficas, su posterior recuperación de forma selectiva y la generación de bibliografías con diferentes formatos.

Como es sabido, una de las características del trabajo académico es la necesidad de publicar. Es el famoso *publish or perish* (o publicas o pereces). Los miembros de la academia, ya sean estudiantes de ciclos superiores, profesores de universidad o investigadores solamente pueden justificar su carrera académica y sus avances en la investigación a través de la publicación de artículos en revistas científicas.

Los académicos y los investigadores

Diferenciamos entre académicos e investigadores porque ni todos los académicos son investigadores ni todos los investigadores son académicos. Por ejemplo, los estudiantes de los últimos ciclos de universidad sin duda son académicos, pero no necesariamente sus trabajos son frutos de la investigación, sino más bien de estudios avanzados. Por otra parte, hay muchos investigadores de empresas y corporaciones que no están en la academia, ya que no trabajan en la Universidad ni imparten docencia.

El punto importante aquí consiste en que estos artículos deben basarse en conocimientos anteriores. La ciencia se define como conocimiento acumulativo. És más, uno de los criterios que se manejan para separar a la ciencia de la pseudo ciencia es este: si alguien afirma que sus descubrimientos no deben nada a los conocimientos anteriores, es seguro que estamos ante un estafador.

La cuestión es que a ningún científico ni académico serio se le ocurre trabajar sobre el vacío. Al revés, parte del éxito de una investigación en concreto, o de toda una vida académica en general, descansa en la habilidad del investigador para analizar la producción científica anterior en su campo.

Lo cual nos conduce a la necesidad clásica de investigadores y académicos de manejar citaciones y bibliografías. Este colectivo es muy numeroso, al menos a nivel internacional, y esta necesidad es imperiosa. Probablemente, estos dos factores explican que los sistemas de gestión bibliográficos se cuenten entre los más populares en el mundo de la gestión documental (y de las aplicaciones de la Web 2.0).

1.4.1. Estructura y características

Un sistema de gestión bibliográfica suele disponer de estos cuatro grupos funcionales:

1) Un conjunto de medios para la entrada de referencias bibliográficas

Esto incluye la entrada manual a través de registros predefinidos para los tipos documentales habituales en el mundo académico: artículos de revista, libros, ponencias y comunicaciones en congresos, capítulos de libro, etc. Pero lo más importante es que cada vez más los sistemas de gestión bibliográfica pueden importar referencias de forma automática, desde diferentes fuentes de información, típicamente bases de datos, catálogos, buscadores, etc.

2) Un sistema de búsqueda

Ningún sistema documental está completo sin un sistema de búsqueda que permita crear grupos selectivos de referencias. Esta función queda asumida por la búsqueda simple o avanzada del sistema que permitirá al usuario crear páginas de resultados con referencias bibliográficas basadas en palabras clave, en nombres de autores, en títulos de revistas, etc.

3) Un sistema para generar bibliografías

A diferencia de otros sistemas documentales, el objetivo final no es sólo la recuperación de información, todo y su importancia capital, sino generar una bibliografía configurada de acuerdo con una norma o formato específico. Por ejemplo, el formato de bibliografía que nos exige la revista donde pretendemos publicar el artículo o la que marca la Universidad donde queremos defender un trabajo académico.

Ejemplo

El formato de bibliografía que nos exige la revista donde pretendemos publicar el artículo o la que marca la universidad donde queremos defender un trabajo académico.

La bibliografía como resultado final, a su vez, puede ser generada en formato impreso o como archivo informático. En este último caso, suele existir la posibilidad de elegir entre diversas formas de codificación: archivo ASCII, DOC, ODT, RTF, HTML, XML, etc.

4) Eventualmente, un sistema para introducir citas y posteriormente generar la bibliografía de forma automática

Los sistemas de última generación suelen proporcionar un plugin para editores de texto (*Word* u *OpenOffice*, típicamente). Gracias a este plugin, es posible incorporar citas en el texto de un trabajo académico (artículo, tesis, etc.) mediante búsquedas en la base de datos realizadas desde el editor de textos. Posteriormente, el sistema genera la bibliografía completa, formateada y ordenada de manera adecuada, al final del documento.

1.4.2. Mercado: sistemas de escritorio *vs* sistemas online

Los programas más importantes de esta categoría se dividen en:

1) Sistemas de escritorio, o sea software convencional que reside en, y se carga desde, el disco duro del usuario y

2) Sistemas en línea, es decir, software que no es necesario instalar en el ordenador del usuario, sino que se ejecuta a través de un navegador de Internet.

Al final, ambos acaban ejecutándose en la memoria RAM del ordenador del usuario, pero a efectos prácticos hay grandes diferencias entre los programas de escritorio y los programas en línea.

Posiblemente, las aplicaciones en línea están llamadas a cambiar (¿a revolucionar?) la ofimática de los próximos años. Dicho de otra forma: probablemente, la futura versión de Office (*MS* u *OpenOffice*) ya no se ejecutará en modo local, sino que nos conectaremos a un servidor web y la ejecutaremos desde un navegador; después, naturalmente, de identificarnos y entrar en nuestro espacio personal. Solo ocasionalmente haremos uso de la versión de escritorio correspondiente, justo al contrario que ahora.

De hecho, es ya la forma más habitual de trabajar para los sistemas que nos ocupan ahora, los gestores de información bibliográfica, pero también de los gestores de contenidos y cada vez más de los gestores y editores de imágenes.

La tabla siguiente es un resumen de características de ambos tipos de aplicaciones (online y escritorio).

Tabla 5. Características de los gestores bibliográficos

Característica	Escritorio	Online
Mantenimiento de la aplicación	A cargo del usuario final	A cargo del proveedor del servicio
Uso de la aplicación	Modo local desde un ordenador con una aplicación necesariamente pre instalada y configurada	Cualquier ordenador de cualquier lugar del mundo sin necesidad de instalación previa de ninguna aplicación
Ubicación de los datos	Disco duro de un ordenador concreto accesible solo en modo local	Disco duro de un servidor web accesible desde cualquier navegador y por diversos usuarios. Disco duro del usuario si lo desea
Seguridad	Típica de usuarios finales (débil y contradictoria)	Típica de organizaciones profesionales (fuerte y sistemática)
Velocidad potencial	Muy alta (solamente limitada por el hardware del usuario)	Alta/Media/Baja (limitada y dependiendo del tipo de conexión a Internet)
Funciones	Sin límites a priori	Limitadas por problemas logísticos

Aplicaciones en línea

Dentro de los gestores bibliográficos en línea, dos de las aplicaciones más importantes, por su presencia en universidades y centros de investigación, son *RefWorks* (www.refworks.com) y *EndNote Web* (www.endnoteweb.com). En este caso, se trata de aplicaciones comerciales, es decir, su utilización requiere una subscripción previa por parte de la institución (universidad o centro de investigación), que a su vez queda a disposición de los investigadores o usuarios individuales de la institución.

En el caso de *RefWorks* y *EndNote* se trata de sistemas funcionalmente muy similares y que representan probablemente el nivel más alto del estado de la cuestión en esta tecnología por la excelencia de sus funciones y prestaciones. Sus únicas diferencias proceden de las empresas respectivas, ambas ligadas a la industria de las bases de datos.

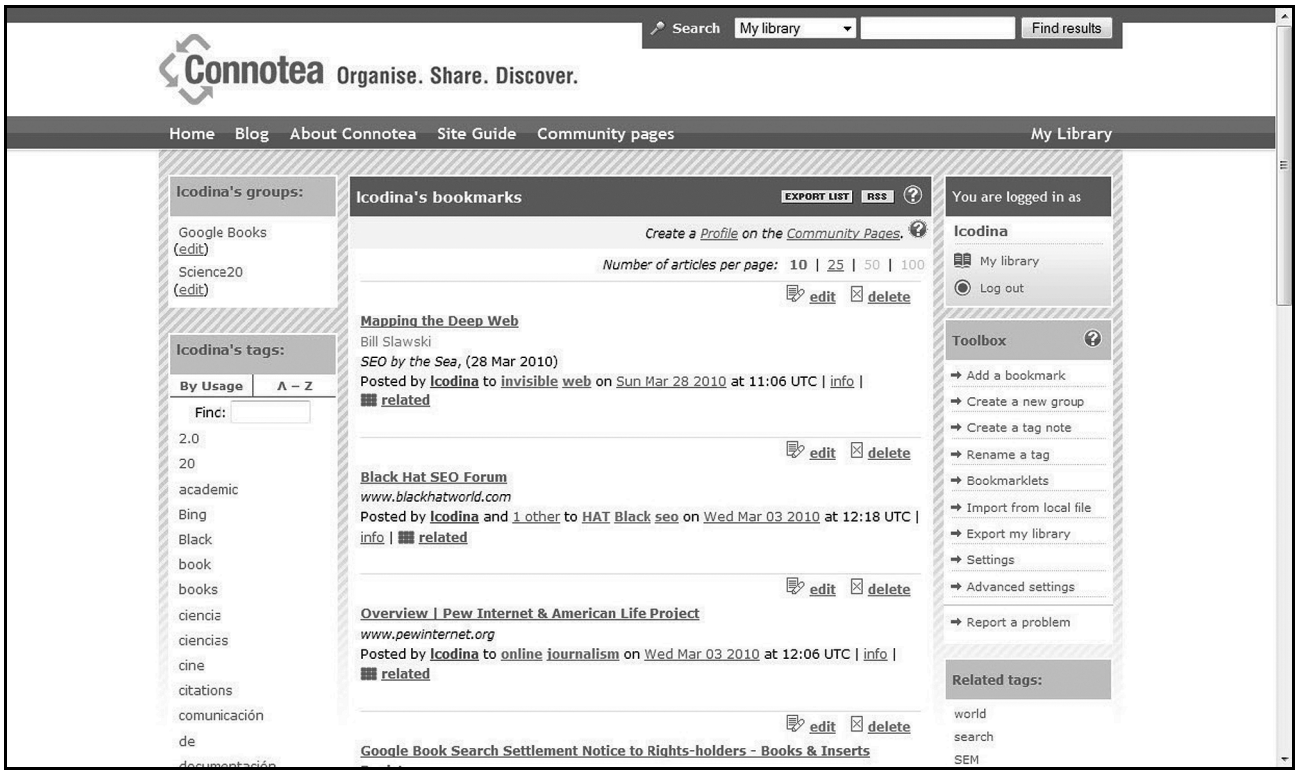
RefWorks es un producto del productor y distribuidor de bases de datos académicas ProQuest, mientras que *EndNote Web* es un producto de Thomson Reuters. Ambas son empresas productoras y distribuidoras de las bases de datos científicas más importantes del mundo.

Además de las dos anteriores, existe un grupo de aplicaciones en línea para gestión bibliográfica de coste cero que presentan prestaciones muy notables, en algunos casos comparables a productos comerciales como los señalados antes. Los más importantes son *Zotero* (www.zotero.org) –en realidad es una aplicación que combina un software de escritorio y un sistema en línea–, *Connotea* (www.connotea.org), *CiteULike* (www.citeulike.org) y *Mendeley* (www.mendeley.com).

Las características comunes a estos cuatro programas gratuitos son las siguientes:

- Extremada facilidad de uso.
- Plena integración con el navegador de Internet.
- Extrema facilidad para importar información de páginas web.

Figura 4. Interfaz del sistema de gestión bibliográfico a *Connotea*



Aplicaciones de escritorio

Las aplicaciones de gestión bibliográfica de escritorio más importantes se da la circunstancia que pertenecen a la misma empresa (por un proceso de adquisiciones a los largo de los últimos años), en concreto a una de las divisiones de software de Thomson Reuters, y se trata de las siguientes:

- *Reference Manager* (www.refman.com)
- *ProCite* (www.procite.com)
- *EndNote* (www.endnote.com)

El motivo por el cual una misma empresa produce y mantiene tres productos similares es un misterio para estos autores. Todo lo que puede decirse es que existe una ligera diferenciación del usuario final de cada producto. *Reference Manager* es el sistema más completo y complejo, y por tanto destinado a un usuario experto, mientras que *ProCite* es probablemente el sistema más fácil de utilizar a costa de prescindir de algunas prestaciones y por tanto destinado a usuarios que no necesitan profundizar en el uso de la aplicación. Por último, *EndNote* se sitúa en algún punto intermedio y es, a la vez, el que dispone de una versión en línea, como hemos visto antes.

1.5. Sistemas de indexación

Los motores de búsqueda, también denominados *indizadores* o *sistemas de indexación*, se han hecho justamente famosos a raíz del importante papel que están

jugando los buscadores de Internet, especialmente *Google*. Estos servicios, que facilitan el acceso al texto completo de los documentos que se encuentran en Internet, disponen de un motor de búsqueda (*search engine)* que facilita la consulta de cualquier término o combinación de términos que aparezcan en las páginas web y otros documentos (PDF, por ejemplo).

Lo cierto es que en el inicio de la Web, ya existían estos sistemas de indexación. Sus antecedentes se remontan a las primeras bases de datos de texto completo. *Lexis* fue uno de los primeros sistemas que ofreció acceso al texto completo de los documentos que contenía. Esto pasaba entre finales de los años setenta y principios de los ochenta. Existieron otros programas de este estilo que funcionaban con grandes sistemas al menos desde la década de los años ochenta como *STAIRS*, *Basis*, *DOCU-MASTER*, etc. También durante los años ochenta aparecieron las primeras versiones de esta clase de programas para microrordenadores: *AskSam*, *Personal Librarian* o *ZyIndex* son algunos ejemplos.

Así pues, los motores de búsqueda son un tipo de SGD que sirven para crear bases de datos de texto completo, que elaboran unos voluminosos índices que permiten recuperar la información a partir de cualquier palabra que forme parte de los documentos de la base de datos.

No existe una denominación consolidada para referirse a este tipo de programa. En inglés se utiliza el término *text retrieval software*, juntamente con *full-text retrieval system* o *text information management system*, entre otras, para referirse a este tipo de SGD. En francés se utiliza la expresión *moteurs de indexation et de recherche*.

La diferencia esencial entre un sistema de gestión de bases de datos documentales (SGBDD) y un motor de búsqueda (o sistema de indexación) es que este último no tiene ningún módulo para diseñar y administrar modelos de registro. De hecho, los motores de búsqueda no utilizan registros en el sentido de representaciones de los documentos (documentos secundarios), sino que generan índices directamente a partir del análisis de los documentos originales, al estilo de los programas que están detrás de *Google*.

1.5.1. Estructura y características

Para describir el funcionamiento de este tipo de programa partiremos de los módulos básicos que se han descrito anteriormente para los SGBDD (véase 1.3.1.), es decir, administración de la colección (o base de datos), mantenimiento, indexación, recuperación, y salida e intercambio de información.

1) Administración de la colección

Los documentos que forman parte de la colección o del fondo documental se mantienen en la máquina original (bien sea un ordenador local o remoto). El

programa de indexación genera unos índices a partir de los cuales se puede acceder a los documentos de forma selectiva a partir del contenido del texto completo de la colección.

De este modo, la colección está formada por dos tipos de datos. Por un lado, los ficheros con los documentos y por otro, los índices que remiten a estos documentos. Los documentos pueden estar localizados en diversas unidades de almacenaje o en servidores externos, y lo único que hay que tener en cuenta es su ubicación precisa en el momento de definir la colección (en qué unidades de disco o cuáles son las direcciones de los servidores remotos donde se encuentran los ficheros con los documentos) que hay que indizar. Cuando ejecutemos el programa utilizaremos los índices y, con el apuntador del documento, podremos visualizarlo a través de la aplicación original con la que fueron creados.

Ejemplo

Si el documento recuperado es una página web, podremos verla en un navegador, pero si se trata de un documento de texto, podremos verlo en un editor de texto.

Aunque esta clase de aplicaciones no acostumbra a estructurar los documentos, es cada vez más frecuente el uso de campos o de etiquetas que permiten dar una apariencia de estructura de campos a la colección y facilitan el acceso a partes concretas del documento; habitualmente, el título, la fecha de creación o el autor. Esta estructuración pueden realizarla, pese a no utilizar una auténtica estructura de registros, por derivación de los metadatos que suelen generar cada vez más aplicaciones.

2) Mantenimiento (entrada de datos)

Tal y como se deduce de lo que se ha descrito en el anterior apartado, la entrada de datos al sistema no se acostumbra a hacer desde el teclado porque se dispone ya de los ficheros informáticos con la información que se ha de procesar (documentos HTML, documentos de texto, de hojas de cálculo, gráficos, etc.).

El problema puede provenir de la diversidad de formatos en los que pueden estar los documentos que han de formar parte de la base de datos (o colección), que pueden ser de todo tipo (DOC, ODT, RTF, HTML, XLC, PDF, EDS, TIFF, etc.).

3) Indexación

El programa indiza el texto completo de los documentos que forman parte de la base de datos o colección y también, si los hubiera, los indicadores, marcas de campo o metadatos. De esta manera se pueden acotar las consultas a un campo determinado del registro.

4) Recuperación

En general, el proceso de consulta en sistemas de indexación se realiza de manera similar a la consulta de bases de datos de tipo referencial, es decir, se usa el álgebra booleana, y se dispone de una serie de operadores complementarios (truncamiento, proximidad, etc.). Ahora bien, además de este tipo de consulta, que es la tradicional en todos los programas de recuperación de la información, los motores de

búsqueda están experimentando con otros tipos de prestaciones, fundamentalmente, las búsquedas semánticas y las búsquedas por patrones.

a) Búsqueda semántica

Se trata de poder ampliar la consulta de un término a todos aquellos que estén relacionados de alguna manera con él –derivación morfológica, equivalencia lingüística, sinonimia, generalización, especialización, etc.

Ejemplo

Si un usuario busca información sobre *copyright* el sistema muestra un conjunto de términos relacionados con el solicitado (*propiedad intelectual, derechos de autor, derechos de explotación,* etc.).

b) Búsqueda por patrones

Se trata de un sistema totalmente opuesto al anterior. En este caso nos referimos a un análisis que no tiene en cuenta la morfología (la forma de las palabras), ni la sintaxis (el orden y coordinación de las palabras), ni la semántica (los significados), sino que la indexación de la información se basa en patrones de bits. De esta manera, cualquier tipo de información, ya sea texto, sonido o imagen, está indizada y se recupera empleando el mismo sistema de representación. Esta tecnología se basa en la apariencia física de los términos (su código binario) y no en la semántica (su significado).

Las búsquedas por patrones permiten comparar textos o imágenes a partir de patrones binarios, es decir, permiten encontrar textos o imágenes que comparten una serie de características estructurales comunes.

Ejemplo

Si buscamos *Eltsin*, el programa nos facilitará todos los documentos en los que aparezca exactamente esta palabra y también aquellos en los que consten otras palabras parecidas: *Yeltsin*, *Elsin*, *Ieltsin*, etc. Lo mismo podría pasar con *Gadafi*, *Kadhafi*, *Kadafi*, *Gadaffi*, etc. Las variaciones pueden ser debidas a que los términos hayan sido mal escritos, mal reconocidos por un OCR, o a que se trate de transliteraciones realizadas con criterios diferentes.

5) Ponderación de resultados

La utilización sin control de las búsquedas semánticas y por patrones comporta la aparición de un buen número de resultados no deseados. Los mecanismos de ponderación de términos constituyen un instrumento complementario muy útil y prácticamente imprescindible para minimizar los efectos no deseados de este tipo de consultas. Hay que tener presente que, en estos entornos, normalmente se recupera un número muy alto de documentos y hay que disponer de instrumentos que ayuden a determinar cuáles son los más relevantes.

1.5.2. Aplicaciones

Las aplicaciones de los sistemas de indexación son básicamente de dos tipos: buscadores de páginas web y bases de datos de texto completo.

1) Buscadores de páginas web

Como ya hemos apuntado, los buscadores de páginas web han popularizado a los sistemas de indexación, dándolos a conocer al gran público. Estos pueden ser generales o a escala de toda la web (*Google*, *Yahoo*, *Bing*, etc.), o particulares de una sola sede web, como puede ser el caso del buscador del sitio

de la Universitat Oberta de Cataluña (www.uoc.edu) o de cualquier otro organismo público o privado.

2) Bases de datos de texto completo

Los sistemas de indexación también pueden utilizarse para crear bases de datos de texto completo. En este caso puede existir una cierta unidad temática o de publicación en el contenido de la base de datos, a diferencia del anterior, en que podemos encontrar una amalgama y variedad muy dispar.

En la mayoría de los casos se utilizan metadatos descriptivos del contenido (autor, título, fecha, etc.) y en todas ellas utilizan algún tipo de metadatos administrativos, estructurales o de derechos de propiedad. Podemos encontrar ejemplos diversos en prensa, revistas científicas y fondos editoriales.

Ámbito temático	Ejemplo	Sitio web	Comentarios
Prensa	*MyNews*	www.mynewsonline.com	
	El País	www.elpais.com/archivo/	
	La Vanguardia	www.lavanguardia.es/hemeroteca/	
Revistas académicas y científicas	*Ariadne*	www.ariadne.ac.uk/search/	
	Information Research	http://informationr.net/ir/search.html	
Fondos editoriales	*Ocenet* (Editorial Océano)	http://consulta.oceano.com	Accesible desde la red de bibliotecas de la Diputación de Barcelona
	V-lex	http://vlex.com/	BD de legislación con documentos de 130 países

1.5.3. Mercado

Aunque la mayoría de aplicaciones requieren de la estructura cliente-servidor, es posible encontrar también algunas versiones personales que funcionan con un microordenador. A continuación se describen los principales programas de este estilo que se encuentran en el mercado español.

Nombre	Apache Lucene
Productor	Apache (http://lucene.apache.org/)
Ejemplo	Lista en: http://wiki.apache.org/lucene-java/PoweredBy

Nombre	askSam
Productor	askSam (http://www.asksam.com)

Nombre	Autonomy
Productor	Autonomy <http://www.autonomy.com>

Nombre	Google Custom Search
Productor	Google <http://www.google.com/coop/cse/>
	Sirve para indexar una parte de la web que sea del interés del usuario. Versión de pago (*site search*) y versión gratuita.

Nombre	Google Search Appliances
Productor	Google <http://www.google.com/enterprise/public_search.html>
	Motor de búsqueda para servidores propios. Aplicación para intranets o sedes web. Hay una versión reducida (*mini search*).

Nombre	Greenstone
Productor	New Zealand Digital Library Project at the University of Waikato (www.greenstone.org). Software libre.
Ejemplo	New Zealand Digital Library <http://nzdl.sadl.uleth.ca/cgi-bin/library>

Nombre	Swish-e
Productor	Comunidad de desarrolladores (http://swish-e.org/). Software libre.
Ejemplo	Ariadne (http://www.ariadne.ac.uk/search/)

2. Distribución de bases de datos

La producción y la distribución de bases de datos son dos procesos complementarios que en ocasiones realizan agentes distintos, con tecnología y herramientas bien diferenciadas. En el apartado anterior nos hemos centrado en la producción de bases de datos, es decir, en el proceso de creación y elaboración de unos contenidos informativos que quedan estructurados de una determinada forma y que son explotables con el concurso de un sistema informático. La distribución, en cambio, es el conjunto de operaciones que facilita a los usuarios el acceso a estos contenidos informativos. Así pues, mientras el proceso de producción permite elaborar un contenido único, el proceso de distribución permite que pueda llegar a su público por distintos canales.

Estas diferencias esenciales entre ambos procesos explican que las estrategias y los programas informáticos relacionados con la producción, normalmente tienen poco que ver con los mecanismos e instrumentos que se utilizan para la distribución.

Hasta hace pocos años, los productores y los distribuidores de bases de datos (estos últimos en particular) acostumbraban a tener un carácter especializado y a disponer, por tanto, de una estructura empresarial y profesional que les apoyaba. Esta situación ha cambiado radicalmente con la eclosión de Internet y el desarrollo de distintas herramientas fácilmente configurables y adaptables que ponen al alcance de centros de información y documentación pequeños y medianos, e incluso de usuarios personales, la posibilidad de convertirse en productores y distribuidores de bases de datos. Estas funciones, con la tecnología anterior, eran muy difíciles de desempeñar sin una infraestructura muy especializada y costosa. En la exposición que hacemos aquí tendremos en cuenta este cambio importante.

2.1. Antecedentes

La consulta local fue el primer sistema que se utilizó para facilitar el acceso de los usuarios a la base de datos. Tiene importantes limitaciones, fundamentalmente de acceso, ya que obliga a los usuarios a desplazarse expresamente al centro de documentación o unidad donde está disponible la aplicación.

De forma complementaria, especialmente en centros del ámbito de las humanidades y ciencias sociales, también se utilizó la publicación de bibliografías impresas como forma para la distribución de los contenidos de bases de datos. Estas bibliografías se pueden elaborar automáticamente desde algunos SGD y constan, básicamente, de dos partes: en primer lugar, una lista global correlativo de los registros numerados u ordenados por algún elemento descriptivo, normalmente el autor, y que incluye la descripción completa de cada uno de

los registros; en segundo lugar, se pueden encontrar índices diversos –autores, títulos, materias, etc.– que remiten al número de registro de la lista general. Los inconvenientes más destacables son los altos costes de impresión y de distribución y las dificultades para actualizar las obras.

Finalmente, otra vía que se ha utilizado en el pasado es el soporte óptico, que implicaba incorporar al disco compacto el programa de recuperación de la información (o, al menos, el módulo de consulta). Esta vía fue notablemente usada a finales del s. XX y luego pasó rápidamente a la historia con la eclosión del web.

Bibliografías impresas

Inmagic, *ISIS*, *Pro-Cite* y, en menor medida, *Knosys*, disponen de sistemas para facilitar, más o menos, esta tarea de acceso a la base de datos.

2.2. Web

El Web es el sistema de distribución de bases de datos documentales más utilizado. El motivo es sencillo: el usuario que consulta la base de datos sólo tiene que contar con un navegador para poder acceder a los registros de forma actualizada y disponiendo, en algunos casos, de las mismas prestaciones de consulta y explotación que tienen los sistemas de gestión documental.

En este sentido, el usuario no necesita instalar ninguna versión cliente del programa que gestiona la base de datos, sino que es el propio navegador de Internet (*Internet Explorer*, *Mozilla* u *Opera*) el que actúa como cliente de la base de datos. Desde el navegador, tan sólo tendrá que indicar su petición mediante un formulario HTML para recibir las respuestas también en este formato que el navegador no tendrá ninguna dificultad en reproducir en el monitor del usuario.

Formulario HTML

Es una sección de una página web que contiene elementos de control (casillas de verificación, botones de elección, etc.) y permite introducir texto para ser procesado por un servidor web.

Ahora bien, para que este método de acceso sea posible, es necesario en el lado del servidor un programa o un conjunto de programas que permita establecer la comunicación entre dos entornos en principio incompatibles o distintos: la base de datos gestionada por el SGBD, por un lado, y el navegador web, que utiliza el usuario y que sólo es capaz de interpretar páginas HTML transmitidas mediante el protocolo http, por el otro. Estos programas suelen recibir la denominación CGI (*common gateway interface*) o interfaz de pasarela.

Término *pasarela*

Se utiliza el término *pasarela* (*gateway*) porque hace referencia a la función de relación entre el servidor web y las aplicaciones externas.

2.2.1. Estructura

Los elementos básicos que intervienen en este proceso son los siguientes:

- Navegador (por ejemplo, *Explorer*, *Mozilla*, etc.).
- Servidor httpd (por ejemplo, *Apache*, *Internet information server*, etc.).
- Programa CGI (por ejemplo, *WwwIsis*, *WebPublisher*, etc.).

- Interfaz de consulta.
- Base de datos.

Los programas (el servidor httpd y el CGI) estarán instalados en un servidor, que contará con tarjeta de red y una dirección IP.

De todos los elementos enumerados, quizá el menos conocido sea el programa CGI (Common Gateway Interface) que actúa de sistema de comunicación o pasarela entre los registros de la base de datos, que no están codificados en HTML, y el navegador web, que sólo puede interpretar información codificada en HTML. El protocolo CGI es un estándar desarrollado originalmente para Unix. La creación de esta especificación fue obra de los principales autores de los servidores http (Tony Saunders, entre otros) y se explica porque no querían tener que ir ampliando constantemente las funciones de los servidores para irlos adaptando a los nuevos programas. Es por ello que prefirieron crear un núcleo para el servidor web y proporcionarle un instrumento que le permitiera extender sus servicios y capacidades.

Así pues, el protocolo CGI es un estándar por medio del cual un servidor web (httpd) se puede comunicar con un programa externo, obteniéndose documentos HTML dinámicos (es decir, que se generan al momento, ya que varían según cuál haya sido la petición del usuario). Este protocolo establece una forma de enviar datos desde una página web –por medio de un formulario– y de procesarlos mediante un fichero ejecutable –programa CGI– que está situado en el directorio cgi-bin, o equivalente, de un servidor.

Por otro lado, un programa CGI es una aplicación informática escrita en lenguaje de programación (*Perl*, *C*, *C++*, etc.) que posteriormente es ejecutada e interpretada por un servidor web para poder contestar peticiones de información de los usuarios. El programa CGI es capaz de leer e interpretar las órdenes que se le transmiten desde un formulario HTML, algunas de ellas introducidas por el usuario (por ejemplo, los términos de búsqueda) y otras correspondientes a parámetros generales (por ejemplo, la ubicación del programa y de la base de datos en el servidor, el formato de visualización, el número de documentos a visualizar, etc.). A continuación, los ejecuta y el resultado lo transfiere al usuario en formato HTML.

Además del programa CGI, es necesario preparar una interfaz de consulta adaptada a la base de datos que tenga en cuenta los campos que se han definido, los formatos de visualización, etc. Esta interfaz se construye con el lenguaje de programación del programa CGI, entremezclada con código html y consta básicamente de tres elementos: pantalla de consulta, pantalla de visualización de resultados (lista) y pantalla de visualización del documento completo.

Sobre esta interfaz, podéis ver el apartado 2.3 de este módulo didáctico.

2.2.2. Mercado

Los programas CGI que se indican a continuación sirven para poder consultar bases de datos documentales que han sido creadas con el SGBD correspondiente (ya sea *FileMaker*, *Knosys*, *Inmagic*, o *CDS/ISIS*).

Nombre	*FileMaker*
Productor	Claris <www.filemaker.com>
Distribuidor	Claris <www.filemaker.com/es>
Comentarios	• Dispone de un asistente que permite elaborar rápidamente la interfaz de consulta.
Ejemplos	BIGPI (Geología de la Península Ibérica): http://www.bib.ub.edu/fileadmin/bigpi/bigpi.htm

Nombre	Knosys Internet
Productor	Micronet <http://www.micronet.es/menu/prof/mki.htm>
Distribuidor	Micronet
Comentarios	• Dispone de un asistente aunque es un poco limitado. • No permite ordenar los registros por ningún criterio.
Ejemplos	En el apartado "Clientes" de las páginas dedicadas a KnosysInternet se puede encontrar una lista de usuarios.

Nombre	WebPublisher
Productor	Inmagic <www.inmagic.com>
Distribuidor	Doc 6 <www.doc6.es>
Comentarios	• Dispone de un buen asistente. • Se pueden mostrar los índices de campo.
Ejemplos	Directorio de bases de datos: <http://www.andornot.com/webpublinks>

Nombre	WwwIsis + GenIsis
Productor	Bireme <http://regional.bvsalud.org/>
Distribuidor	Bireme
Comentarios	• Dispone de asistente (GenIsis). • Se pueden mostrar los índices de campo.
Ejemplos	Bases de datos de BVS: <http://bases.bvs.br>

2.3. Interfaz de consulta

La interfaz de consulta de una base de datos sirve para establecer la comunicación entre personas que buscan información y los sistemas de recuperación de la información, y es una de las partes más importantes del proceso de distribución de una base de datos. Como ya hemos avanzado, en el caso de base de datos distribuidas a través de la web, esta interfaz está

formada por un conjunto de páginas HTML de las cuales podríamos destacar las cinco clases siguientes:

- El formato de consulta.
- Los resultados.
- La visualización del documento completo.
- La información general.
- Las ayudas.

En el caso de las interfaces de consulta de bases de datos web, destacamos dos textos que han abordado el proceso de consulta a bases de datos desde el punto de vista del proceso seguido por el usuario (Marchionini, 1995; Shneiderman 1997) y también a los dos textos clásicos en el ámbito de la usabilidad y la arquitectura de la información (Nielsen, 2000, 2006; Morville y Rosenfeld, 2006). Estos últimos disponen de apartados dedicados a la interfaz de consulta de bases de datos en los cuales se destacan los principales aspectos que debe cumplir una buena página de este tipo. A partir de estos precedentes, se presentó una propuesta de indicadores sobre bases de datos (Abadal, 2002).

2.3.1. Qué es una interfaz de consulta

Una interfaz de consulta es un conjunto de elementos de software y de hardware que sirve para establecer la comunicación entre personas que buscan información y uno o más sistemas de recuperación de información.

Hay otras dos definiciones complementarias que expresan de forma más precisa esta aproximación. En la primera de ellas, Marchionini se refiere a la interfaz desde un punto de vista conceptual, y la vincula con el proceso de búsqueda, en general:

> "La interfaz debe proporcionar un mapeo (*mapping*) robusto entre el contenido de la base de datos y las representaciones conceptuales que el buscador de información manipula." (Marchionini, 1995: 39)

Por tanto, Marchionini pone el énfasis en que la interfaz sirve para establecer la comunicación entre personas con necesidades de información y un sistema de recuperación de información.

En la segunda, Marti Hearst realiza una aproximación más pragmática, precisando con más detalle cuál es la función concreta, los objetivos, de una interfaz de consulta y la vincula con las fases del proceso de búsqueda:

Proceso de búsqueda

Comprensión (definición del problema), planificación (selección de un sistema de búsqueda, formulación de una pregunta, ejecución de la búsqueda), evaluación y uso.

> "La interfaz de usuario debe ayudar a comprender y expresar las necesidades de información. Debe ayudar también al usuario a formular sus preguntas, seleccionar entre las fuentes de información disponibles, entender los resultados y seguir el progreso de su búsqueda." (Hearst, 1999: 257)

Como se ha descrito anteriormente, nuestro objetivo consiste en determinar cuáles son los elementos que han de formar parte de la interfaz de consulta.

En nuestro caso partiremos de las clases de modelos de página que deben elaborarse para que la aplicación funcione de manera adecuada como interfaz de consulta. Los agruparemos de la siguiente manera:

- Consulta.
- Lista de resultados.
- Visualización del documento completo.
- Otras páginas (información general, ayuda, etc.).

Así pues, vamos a determinar cuáles son los elementos básicos que han de estar presentes en cada una de las cinco clases de páginas antes citadas para contribuir a facilitar el proceso de recuperación de la información por parte de los usuarios.

Tabla 6. Elementos de una interfaz de consulta a bases de datos

Página	Componentes
Consulta	• Identificación de la página o base de datos. • Niveles: simple, avanzada, índices. • Especificación de la base de datos (o del fondo o colección o subsede web). • Sistema de recogida de información del usuario. • Acotación de la búsqueda a un campo o conjunto de campos. • Utilización de los operadores (booleanos y otros). • Visualización de los índices. • Informaciones breves para ayudar en la consulta. • Elección de la forma de presentación de los resultados: – Formato de visualización. – Número de registros a visualizar. • Elección del sistema de ordenación de los resultados. • Botones para la ejecución de acciones. • Registro de las búsquedas realizadas (historial). • Acceso multilingüe. • Navegación entre páginas de la interfaz. • Datos identificativos (productor, fecha, lugar, etc.).
Lista de resultados	• Identificación de la página o base de datos. • Información sobre el término de búsqueda y los resultados obtenidos. • Lista con la descripción básica de los documentos. – Estructura. – Inclusión del nombre del campo. – Casilla de selección. • Indicación de tipo de documento (objeto). • Agrupar los resultados por categorías. • Forma de presentación de los resultados. – Formato de visualización. – Número de registros a visualizar. • Información sobre errores o ausencia de resultados. • Opciones de gestión de los registros o documentos. • Elección del sistema de ordenación de los resultados. • Reformulación de la búsqueda. • Posibilidad de encontrar documentos similares. • Navegación entre registros de la base de datos. • Avance y retroceso en las páginas de resultados. • Navegación entre páginas de la interfaz.

Página	Componentes
Visualización de los documentos o registros	• Identificación de la página o base de datos. • Indicación del número de registro que se está visualizando. • Opción de cambio de formato de visualización. • Resaltar los términos de búsqueda. • Distintas resoluciones. • Navegación entre registros de la base de datos. • Avance y retroceso entre los registros seleccionados. • Navegación entre páginas de la interfaz.

2.3.2. Consulta

Las páginas de consulta contienen los formularios que tienen por objetivo facilitar la recuperación de la información contenida en la base de datos. Su función es permitir que el usuario formule su necesidad de información (una de las fases fundamentales del proceso de búsqueda) y es por ello por lo que contendrá diversos recuadros de texto para que se puedan introducir los términos de búsqueda, así como también incluirá los operadores de búsqueda disponibles.

Ahora bien, atendiendo la recomendación de adaptarse al nivel del usuario, ya sea experto o principiante, sería deseable que se pudiera disponer, al menos, de dos tipos de página de consulta: una consulta simple, con pocas opciones de búsqueda, y una consulta avanzada, en la cual se puedan usar todos los operadores y, además, combinar diversos términos. Por otro lado, también es recomendable poder consultar los índices de campo y acceder por categoría temática.

Figura 5. Ejemplo de página de consulta simple (*JStor*)

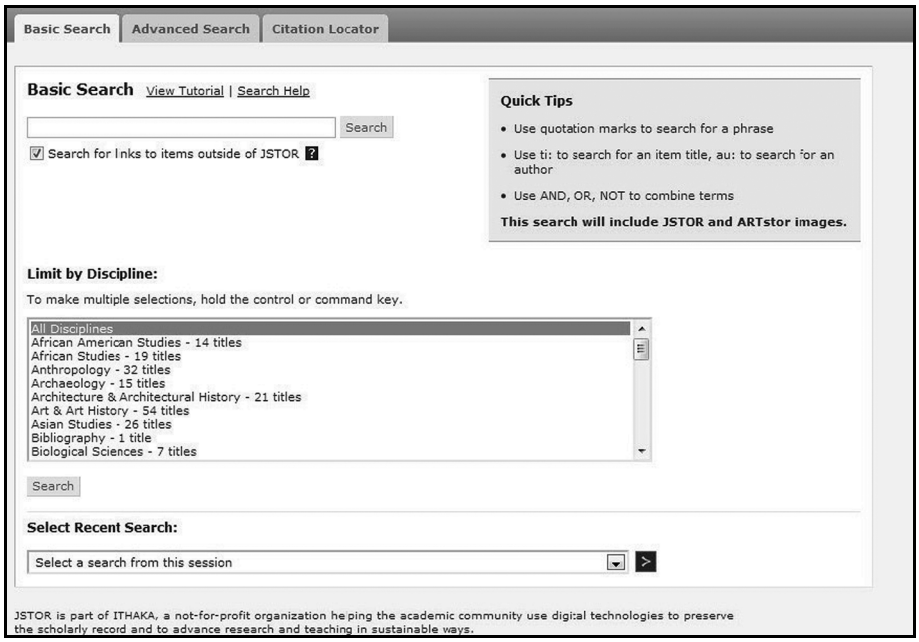


Figura 6. Ejemplo de página de consulta avanzada (*JStor*)

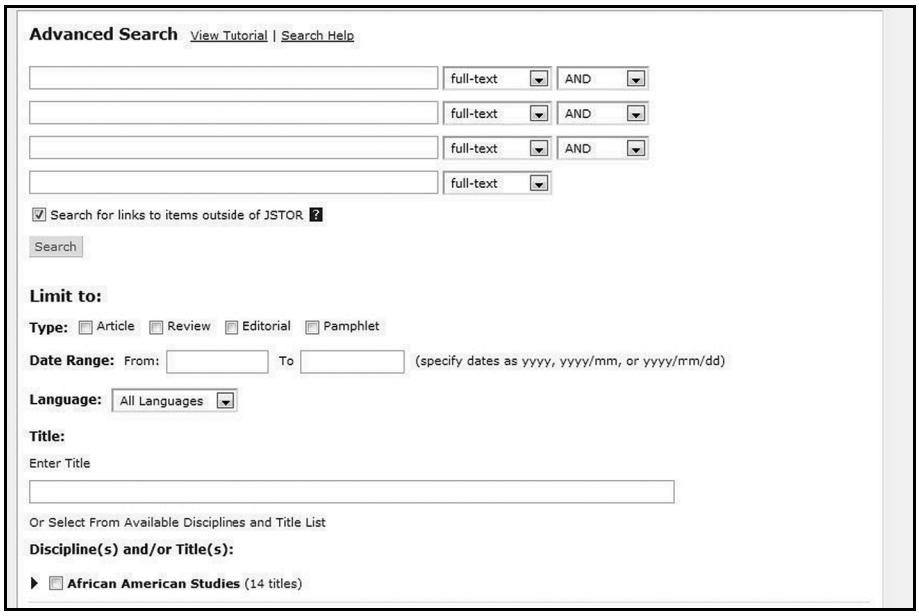


Figura 7. Ejemplo (vista parcial) de página de consulta avanzada de un buscador académico (*Scirus*)

Como hemos detallado en la tabla 6, los elementos de una página de consulta pueden ser bastante numerosos y su composición puede generar confusión si no están bien estructurados. Es por ello conveniente establecer una serie de zonas o áreas ordenadas jerárquicamente para facilitar la secuencia de acciones que sigue un usuario cuando formula una pregunta a una base de datos. En este sentido, hay que resaltar de forma especial, el sistema de recogida de datos del usuario (formulación de la pregunta) y, en segundo lugar, las especificaciones de visualización, ya que el usuario puede tener un interés especial en escoger algunas opciones relacionadas con las características de visualización de

la lista de resultados: cuántos registros se presentarán, en qué formato, en qué orden aparecerán, etc.

2.3.3. Lista de resultados

La primera respuesta del sistema a una consulta expresada por el usuario debe ser una página con la lista que contiene la información básica de los documentos o registros que satisfacen la pregunta, es decir, que son relevantes a la necesidad de información. El objetivo de esta página debe ser presentar una visión global de los resultados y facilitar al usuario la valoración del interés de cada documento a partir de su descripción resumida.

La visualización de los resultados se puede presentar de forma textual (alfanumérica), con la información de los campos que se visualizan uno detrás de otro o dentro de una tabla para estructura mejor el espacio de respuesta. También se pueden utilizar presentaciones de carácter gráfico.

La selección de los registros que son de mayor interés para el usuario es más compleja cuanto mayor es el número de documentos recuperados. En este punto es muy útil disponer de sistemas de ordenación de los resultados basados en la la relevancia u otros criterios libremente configurables por el usuario (fecha, autor, título, etc.).

Figura 8. Ejemplo de página de resultados (*ISI*)

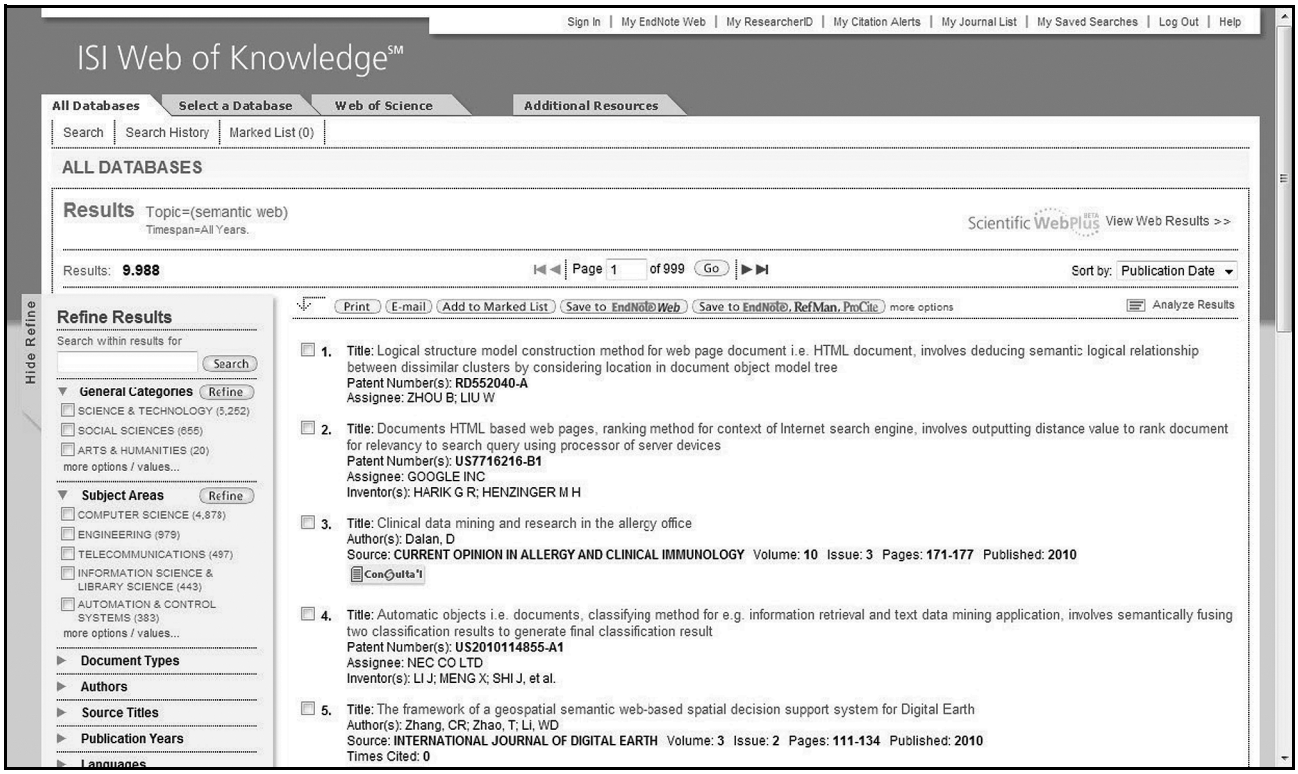
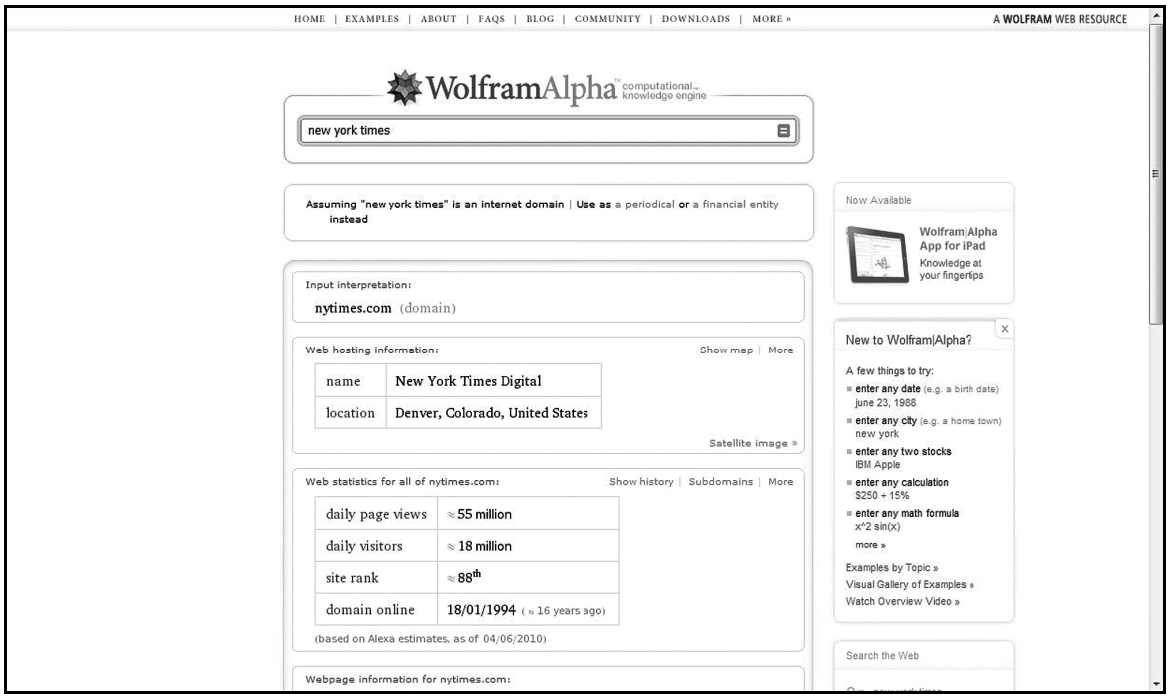


Figura 9. Ejemplo de página de resultados de tipo tabular (*Wolfram Alpha*)

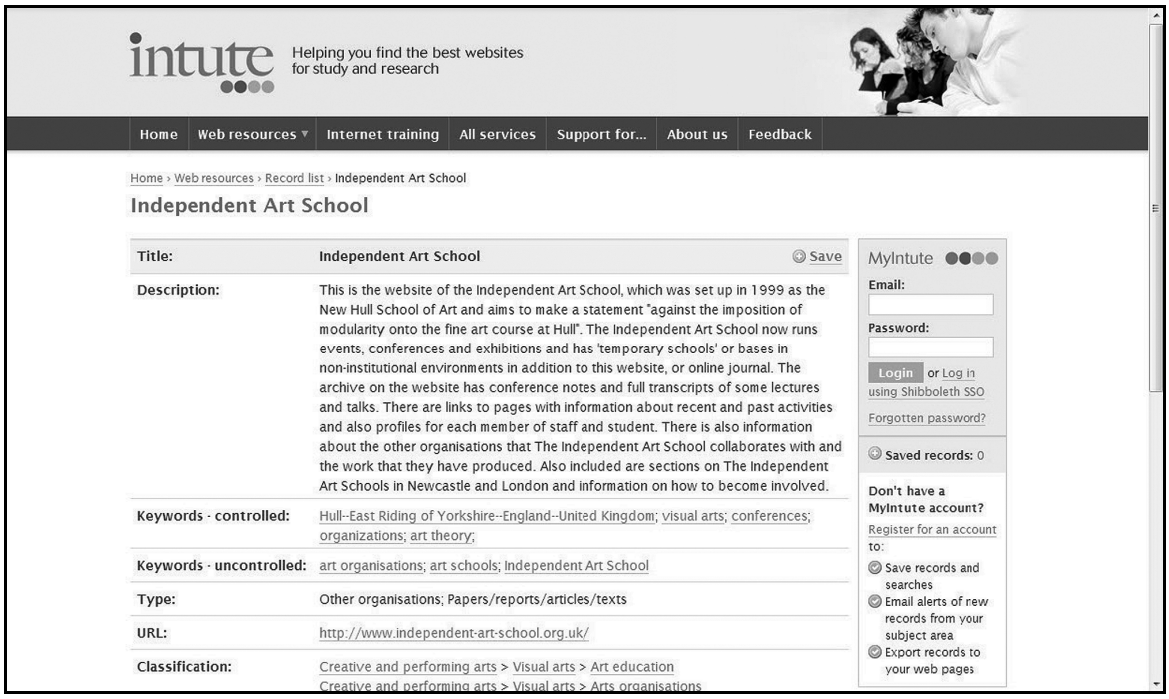


La unidad de información de la lista de resultados acostumbra a ser el documento (página web, PDF, etc.) o el registro (metadatos). En este último caso, cuando buscamos en bases de datos estructuradas.

2.3.4. Visualización de los documentos (o registros)

Desde la página de lista se debe pasar a otra página que incluya la visualización del documento unitario solicitado. Este documento puede ser de tipo textual, gráfico o sonoro o combinar alguno de estos tipos de información. Además, se puede pedir la visualización de la referencia o de los metadatos.

Figura 10. Ejemplo de página de documento, en este caso un registro con metadatos (*Intute*)



2.3.5. Otras páginas

El grupo de páginas descritas anteriormente constituye el núcleo fundamental de la interfaz de consulta. De todas formas, existen otras páginas que las complementan y entre las cuales destacamos las siguientes:

1) Descripción general del contenido

Esta página informa al usuario sobre el ámbito geográfico, temático y lingüístico de la base de datos. Además incluye datos sobre su estructura (campos, etc.), número de registros, etc. Los datos que proporciona han de permitir contextualizar el contenido de la base de datos, así como mostrar su alcance.

2) Ayudas

En este apartado se incluyen, por un lado, los textos que informan al usuario sobre el funcionamiento de la aplicación (es decir, cómo hay que realizar las consultas, cuáles son las opciones disponibles del sistema, etc.) y, por otro lado, los mensajes de ayuda y de error que el sistema va facilitando al usuario a medida que éste va realizando sus acciones.

Nota

Como es bien sabido, las características fundamentales que ha de cumplir el sistema de ayuda son las siguientes: fáciles de localizar, bien organizadas y contextualitzadas.

3) Página de identificación (conexión/desconexión)

En algunas aplicaciones es necesario incluir una página que permita al usuario conectarse al sistema por medio de una identificación (login) y una contraseña (password) y, posteriormente, desconectarse.

La falta de inclusión de alguno de los elementos antes citados resta efectividad a la interfaz de consulta y dificulta al usuario las operaciones de acceso y recuperación de los contenidos de la base de datos. En cualquier caso, hay que tener presente dos cuestiones adicionales:

a) Jerarquización

Es evidente que no todas estas funcionalidades tienen la misma importancia, y que hay unas que pesan mucho más que otras. Esta cuestión, sin embargo, ha sido soslayada en la discusión que se ha presentado.

b) Universalidad

Por otro lado, todos estos elementos tampoco son útiles ni necesarios para todo tipo de usuarios. Si consideramos, como mínimo, dos niveles de experiencia entre los usuarios –noveles y expertos– se comprenderá rápidamente que, para los primeros, la mayoría de los elementos se deben presentar ya configurados, sin dejarles libertad de elección para personalizarlos.

2.3.6. Tendencias

Además de los elementos comentados hasta aquí, tenemos que hacer referencia a otros elementos que todavía no están tan generalizados como los anteriores.

1) Filtrado de los resultados por características

Es cada vez más frecuente disponer de la posibilidad de filtrar los resultados encontrados por tipo de documento (libros, artículos, vídeos, etc.), fecha, idioma, autor, etc.

Ejemplos

- Buscador de noticias de *El País*
  Dispone de una búsqueda básica a partir de la cual se pueden limitar los resultados por tipo de artículo, fecha, etc.
- *Google*
  Permite filtrar los resultados por tipo de documento, fecha, etc.
- Catálogo de la UPC (http://bibliotecnica.upc.edu/search)
  Dispone de filtros por idioma, tipo de documento, etiqueta (tag), autor, etc.

2) Presentación gráfica (visual) de los resultados

Pretende hacer más fácil, simple e intuitiva la presentación de los resultados. Se trata de un ámbito en el que se lleva muchos años investigando aunque se han obtenido aún pocos resultados.

Ejemplos

- *Newsmap* (http://marumushi.com/apps/newsmap/newsmap.cfm)
  Aplicado a noticias de actualidad.
- *Liveplasma* (www.liveplasma.com/)
  Consulta de música y cine.
- *WebBrain* (www.webbrain.com)
  Muestra gráficamente las relaciones de jerarquía entre contenidos de páginas.

3) Agrupaciones temáticas de los resultados

En algunos casos, los motores de búsqueda son capaces de agrupar los resultados en base a categorías temáticas generadas automáticamente.

Ejemplos

- *Clusty* (http://clusty.com/)
  Presenta, a la izquierda, los grupos que se han creado automáticamente con los resultados obtenidos (antes se denominaba *Vivisimo*).
- *IBoogie* (www.iboogie.com)
  Presenta los grupos en los que se dividen los resultados en función de la coocurrencia de los términos. Las jerarquías se subdividen.
- *Moot* (www.mooter.com)
  Indica las búsquedas relacionadas. Dispone de un interesante apartado pregunta-respuesta.

4) Herramientas de descubrimiento: los nuevos OPAC de bibliotecas

Los catálogos de biblioteca están incorporando interfaces de consulta que intentan reproducir los modelos de funcionamiento de las sedes web más exitosas, conocidas y más valoradas por los usuarios (es decir, *Google* y *Amazon*, por poner dos ejemplos). Estas interfaces permiten una consulta conjunta en las diversas colecciones de la biblioteca (catálogo, repositorio, colecciones suscritas) y no sólo del catálogo como era habitual hasta hace poco. También permiten el filtrado de resultados (por temática, autor, tipo de documento, etc.) y facilitan la visualización de las portadas. En inglés se les llama *discovery tools*.

OPAC

Online public access catalog.

Ejemplos

- *Aquabrowser*
  Ejemplo: Queens library (www.queenslibrary.org/)

- *Encore*
  Ejemplo: Michigan State University (www2.lib.msu.edu/)

3. Metodología para la creación de bases de datos documentales

El principio general de creación de sistemas de información indica que todo proyecto comienza siempre por un diseño lógico y que, una vez aprobado éste, se procede al diseño físico o implantación, en un proceso que es tan interactivo como lineal, ya que la fase de diseño, por ejemplo, puede obligar a repensar aspectos de la fase de análisis.

El aspecto importante aquí es que la metodología nos dice claramente que el proceso de creación de una base de datos debe ir siempre desde los aspectos lógicos hacia los aspectos físicos, y no al revés como suele suceder, ya que en la práctica existen muchas formas de violar ese principio general a causa de malos hábitos de trabajo.

Así pues, el proceso de creación de un sistema de información debe ajustarse siempre al siguiente ciclo de vida:

1) Análisis.
2) Diseño.
3) Implantación.

Otra forma de enfocar un proyecto de desarrollo es indicar que la dirección del diseño debe proceder de lo conocido a lo desconocido, y no al revés, como sucede cuando se desea visualizar el sistema de información antes de conocer el sistema de actividades humanas y el sistema de conocimiento.

Finalmente, y por la misma razón, la dirección del diseño debe ir de lo general a lo específico y de los aspectos lógicos a los aspectos físicos, y nunca al revés, es decir, nunca se debe empezar a discutir o a considerar cuestiones concretas (¿cómo se imprimirá la información?) o físicas (¿qué tamaño tendrán las estanterías de los documentos?) antes de plantear las cuestiones generales (¿cuál es el propósito de la base de datos?) o lógicas (¿qué entidades formarán parte de la base de datos?). El siguiente cuadro sinóptico sintetiza estas ideas:

Cuadro 1. Dirección del diseño en el proceso de creación de un sistema de información

- De lo conocido a lo desconocido.
- De los aspectos lógicos a los aspectos físicos.
- De lo general a lo concreto.

En cuanto al proceso de creación, cada una de las tres fases enunciadas antes (análisis, diseño, implantación) puede dividirse en cuantas subfases sean necesarias según el proyecto concreto y la clase de sistema que se está diseñando.

En el caso de una base de datos documental, las dos primeras fases se pueden subdividir en otras dos subfases (a y b). La fase de implantación puede subdividirse en cinco subfases (a, b, c, d y e). Nuevamente debe indicarse que tales divisiones tienen siempre algo de arbitrario. Aquí se hace una propuesta concreta, pero pueden ser válidas otras formas de dividir el proceso de creación. En concreto, en esta metodología se propone la división de fases que muestra el cuadro siguiente:

Cuadro 2. Proceso de creación de una base de datos documental

1. Análisis

1a. Análisis de la empresa u organización (sistema de actividades humanas) y de su entorno.

1b. Análisis de los objetos candidatos a ser registrados (sistema de entidades registrables).

2. Diseño

2a. Diseño del modelo conceptual.

2b. Determinación del tratamiento documental (descripción, análisis e indexación documental...).

3. Implantación

3a. Selección del soporte informático *(software* y *hardware)* de acuerdo con los requerimientos expresados en el modelo conceptual de la base de datos producido en la fase 2a y de acuerdo con los requerimientos expresados en 2b.

3b. Elaboración del presupuesto y del calendario de implantación.

3c. Instalación, pruebas de rendimiento y re-elaboración, en su caso, de los puntos previos de este proceso de creación.

3d. Elaboración del libro de estilo de la base de datos.

3e. Carga de datos, formación de usuarios y promoción del producto.

Aunque expresado en fases y enumeradas secuencialmente el proceso de creación parece estrictamente lineal, en realidad también tiene mucho de iterativo, porque aunque siempre se empieza por la fase de análisis y se sigue con la de diseño, llegados a la fase 2b, por ejemplo, es posible que el diseñador desee considerar de nuevo algunos aspectos de 2a o que necesite aclarar mejor algunas cuestiones de 1b, etc.

En este sentido, debe hacerse notar que la metodología no excluye totalmente el procedimiento de ensayo y error, como ya se advirtió, sino que lo integra de un modo controlado para refinar el producto.

En particular, es prácticamente imposible producir un modelo conceptual correcto en el primer intento, y la experiencia indica que lo más probable es que el modelo elaborado en los puntos 2a y 2b se tenga que rehacer más de una vez, por lo menos en alguno de sus aspectos, principalmente a la vista de las primeras pruebas de rendimiento (3c).

Naturalmente, debe llegar un momento en el cual el diseñador dé por finalizado el proceso, pero la cuestión de cuántas veces conviene iterarlo antes de darlo por bueno no puede establecerse *a priori,* sino que, antes bien, es una cuestión sensible al contexto y que debe decidir el diseñador en cada caso.

De todos modos, es importante que se llegue a la fase de implantación con un modelo lo más sólido posible porque a partir de tal fase ya no resulta tan fácil reconsiderar el proyecto, por lo menos no sin pagar algún precio, de manera

que el punto 3c debería considerarse el punto de despegue; de alguna manera, el punto de no retorno del proyecto.

La fase de implantación puede llevarla a cabo un equipo distinto del que hizo el diseño. De hecho, en algunas empresas, sobre todo en empresas medianas y grandes, puede ocurrir que la fase de implantación corra a cargo del departamento de informática, aunque el análisis y el diseño lo haya hecho el de documentación. En empresas pequeñas, lo más habitual es que todo el proceso lo ejecute un mismo equipo o una misma persona.

Cada una de las fases precedentes (análisis, diseño e implantación) tiene unos objetivos, debe producir unos resultados concretos y utilizar unas herramientas determinadas.

3.1. La fase de análisis

El objetivo de esta fase es conocer bien aquella parte del mundo real, al que denominamos *sistema objeto*, que justifica y requiere la creación del sistema de información; en este caso, una base de datos.

A efectos de análisis, el sistema objeto se considera dividido en:

- **Un sistema de actividades humanas (SAH):** la empresa, organización, sistemas social, etc., que necesita o justifica la base de datos.

- **Un sistema de entidades registrables (SER):** las cosas, personas o conceptos que estarán representados en la base de datos.

Por lo tanto, y dado que las características del sistema de actividades humanas (SAH) determinarán las características de la base de datos, deberán conocerse lo mejor posible antes de iniciar cualquier actividad de diseño.

El sistema de actividades humanas (SAH) se refiere a la organización, la empresa o, en términos generales al sistema social –es decir, un sistema formado por personas y cosas– que justifica o exige la existencia de la futura base de datos. En esta organización social desarrollan sus actividades los futuros usuarios que necesitarán que exista un sistema de información. En ocasiones, puede ser conveniente considerar que, a su vez, dentro del SAH podemos distinguir entre el poseedor o propietario del sistema y los usuarios o beneficiarios del mismo (Checkland, 1981).

Por ejemplo, si pensamos en el OPAC (catálogo online) de una biblioteca universitaria como en un sistema de información, entonces el sistema objeto al cual modela es la universidad de la que forma parte, la cual necesita a la biblioteca (así como otros recursos documentales) para sus actividades de creación

y difusión del conocimiento. ¿En qué sentido, entonces, el OPAC de la biblioteca modela en alguna forma a la universidad? En el sentido en que el lenguaje documental con el cual describe los documentos, la propia selección de los documentos que adquiere, los procedimientos de trabajo, los servicios que presta, etc., son un reflejo de las características de la universidad.

Si consideramos ahora la base de datos de una empresa periodística, la propia empresa periodística es el SAH del sistema objeto, pero el público interesado en la consulta de esa base de datos formará parte también del SAH, en este caso, como beneficiario del sistema.

Dado que el entorno siempre influye en el sistema, a veces de forma decisiva, los diseñadores de la base de datos también deberán conocer las características del entorno de la empresa (o del SAH, en términos más abstractos).

Por su parte, el conjunto de cosas, entidades o documentos que deberán ser descritos y representados en la base de datos forma el llamado *sistema de entidades registrables* (SER). Cuando pensamos en una base de datos documental, es normal pensar en documentos (p.e., en documentos impresos), pero desde un punto de vista abstracto esto es inexacto. En primer lugar, en rigor, una base de datos contiene representaciones de entidades y no necesariamente las entidades en sí mismas (pensad en una base de datos de patrimonio arquitectónico). En segundo lugar, en una base de datos documental podemos tener los siguientes tipos de *entidades representadas*:

a) Cosas: documentos en papel (bases de datos bibliográficas), films (bases de datos de cinematografía), obras de arte (bases de datos de museos), monumentos (bases de datos de patrimonio arquitectónico), etc.

b) Personas: datos biográficos de personajes históricos o de personalidades contemporáneas, cargos de la Administración, etc.

c) Conceptos: ideas y teorías (bases de datos de enciclopedias y diccionarios).

Por tanto, lejos de limitarse a documentos impresos como su único objeto, las bases de datos documentales pueden contener representaciones de un número ilimitado de clases de cosas. A estas posibles clases de cosas susceptibles de estar representadas en una base de datos las denominamos en el argot técnico *entidades*. Por tanto, además de considerar a la empresa (el SAH), en el proceso de diseño hemos de considerar también al conjunto de entidades que deberemos representar en la futura base de datos (SER).

Los conceptos *objeto* y *entidad* se definen en el apartado "Algunos conceptos básicos" del módulo "Sistemas de base de datos".

En el caso de la base de datos de una empresa periodística, por seguir con un ejemplo ya mencionado, el SER consistirá en las informaciones de actualidad que publica esa empresa, sin perjuicio de otros tipos de entidad. Por ejemplo, una de las agencias de noticias más importantes de nuestro país, la Agencia

EFE, produce bases de datos no solamente sobre noticias de actualidad sino sobre personajes (biografías), sobre organismos y legislación de la Unión Europea (directorio, disoposiciones legales), etc.

Con los dos principios fundamentales anteriores se dispone ya de un mínimo aparato conceptual que permite iniciar la discusión de los otros elementos de la metodología. Se observará que algunas herramientas del aparato instrumental, tal como el *modelo E/R* (*entidad-interrelación*) incluyen también aspectos conceptuales. En realidad, es en buena parte arbitrario decidir qué elementos pertenecen al aparato conceptual y qué elementos pertenecen al procedural o al instrumental. Aquí se he hecho una elección concreta, pero probablemente son posibles otras interpretaciones.

El resultado de esta fase de análisis es una descripción textual que suele denominarse *informe de funciones* o *informe de oportunidad,* y que debe incluir, como mínimo, los siguientes aspectos:

1) Propósito y objetivos de la empresa u organización (SAH).
2) Propósitos y objetivos de la futura base de datos o sistema de información.
3) Identificación y características principales de las entidades registrables (SER).
4) Sistemas similares ya en funcionamiento, si es el caso.

La herramienta principal aquí es la realización de entrevistas con representantes de la empresa u organización (SAH), así como el análisis de cualquier documentación sobre la empresa que pueda aportar una comprensión global del sistema. Entre tales documentos podemos citar organigramas, documentos fundacionales, memorias, etc. Por supuesto, serán básicas las entrevistas con los futuros usuarios del sistema, así como con la persona o los representantes de la empresa que hayan realizado el encargo.

En muchos casos, nos podremos beneficiar de un estudio de tipo *benchmarking*. Si existen ya otras bases de datos similares, será conveniente proceder a algún tipo de estudio o análisis.

El resultado de esta fase debe consistir en la identificación clara y sin ambigüedades no solamente de las cosas, las personas o los conceptos (entidades) sobre los cuales la base de datos deberá mantener información, sino también de las funciones y beneficios que se espera de la futura base de datos.

Ejemplo

Si el encargo consiste en el diseño de una base de datos documental para un museo, sería coveniente programas visitas a algun museo que ya disponga de ellas. En último extremo, casi siempre podremos encontrar modelos de bases de datos en funcionamiento a través de Internet.

El *informe de funciones* debería ser aprobado por la persona que realiza el encargo de la base de datos, como forma de asegurarnos de que las dos partes: la empresa y los diseñadores de la base de datos comparten las mismas ideas básicas.

3.2. La fase de diseño

El propósito de la fase de diseño es obtener un modelo conceptual de la base de datos y que contenga también una propuesta de tratamiento documental. El primer elemento contiene las indicaciones necesarias para orientar el proceso de implantación. El segundo elemento establece criterios y orientaciones sobre el proceso de descripción y de representación del contenido semántico de las entidades de las que tratará la base de datos.

Los dos componentes mencionados son el resultado de la fase de diseño y deben ser aprobados también por quien encargó el proyecto, antes de que puedan servir como guías de implantación. Por tanto, el modelo conceptual no sólo debe ser acertado, sino que, además, debe parecerlo.

El *modelo conceptual* debe contener, por lo menos, los siguientes elementos:

1) Objetivo y propósitos de la base de datos con identificación de los usuarios del sistema (puede repetir partes esenciales del informe de funciones, si es necesario).

2) Una definición de los ámbitos o contenidos temáticos de la base de datos.

3) Una identificación de las entidades representadas en la base de datos.

4) El diccionario de datos.

5) Una descripción funcional que debe considerar los siguientes elementos:

a) Qué clase de información se tratará y cómo entrará la información en el sistema.

b) Qué procesos documentales se llevarán a cabo.

c) Qué servicios y productos generará el sistema, o a qué aplicaciones podrá dar soporte.

6) Una propuesta de tratamiento documental.

Dado su contenido, las herramientas para producir el documento anterior son, entre otras, las siguientes:

1) El informe de funciones elaborado previamente.
2) El modelo entidad-interrelación.
3) El diccionario de datos.

Ámbito temático

El *ámbito temático* de la base de datos es el conjunto de los temas o entidades sobre los que mantiene información la base de datos. Como todo ámbito, puede definirse por extensión o por comprensión. Por tanto, puede ser tan breve como el nombre de una o más disciplinas científicas, por ejemplo, el ámbito de la base de datos LISA Plus son las ciencias de la documentación. O puede consistir en una frase, por ejemplo, el ámbito o contenido de la base de datos TESEO se enuncia diciendo que está formado por las tesis doctorales publicadas por universidades españolas.

3.2.1. El diccionario de datos

El diccionario de datos (*data dictionary*) es una herramienta que ayuda al diseñador de una base de datos a garantizar la calidad, la fiabilidad, la consistencia y la coherencia de la información introducida en la base de datos, y que condiciona decisivamente el rendimiento y la calidad global del sistema de información.

Consiste en la lista detallada de cada uno de los campos de la base de datos con la especificación, para cada uno de ellos, de un conjunto de parámetros que incluyen, como mínimo, los siguientes aspectos:

1) Etiqueta.
2) Dominio.
3) Tipo.
4) Indización.
5) Tratamiento documental.
6) Lengua.
7) Otros controles de validación u observaciones.
8) Obligatoriedad.
9) Repetibilidad.
10) Instrucciones para la entrada de datos.

Ejemplo

Supongamos, a efectos de esta explicación, una base de datos documental imaginaria sobre noticias de actualidad con sólo tres campos: <Título>, <Descriptores> y <Fecha de publicación>. El diccionario de datos tendría entonces esta forma (el diccionario de datos real tendría más campos):

Etiqueta: Título

Dominio:
Título del documento.
Tipo:
Alfanumérico.
Indización:
Indizado.
Tratamiento documental:
Lenguaje libre.
Lengua:
Lengua del documento.
Controles de validación:
No puede quedar vacío. Si por alguna razón, el documento careciera de título, el documentalista asignará un título descriptivo.
Obligatoriedad:
Obligatorio.
Repetibilidad:
No es un campo repetible.
Instrucciones para la entrada de datos:
Las diversas partes del título se transcribirán de la siguiente forma: *Título: antetítulo: subtítulo*. Los artículos iniciales no se pospondrán. Por ejemplo, "Desacuerdo en Bruselas: Reunión de los ministros de economía: Se cuestiona el pacto de estabilidad".

Etiqueta: Descriptores

Dominio:
Los descriptores deberán obtenerse del tesauro de la base de datos.
Tipo:
Alfanumérico.
Indización:
Indizado.
Tratamiento documental:
Lenguaje controlado.
Lengua:
Del centro de documentación.
Controles de validación:
No puede quedar vacío y sólo admite valores extraídos de una lista de términos autorizados.
Obligatoriedad:
No.
Repetibilidad:
Sí. Pueden asignarse diversos valores a este campo.
Instrucciones para la entrada de datos:
Se asignarán descriptores (esto es, términos de indización) que expresan los conceptos principales contenidos en el documento, según el siguiente principio general: si el artículo contiene *n* conceptos relevantes se asignarán *n* descriptores (hasta un máximo de 20 descriptores por documento). Se seguirá las normas ISO/UNE de determinación de temas de documentos y de asignación de descriptores. Los términos se separarán con ",". Por ejemplo, edición óptica, publicación digital, documentación.

Etiqueta: Fecha de publicación

Dominio:
La fecha de publicación de la noticia.
Tipo:
Fecha.
Indización:
Indizado.
Tratamiento documental:
No procede.
Lengua:
No procede.
Controles de validación:
No admite valores fuera de rango.
Obligatoriedad:
Sí.
Repetibilidad:
No.
Instrucciones para la entrada de datos:
Los datos tienen que introducirse en el formato: dd/mm/aaaa. Por ejemplo, 28/11/2003

Estudiando el ejemplo de diccionario de datos anterior, formado únicamente por tres campos a efectos didácticos, podemos observar cuatro aspectos importantes para el diseño de bases de datos:

1) Que el **dominio**, en el contexto del diccionario de datos, se refiere al conjunto del que un campo puede obtener sus valores.

2) Que el **tipo** se refiere, en cambio, al tipo de dato que admite el campo. Los tipos de datos suelen ser: numérico, alfanumérico, fecha y lógico.

Recordemos que un tipo de dato (*data type*) define un conjunto de operaciones válidas y un rango de valores aceptable. Por ejemplo, el tipo de datos "alfanu-

mérico" define operaciones de comparación de cadenas de caracteres, entre otras, así como cualquier letra de la *a* a la *z* y cualquier número del *0* al *9*, así como cualquier combinación de esos caracteres en palabras, frases, párrafos, etc. En cambio, no admite operaciones aritméticas, aunque admita números. Por el contrario, un tipo de dato "numérico" admite sólo números así como cualquier operación aritmética, etc.

Por su parte, un *campo de fechas* sólo admite fechas en un formato establecido y permite búsquedas por rangos de fecha o por valores superiores o inferiores a una fecha dada. Un *campo lógico* sólo admite uno de dos valores: sí o no; verdadero o falso.

3) Que el **tratamiento documental** establece si se debe utilizar algún lenguaje documental para entrar los valores del campo, como así sucede en el campo Descriptores, donde el diccionario de datos establece que ese campo sólo admite palabras clave autorizadas extraídas de un tesauro de una lista de autoridades.

Lista de autoridades

Permite asegurar para cada campo que cualquier punto de acceso sea único y no se pueda confundir con ningún otro punto de acceso.

4) Que la **lengua** puede ser, o bien la lengua del documento, o bien la del centro de documentación. Eso significa, en el caso de un documento escrito en inglés, que el título estaría en inglés, pero los descriptores en castellano, siempre de acuerdo con el diccionario de datos precedente.

A modo de síntesis, la tabla siguiente recoge los grupos de campos que normalmente encontraremos en una base de datos de tipo documental. Cuando realizamos el diseño del diccionario de datos, es aconsejable chequearlo con esta tabla y comprobar que no olvidamos alguna categoría o grupo de campos:

Tabla 7. Grupos de campos en una base de datos documental

Grupos de campos	Explicación	Ejemplo de campos
De control	Controlan la gestion interna del registro.	Número de registro (ID), fecha de entrada, fecha de modificación, nombre de analista
Descriptivos	Describen las características de las entidades o documentos de la base de datos.	Autor, título, idioma, fecha de publicación, formato, soporte, descripción física
Temáticos	Representan el contenido o tema del documento o entidad representada en la base de datos.	Resumen, descriptores temáticos
De derechos	Indican, en su caso, que restricciones o derechos limitan la utilización del documento o quienes están en posesión de los mismos.	Licencia CC, restricciones y derechos de uso
De ubicación	Indican, en su caso, la ubicación o localización del documento original. Pueden referirse a la ubicación física de un documento (un libro en una estantería), o puede consistir en un puntero informático que abre el documento original en el caso de documentos digitales.	Ubicación recurso físico, ubicación recurso digital

3.2.2. ISBD y modelos canónicos

No deberíamos olvidar que, en documentación, la experiencia previa ha dejado bien sentados cuáles son los atributos de algunas entidades e, incluso, cuál es la forma más conveniente de representarlos. Podemos hablar entonces de situaciones canónicas que han generado un modelo. La mejor herramienta de análisis y de diseño, en tal caso, consiste precisamente en aplicar ese modelo bien conocido y probado.

Ejemplo

Los atributos estructurales de cualquier clase de documento pueden ser adecuadamente modelados siguiendo las normas internacionales ISBD. Esas normas internacionales representan un gran esfuerzo de abstracción para proporcionar un marco general de descripción, válido para cualquier clase de documento, desde una partitura musical, hasta una filmación audio-visual, pasando por un archivo de ordenador, un fonograma o un artículo de revista, de manera que las ISBD constituyen una herramienta de diseño de primera magnitud para cualquier problema documental donde debamos representar documentos.

Sobre el uso de las ISBD, cabe advertir que algunos centros de documentación se han sentido intimidados ante la aparente complejidad de la norma y la supuesta obligación de adoptarla como un todo, incluyendo la prolija puntuación que prescribe y, en tal sentido, se ha argumentado que utilizar la norma ISBD solo tiene sentido en el contexto de las bibliotecas normalizadas, aquellas que necesitan intercambiar registros y que, por tanto, siguen estándares internacionales.

ISBD

(*International standard bibliographic description*) es un estándar que determina la forma y el contenido de la descripción bibliográfica.

Entendemos que tal postura es un error: primero siempre podemos utilizar la estructura de las ISBD como una orientación en el análisis de los documentos convencionales, así como una fuente de inspiración para situaciones más exóticas, independientemente de que incorporemos o no la norma en toda su complejidad, es decir, incluyendo todos los niveles de descripción y todas las prescripciones de puntuación, máxime cuando el hecho de separar zonas mediante campos libera de la necesidad de utilizar la puntuación prescrita.

Además, en caso necesario, debería permitir, (como es el caso de diversos de ellos; por ejemplo, *Inmagic* o *CDS/ISIS*) presentar la salida de los datos en formato ISBD (o en cualquier otro formato), desde el momento en que la estructura repetitiva de los registros permite incorporar instrucciones del tipo: "el valor del campo *título* se transcribe seguido por un punto, espacio y un guión", etc.

3.3. La fase de implantación

Una vez aprobado el modelo conceptual de la base de datos, puede procederse a su implantación, la cual puede seguir el siguiente proceso:

1) Preselección del sistema informático (software + hardware)

A menos que el equipo informático forme parte de las restricciones iniciales, probablemente será necesario examinar varios programas candidatos hasta que exista una razonable certeza de que el programa elegido se ajusta bien a los requerimientos del modelo conceptual.

Para seleccionar el programa más adecuado será necesario contactar con diversas empresas del sector para solicitar documentación sobre las prestaciones de sus programas, presupuestos, etc. Entre los criterios que nos ayudarán a tomar

una decisión deberemos considerar los siguientes (además de otros criterios *ad hoc* según la naturaleza específica del proyecto):

a) Grado de compatibilidad con la plataforma informática de la empresa o corporación.

b) Grado de satisfacción de los requerimientos establecidos en el diseño conceptual.

c) Posibilidades de parametrización y disponibilidad de herramientas de desarrollo disponibles con la aplicación (lenguaje de guiones, herramientas de programación adicionales, etc.).

Lenguaje de guiones

Lenguaje de programación simple usado por el usuario final que utiliza conjuntos de instrucciones por lotes (*scripts*) para interactuar con el sistema operativo, controlar varias aplicaciones, etc.

d) Valoración de otros clientes y usuarios. Base instalada de clientes: ¿pueden proporcionar referencias de otros clientes?, ¿existe un club de usuarios de la aplicación?

e) Utilización de estándares bien establecidos, ya sean *de facto* o *de jure,* y compatibilidad con sistemas abiertos (por ejemplo, compatibilidad con el formato PDF y con el lenguaje HTML; o en otro extremo, compatibilidad con el uso de metadatos y normas como Dublin Core, etc.).

f) Coste económico (presupuesto)

El orden indicado no es significativo: en algunos casos, el punto f) puede ser primordial mientras que el punto a) puede carecer de importancia, etc. En cada proyecto concreto, los responsables decidirán cuál es el orden más adecuado según el contexto. En todo caso, los puntos a) al f) constituyen un buen conjunto de elementos de partida que se deberán considerar en casi cualquier proyecto.

2) Elaboración del presupuesto y del calendario de implantación

a) Una vez seleccionada una aplicación candidata, se procede a la instalación del programa y a una primera implantación de la base de datos aplicando el modelo conceptual creado en la fase de diseño para realizar las primeras pruebas.

b) Si se aprueba finalmente el uso de la aplicación elegida, se procede a la designación de un administrador de la base de datos que, a partir de ahora, será el máximo responsable de la misma y comienza a desarrollar la primera versión de la base de datos según se indica en el punto siguiente.

3) Implementación de los controles terminológicos; por lo menos de aquella clase de controles de los que se tenga la posibilidad de instalación anticipada a la carga de datos (palabras vacías, sinónimos, valores predefinidos, etc.).

4) Realización de pruebas con una colección de prueba (conjunto de documentos o de entidades, candidatos/as a ser representados/as) para compro-

bar la consistencia de los modelos y esquemas de registros detallados en las fases previas y contenidos en el diccionario de datos.

5) Realización de cambios o ajustes según el resultado de las pruebas anteriores, si es el caso, en las definiciones de los campos o en la estructura del registro.

6) Automatización de procesos repetitivos facilidades para dar altas (carga de datos), realizar exportaciones, hacer consultas más frecuentes, etc.

7) Segunda carga de datos con otra colección pequeña de documentos (por ejemplo, con 15 o 20 documentos). En este punto, es conveniente simular todos los procesos que se van a realizar con esta base de datos (consultas, exportaciones, etc.) para obtener la seguridad de que se va por el buen camino. Es normal que en esta fase aparezcan imprevistos: formatos de exportación en los que no se había pensado, tipos de consultas que requieren algún reajuste en los campos, etc.

8) Test de usabilidad. Se puede aplicar ahora un test de usabilidad, encargando a varios futuros usuarios reales de la base de datos (entre tres y cinco son un buen número) que realicen pruebas de uso realistas de la base de datos encargándoles tareas seleccionadas previamente para este estudio y observando cómo las resuelven, además también les pediremos que den su opinión sobre su rendimiento: ¿es lo que esperaban?, ¿falta alguna opción?, ¿es fácil de usar?

Se incorporarán los cambios en el diseño si se detecta alguna insuficiencia y, posiblemente, ya habremos llegado al diseño definitivo. En todo caso, no podemos estar modificando el diseño de manera indefinida y más pronto que tarde deberemos dar por bueno el modelo.

9) Diseño de las vistas de los usuarios y de las carátulas de portada o inicio.

10) Definición de los grupos de usuarios y de otros responsables de la base de datos. Cada grupo de usuarios debe tener privilegios y, a ser posible, vistas diferentes de la base de datos. En este sentido, es conveniente considerar que suele haber, por lo menos, cuatro tipos de personas involucrados en la base de datos, que son los siguientes:

a) El administrador o director de la base de datos. Es la persona que tiene la máxima responsabilidad en la base de datos. Esta figura ya ha sido designada antes.

b) Los documentalistas/analistas. Son quienes realizan el análisis de la información. Suelen ser profesionales expertos en la temática de la base de datos quienes producen resúmenes o asignan descriptores.

c) Los operadores. Son quienes realizan la carga de datos. Operadores y analistas pueden ser las mismas o distintas personas. En este último caso, a veces,

los analistas realizan su trabajo sobre plantillas que después vuelcan los operadores en la base de datos.

d) Los usuarios finales. Son quienes explotarán y utilizarán la información. Puede haber diversas categorías de usuarios finales. Por ejemplo, en Internet es habitual que haya usuarios que solamente pueden hacer consultas, pero no pueden ver los resultados completos, a menos que esten registrados o que abonen una cuota, etc.

11) Inicio del proceso de carga de datos sistemática y de explotación del sistema.

Como es lógico, llegará el momento en el cual tendremos que empezar la carga de datos de manera masiva y sistemática. Para ello, deberemos establecer de manera explícita, clara y sin ambigüedades los siguientes extremos:

a) Rutinas de carga de datos. Quién, como y cuando se hace la carga de datos. Los encargados de realizarla deben ser personal entrenado no solamente en el uso del programa, sino en el conocimiento del diccionario de datos.

b) Rutinas de seguridad. Quién, cómo y cuando se crean las copias de seguridad. En todo caso, deberían hacerse por lo menos dos copias de seguridad y en dos formatos distintos. Una copia de seguridad de los trabajos del día y otra, desfasada respecto a la anterior en uno o más días.

Es conveniente tener copias de seguridad en dos formatos: el formato nativo del programa más un formato fácilmente explotable con otras aplicaciones o bases de datos. Lo más fácil es tener una copia de seguridad en formato ASCII y en un formato tipo "campos separados por tabulador", que entienden muchos programas de base de datos.

c) Evaluación y controles de calidad. Periódicamente estableceremos controles de calidad. Los elementos más típicos en este control son: el control de duplicados y el control de la calidad de la indización. Para ambos tipos de controles, las bases de datos documentales suelen proveer opciones diversas. En muchos programas documentales, por ejemplo, podemos exportar y publicar la lista de descriptores y revisarlos periódicamente. También podemos solicitar la detección de duplicados. El apartado siguiente se dedica a analizar, de forma pormenorizada, todo lo referente a la evaluación y el control de calidad en bases de datos.

Además, según la naturaleza de la base de datos, estableceremos otros tipos de controles adecuados a su contenido, etc.

d) Política de mantenimiento y explotación. Se editará la primera versión del *Libro de estilo de la base de datos*, que incluye los elementos siguientes:

- Versión definitiva del modelo conceptual.
- Normativa de tratamiento documental.
- Política de formación del personal técnico y organización de sesiones de formación de los usuarios finales.

12) Acciones de promoción, en su caso.

3.4. Reflexión

El valor de esta metodología radica, como ya se dijo al principio, en que ayuda a que el producto final sea más resultado del diseño consciente que de las fuerzas ciegas del azar o del ensayo y error, pero, particularmente entendemos que su utilidad aumenta conforme se aplica a situaciones poco canónicas o a situaciones atípicas, como las que el entorno cambiante de la profesión de documentalista introduce en cada momento y, al parecer, tal como el nuevo horizonte de las autopistas de la información y de un futuro mundo digital parece prometer.

Esperamos que, entonces, la aplicación de esta clase de metodologías sirva que los profesionales de este campo puedan demostrar los beneficios de una adecuada formación académica, del trabajo bien realizado y de la planificación, porque en este campo de actividades también es rigurosamente cierto que el éxito se debe invariablemente a "un diez por ciento de inspiración y un noventa por ciento de transpiración".

4. Evaluación de bases de datos

En los apartados precedentes se han tratado cuestiones diversas referidas al diseño, producción y distribución de bases de datos documentales que deberían servir para realizar un producto concreto, es decir, cosas como *ERIC*, la base de datos de artículos de revistas de educación; el archivo de prensa de *El País*; la base de datos de fotografías *AGE Fotostock* o la base de datos de recursos de Internet *Intute*.

Ahora bien, una vez estas bases de datos se encuentran en el mercado, las cuestiones son: ¿es una base de datos competitiva?, ¿en qué posición se encuentra en el mercado respecto de otros productos similares?, ¿satisface los requisitos mínimos de calidad? Este tipo de cuestiones, que interesan tanto al usuario o cliente de la base de datos como al productor de la misma, constituyen el eje de este apartado.

Nuestra aportación aquí consiste en presentar una aproximación que procura ser global e integradora y que se compone de dos partes. En la primera, más extensa, se recopilan y organizan los principales indicadores o criterios que han sido considerados en los estudios realizados hasta el presente para la evaluación de la calidad de bases de datos y que, por tanto, no se circunscribe tan sólo a los sistemas de recuperación de información (SR) sino que también incluye el contenido, es decir, la base de datos propiamente dicha. En la segunda, se describen las principales técnicas de recogida de datos sobre el usuario, un elemento esencial para poder medir aquellos criterios de evaluación que se refieren especialmente a aspectos subjetivos del usuario (satisfacción, utilidad, etc.).

Por otro lado, hay que recordar que este tipo de estudios de evaluación han tomado como objeto bases de datos de tipos muy distintos. En primer lugar, los catálogos de biblioteca en línea, para los cuales se dispone de una normativa de carácter internacional que persigue facilitar el intercambio de registros; a continuación las bases de datos científico-técnicas, con carácter especializado, para las cuales, en cambio, no se dispone de normas de común seguimiento y cada productor aborda como ha creído más conveniente; y finalmente, los servicios de búsqueda en Internet. Un repaso a los estudios de evaluación en bases de datos nos muestra como en el transcurso de los años se han ido presentando análisis distintos de estos objetos. El enfoque que aquí presentamos quiere tener un carácter integrador y, por tanto, es aplicable tanto a catálogos de bibliotecas como a bases de datos especializados o a motores de búsqueda.

4.1. Indicadores (criterios de evaluación)

Para establecer una metodología de evaluación y análisis de la calidad de las bases de datos, hay que determinar, antes que nada, cuáles van a ser los indi-

cadores que se tendrán en cuenta. Este es un requisito imprescindible para cualquier tipo de evaluación que se quiera llevar a cabo sobre cualquier servicio o producto (por ejemplo, evaluación de revistas, evaluación de sistemas de recuperación de la información, etc.).

Diversos son los antecedentes. Podemos empezar haciendo referencia al trabajo realizado por el *Southern California Online User Group* (SCOUG) que, en 1990, en el marco del *Fourth Annual Retreat* dedicado a la medición de la calidad de bases de datos (*Measuring the quality of databases*) estableció diez criterios para la evaluación que han sido la base de muchos estudios posteriores, como los de Wilson (1998), Xie (1998) o Rodríguez Yunta (1998), entre muchos otros, que incluyen propuestas de las cuales se pueden extraer un importante número de indicadores o criterios de evaluación. La popularización de los buscadores web a finales de 1990 y principios del siglo XXI matiza la importancia de algunos de estos criterios, que están más bien pensados para el contexto de las bases de datos científico-técnicas, aunque en líneas generales son fácilmente adaptables a este contexto.

A fin de organizar un poco la larga lista de indicadores a la que se debe hacer referencia, los hemos agrupado en tres grandes ámbitos:

1) La base de datos (el **contenido**): incluye todo lo que se refiere a la calidad de la información contenida en la base de datos, su indización, los documentos escogidos, etc.

2) El sistema de recuperación (el **continente** o ***software***): incluye todo lo referente al programa de recuperación y sus prestaciones y también a la interfaz de consulta (no se puede olvidar que una misma base de datos puede estar disponible en diferentes sistemas de recuperación de la información)

3) La gestión y administración de la base de datos (los **servicios adicionales al usuario**): incluye todo lo referente a la documentación sobre la base de datos, los procedimientos de cobro, los precios, y las facilidades otorgadas a los usuarios.

Tabla 8. Criterios de evaluación

Contenido de la base de datos	Grado de exactitud y precisión Errores gramaticales y mecanográficos Errores de omisión Fiabilidad de los datos Registros duplicados Alcance y cobertura Grado de cobertura o alcance temático Cobertura geográfica y lingüística Grado de inclusión Estructura Tamaño Nivel de crecimiento Actualización Grado de actualización Periodo de actualización Consistencia Consistencia de la catalogación Consistencia en el análisis de contenido

Sistema de recuperación	Prestaciones del lenguaje de interrogación Precisión Exhaustividad Tiempo de respuesta Utilidad Formatos de visualización Usabilidad de la interfaz
Gestión de la base de datos	Documentación sobre la base de datos Atención al usuario Precio y sistema de facturación Sistema de distribución

4.1.1. Contenido de la base de datos

En este apartado se evalúa la materia prima fundamental para asegurar la calidad de una base de datos. De poco servirá disponer de un potente SRI con muchas prestaciones, o de una gestión y administración muy eficaz si los contenidos son pobres y de poca calidad. Jacsó (1997) presenta un valioso artículo de revisión bibliográfica que repasa las principales publicaciones que se refieren a los aspectos de calidad del contenido: precisión y fiabilidad, alcance y cobertura, actualización, etc.

La evaluación del contenido empieza cuando el productor de la base de datos selecciona y analiza la información, y afecta a aspectos que son de su competencia directa.

1) Grado de exactitud y precisión

Se trata de un conjunto de indicadores que hacen referencia a problemas relacionados con la falta de precisión en la entrada de datos: errores gramaticales y mecanográficos de las palabras, ausencia de información en los campos de la base de datos, fiabilidad de los datos o duplicación de información. En definitiva, agrupa un conjunto de cuestiones relacionadas con la "calidad de los datos".

a) Errores gramaticales y mecanográficos

Se refiere a la presencia de errores ortográficos, sintácticos (problemas de concordancia de género, de número, etc. de las palabras) y de mecanografía.

Los errores gramaticales y de mecanografía afectan directamente a la recuperación de la información, ya que podemos dejar de obtener una información pertinente o recuperar otra no lo sea. Son especialmente preocupantes en áreas como las finanzas, o la información médica o jurídica, en las cuales se pueden tomar decisiones muy importantes sobre la base del contenido de la información recuperada.

b) Errores de omisión

Los registros incompletos (p.e. falta la fecha de publicación, el idioma, el tipo de documento, etc.) constituyen errores de omisión y pueden ser fácilmente detectables y prevenibles. Para detectarlos, se pueden usar rutinas automáticas.

c) Fiabilidad de los datos

Se refiere a la exacta correspondencia del contenido de los registros con los documentos a los que representan. En la de base de datos de recursos web, hay que estar atentos a la presencia de enlaces inexistentes o sin actualizar.

d) Registros duplicados

El origen de este problema son los errores de exactitud que se han descrito en este mismo apartado y también las inconsistencias (ved el punto 4: "Consistencia"). A efectos de la recuperación, la existencia de registros repetidos dificulta la consulta de la base de datos porque aumenta innecesariamente el número de resultados.

2) Alcance y cobertura

a) Grado de cobertura o alcance temático

Toda base de datos está especializada en una o diversas áreas temáticas. La cobertura indica la proporción de fuentes de esta materia concreta que están disponibles en la base de datos. El sistema que se puede utilizar para medir este indicador consiste en determinar la proporción de fuentes de información consideradas de máximo interés para un área temática que forman parte del conjunto de fuentes de información vaciadas por la base de datos.

b) Cobertura geográfica y lingüística

Teniendo en cuenta al ámbito geográfico y las lenguas se puede valorar el internacionalismo de la base de datos, poco destacable en las anglosajonas. Para el usuario no anglosajón este factor acostumbra a tener un notable valor ya que también le interesa localizar documentos en su idioma.

c) Grado de inclusión

Se refiere a la presencia de determinados tipos de documento: sólo artículos de revista o también monografías, congresos, patentes, normas, etc.

d) Estructura

Se refiere al número de campos definidos y recuperables. Hay que ver si sólo afectan a la parte descriptiva (autor, título, etc.) o también al contenido (descriptores, clasificación, resumen, etc.).

e) Tamaño

El número de registros (o la cantidad de páginas web) indizadas es un criterio del cual acostumbran a alardear los grandes productores de bases de datos y que impresiona de forma notable a los usuarios. Aunque se trata de un indica-

dor importante, hay que interpretarlo de forma adecuada y con la perspectiva correcta. ¿De qué nos sirve tener millones de registros o de páginas web si no se incluyen las fuentes más prestigiosas?

f) Nivel de crecimiento

Se mide el número de registros nuevos por año. Este número tendría que ser igual o parecido al número de documentos producidos de las materias que trata la base de datos en el mismo período de tiempo.

3) Actualización

a) Grado de actualización (o actualidad de la información)

Se mide el tiempo que pasa entre que un documento está disponible y su inclusión en la base de datos.

b) Periodo de actualización

Se refiere a la periodicidad con la que se actualizan los registros de la base de datos.

4) Consistencia

Se refiere a la característica o propiedad que poseen los registros de una base de datos que están confeccionados uniformemente. Para conseguirlo es necesario aplicar estricta y homogéneamente un conjunto de normas comunes.

a) Consistencia de la catalogación (o de la descripción)

Mide el grado de coherencia en lo que respecta al análisis formal de los documentos, es decir, a su descripción bibliográfica (asignación correcta de campos y subcampos) y a la elección de puntos de acceso, siendo esto último lo que acarrea problemas en la recuperación. Los términos que constituyen puntos de acceso fundamentales a los registros (autor/es, título de la revista, materia, etc.) tienen que estar normalizados, es decir, han de asignarse de forma homogénea y consistente porque, de otro modo, limitan las capacidades de recuperación.

La utilización de listas de validación (dominio del campo) es muy útil para asegurar la consistencia en las entradas de autor, título de revista, materia, etc. Por otro lado, es fácil preparar pruebas de consistencia de la información que se ha introducido en un mismo campo. La más sencilla es generar los índices del campo para que, de esta forma, se pueda comprobar si se ha realizado correctamente el control del dominio.

b) Consistencia en el análisis de contenido

Se refiere a la coherencia en la asignación de términos de indización y de códigos de clasificación para asegurar que se utilizan siempre los mismos cuando queremos representar una temática idéntica.

4.1.2. Sistema de recuperación (o SR)

La mayoría de los elementos que se pueden incluir en este apartado ya han sido analizados en los apartados 1 y 2.3 de este mismo módulo. Es por ello que vamos a realizar un breve repaso. Estos indicadores están relacionados con el proceso de búsqueda y visualización de los resultados y dependen, por tanto, de las características del programa informático utilizado.

De hecho, estos criterios son totalmente independientes de los anteriores. Se da el caso de bases de datos que son consultables por distintos SR; con lo cual, si se evalúan, hay que precisar cuál es el programa de recuperación que se tiene en consideración.

Bases de datos con programas diferentes

Medline puede ser un ejemplo. El mismo contenido es accesible en la National Library of Medicine (www.nlm.nih.gov) y también en el portal brasileño Scielo (www.bireme.br/bvs/E/ebd.htm), con programas informáticos distintos.

1) Lenguaje de interrogación

Las principales funcionalidades a considerar son el uso de operadores booleanos, la búsqueda por campos, los operadores de proximidad, la posibilidad de mostrar índices de campos, la posibilidad de mostrar y consultar el tesauro, la búsqueda en lenguaje natural, búsquedas semánticas, etc.

2) Precisión

Mide la capacidad del SR para proporcionar tan sólo los documentos relevantes a la pregunta formulada por el usuario. Los problemas o errores pueden ser debidos tanto a la imprecisión de la consulta como a la inconsistencia en el análisis.

3) Exhaustividad

Se trata de la proporción de documentos relevantes que se suministran en respuesta a una determinada petición respecto del total de documentos que existen en la base de datos.

4) Tiempo de respuesta

Mide el lapso de tiempo transcurrido desde la formulación de la pregunta hasta la obtención de resultados. En algunos casos, no es fácil de contabilizar ya que la intensidad del tráfico en la red es muy variable.

5) Utilidad

Se puede medir objetivamente analizando la consistencia de los resultados, el grado de actualización, la presencia de duplicados, la proporción de enlaces

erróneos o inexistentes, etc. aunque es no deja de ser un indicador un tanto subjetivo (pues depende de la satisfacción del usuario respecto a los resultados).

6) Formatos de visualización

Se refiere a la posibilidad de seleccionar diferentes formatos ajustados a las necesidades de los usuarios. Incluye: formato breve para visualizar la información global, formato amplio, con resumen, para poder seleccionar los registros concretos, etc. También se refiere a prestaciones de impresión, grabación o envío por correo electrónico de los registros.

7) Usabilidad de la interfaz

El objetivo perseguido es la máxima usabilidad, es decir, una presentación clara, sencilla, intuitiva, etc. de los contenidos de la base de datos y de las prestaciones para consultarlas.

Interfaz de consulta

Los elementos que forman parte de la interfaz de interacción del usuario han sido sumarizados y descritos en el subapartado "Interfaz de consulta" dentro del apartado "Distribución de bases de datos" de este mismo módulo: diversidad del sistema de consulta o de búsqueda adecuada a usuarios expertos y principiantes: secuencial, índices, asistida, lenguaje de interrogación; navegación; selección de idioma, sistemas de ordenación de resultados, etc.

4.1.3. Gestión de la base de datos

Los criterios que se indican a continuación miden la eficacia y el grado de calidad del distribuidor de la base de datos o del departamento encargado del marketing y la promoción.

Los distribuidores tradicionales de bases de datos (como sería el caso de *Dialog*) constituyen los ejemplos de mayor calidad en este aspecto. *Google* también pone al alcance del usuario una información muy completa y detallada sobre la estructura y características de su contenido.

1) Documentación sobre la base de datos

Se trata de evaluar si existe una descripción clara y detallada de la base de datos y del sistema de consulta.

2) Atención al usuario

Se tiene en cuenta la existencia de cursos de formación dirigidos a diversos tipos de usuario, o de un servicio más o menos permanente. Esto tan sólo se encuentra en el sector de las bases de datos científico-técnicas, ya que los grandes servicios de búsqueda en Internet no pueden ni plantearse ofrecer un servicio de estas características a sus millones de usuarios.

3) Precios y sistema de facturación

Este criterio se refiere a la relación calidad-precio y los sistemas de facturación establecidos por la base de datos. En muchos casos, es difícil valorar y comparar, ya que los sistemas de cobro utilizados son complejos y no siempre tienen en cuenta los mismos parámetros (los registros visualizados, los descargados, el tiempo, etc.).

4) Sistemas de distribución

Se analiza si existe diversidad de sistemas: web y soporte óptico son los canales principales. Se trata de un criterio con poco peso, ya que el web se ha convertido en el sistema de distribución por excelencia.

4.2. ¿Cómo evaluar la base de datos? Técnicas de recogida de datos

Los indicadores antes reseñados pueden medirse mediante un sistema de evaluación o de cuantificación basado en análisis externos (funcionamiento del sistema, características del contenido de la base de datos, análisis de los registros, etc.). Ahora bien, en su gran mayoría también pueden evaluarse desde el punto de vista del usuario. Así pues, parece claro que el tiempo de respuesta es un criterio que se puede medir objetivamente contando el lapso de tiempo transcurrido desde que se pide ejecutar una petición de información hasta que aparece en pantalla una lista con los resultados. Ahora bien, se puede evaluar un aspecto que podemos denominar tiempo subjetivo, y que indica cuál es la satisfacción del usuario respecto al tiempo de respuesta del sistema.

Para conocer estos valores se necesita utilizar alguna técnica específica de recogida de datos, ya sea directa o indirecta. Las más conocidas son los cuestionarios y las entrevistas, dedicadas a conocer la satisfacción del usuario respecto del uso de la base de datos. También se puede hacer uso de la observación, normalmente grabando las acciones de los usuarios. En los últimos años está cobrando un interés notable el análisis de transacciones (o de *logs*). Los cuestionarios y entrevistas permiten conocer de forma directa lo que piensa el usuario, mientras que la observación y el análisis de transacciones permiten una aproximación indirecta, ya que tan sólo dirigen las acciones que el usuario ha llevado a cabo pero desconociendo su contexto (la pregunta que se formula) ni sus impresiones (satisfacción, utilidad, etc.).

4.2.1. Cuestionarios y entrevistas

Se trata de dos técnicas que recogen los datos directamente del usuario, y que son muy conocidas y utilizadas en amplios ámbitos de investigación. En el tipo de aplicaciones a las que hacemos referencia tienen por objetivo determinar los conocimientos, las opiniones o las actitudes de los usuarios respecto a las bases de datos que han consultado (por ejemplo, su grado de satisfacción).

La diferencia entre ambas técnicas es un tanto sutil ya que, en ambos casos, se parte de un cuestionario previo más o menos estructurado, lo que pasa es que en el cuestionario, propiamente dicho, las respuestas son escritas por el encuestado, y en la entrevista, por el encuestador, que formula las preguntas oralmente. Por otro lado, el cuestionario permite recoger datos a grupos más numerosos de personas.

La principal ventaja de estas técnicas de recogida de datos radican en que permiten un conocimiento más profundo de la opinión y grado de satisfacción del usuario que el que ofrecen métodos indirectos como el análisis de transacciones, ya que se pide directamente por sus opiniones. Por el contrario, se trata de sistemas lentos (no se pueden automatizar) y caros (tienen que pasarse personalmente) y no se pueden administrar a un conjunto muy grande de usuarios (en especial, la entrevista).

4.2.2. Observación

La observación, como técnica de recogida de datos, consiste en tomar nota o registrar el desarrollo de una actividad durante un periodo de tiempo determinado. Se trata de efectuar una vigilancia directa y un registro de las dimensiones del fenómeno que se estudia (en nuestro caso, la consulta a una base de datos o, más genéricamente, el comportamiento en el proceso de búsqueda de información). En el contexto de la recuperación de información acostumbran a utilizarse sistemas de grabación (normalmente, vídeo).

Se trata de una técnica que aporta una mayor objetividad que la entrevista o el cuestionario y que puede complementar la percepción subjetiva del usuario. Por otro lado, prácticamente no incomoda al sujeto observado y se puede aplicar en situaciones en las que los usuarios no son capaces de responder adecuadamente un cuestionario (por ejemplo, niños, personas con poca formación, etc.).

Ahora bien, es una técnica que requiere mucha paciencia, ya que la recogida de datos puede ser larga y lenta, con muchos tiempos muertos. Tiene también un cierto grado de superficialidad, porque no proporciona una visión profunda (causas, etc.) del problema a tratar. Finalmente, hay que tener en cuenta que la deontologia obliga a obtener el permiso de las personas estudiadas.

4.2.3. Análisis de transacciones

El análisis de transacciones (*transaction log analysis* o TLA) es una técnica de recogida de datos que **registra** las acciones realizadas por un usuario en un sistema de recuperación de la información. La designación en inglés incluye la palabra *log* (diario), que evoca el registro cronológico de las operaciones de proceso de datos en un sistema que se registran en un fichero.

Recuperación de la información en la web

Jansen y Pooch (2001) realizan una revisión bibliográfica sobre los diversos trabajos que, sobre estudios de recuperación de información en la web, se han publicado mostrando como la gran mayoría utilizan el análisis transaccional como base para el estudio.

El análisis de transacciones se ha utilitzado de forma extensiva desde los años ochenta para evaluar sistemas de gestión de bibliotecas (la parte pública, los OPAC) y es muy utilizada.

La estructura estándar de un **fichero de registro** acostumbra a incluir los siguientes elementos: fecha y hora; identificador de usuario; expresión de búsqueda (términos de consulta y operadores) y duración de la conexión.

Los analistas diferencian una sesión (entendida como el periodo de tiempo comprendido desde el momento en que el usuario se conecta a una base de datos hasta que la abandona) de una consulta (que es una parte de una sesión y que se refiere a la expresión de búsqueda que el usuario formula al sistema).

Expresión de búsqueda

Está formada por un término o un conjunto de términos unidos, o no, por algún operador.

El análisis de transacciones puede realizarse sobre elementos o indicadores distintos: duración (de la sesión y de la consulta), términos utilizados, operadores booleanos, campos utilizados, número de documentos recuperados, acciones realizadas (impresión, exportación), etc.

Operadores booleanos

Son operadores de búsqueda basados en la lógica como AND, OR, NOT, etc.

Se trata de la técnica más simple para recoger datos sobre la interacción entre el usuario y el SR a distancia, sin necesidad de presencia humana externa y, por tanto, sin molestar ni condicionar las acciones del usuario. Además, se puede aplicar a un gran número de usuarios, tal como se puede comprobar leyendo algunos estudios que manejan millones de interacciones.

Como en el caso de la observación, es un estudio indirecto y un tanto superficial que tan sólo permite conocer las acciones realizadas por el usuario, pero sin saber nada de sus percepciones, opiniones (qué valoración hace del sistema o de los registros obtenidos), conocimientos previos, o cuál es la necesidad de información que quiere satisfacer. Por otro lado, los ficheros que se generan son muy voluminosos y es un poco difícil trabajar con ellos.

Como valoración global podemos decir que se trata de una técnica que resulta limitada para el análisis del comportamiento de los usuarios. A pesar de ello, los datos que proporciona pueden ser muy útiles para realizar propuestas de mejora del acceso a la información en un SR.

4.3. Conclusiones

La evaluación de bases de datos es un proceso que interesa tanto a usuarios como a productores. A los usuarios les ayuda a seleccionar los contenidos más interesantes y completos, a utilizar los mejores SR o aprovechar los mejores precios. Para los productores, el interés por la evaluación y la calidad tiene un alcance mucho más profundo, ya que una preocupación constante por estas cuestiones les permitirá disponer de un producto mucho más competitivo. Si se dispone de unos criterios o indicadores se puede proceder a realizar análisis periódicos del grado de calidad de la base de datos. Estos análisis recogerán datos de carácter objetivo sobre la base de datos y, además, con las técnicas directas o indirectas que se han descrito, se interesarán por recoger la utilidad y el grado de satisfacción de los usuarios.

A partir de los resultados del análisis, pueden surgir determinadas propuestas de cambio que podrán afectar a diferentes aspectos de la gestión y el mantenimiento de la base de datos; ya sean cambios en el programa de recupera-

ción de la información, en la adecuación de los manuales y de las ayudas en línea, o en el método de trabajo establecido. Vemos, por tanto, cómo las modificaciones pueden afectar a cualquiera de los tres niveles analizados: la base de datos (contenido), el sistema de recuperación (programa) o la gestión y administración.

En general, las empresas y organismos productores de bases de datos deberían considerar los elementos discutidos aquí como parte de sus procedimientos de calidad. Si las empresas que producen las bases de datos aplican con determinada periodicidad procedimientos de control de la calidad en otros ámbitos, ¿porqué no hacerlos extensibles a las bases de datos de su departamento de documentación?

Lo anterior aún es más necesario para aquellas empresas u organismos cuya actividad depende en parte (o totalmente) de la calidad de sus bases de datos.

Resumen

A lo largo del módulo, hemos profundizado en las cuatro operaciones fundamentales que se pueden llevar a cabo en relación con las bases de datos documentales. Concretamente, nos hemos ocupado de los aspectos siguientes:

- Administración: estructura y características de los programas que permiten la creación y la explotación de bases de datos documentales.
- Distribución: interfaz de consulta de una base de datos documental e indicaciones para evaluarla y diseñarla.
- Creación: metodología para la creación de bases de datos documentales.
- Evaluación: indicadores fundamentales para la evaluación de bases de datos documentales.

Para empezar, hemos analizado las características generales de los sistemas de gestión de bases de datos (SGBD) y hemos destacado las diferencias entre los sistemas relacionales y los documentales, las cuales resumimos de la manera siguiente:

- Un *sistema relacional* (SGBDR) gestiona información muy estructurada y regular, con datos volátiles que suelen cambiar con frecuencia (cifras de ventas, direcciones postales, etc.). Tiene una estructura tabular homogénea con campos de longitud fija y limitados instrumentos de recuperación. No dispone de índices analíticos ni de controles terminológicos y se utiliza para la gestión, administración, supervisión y/o planificación en cualquier tipo de organización.
- Un *sistema documental* (SGD) gestiona información textual de tipo discursivo (artículos de revista, noticias de prensa, etc.) o descriptivo (objetos multimedia como imágenes, vídeo o sonido) con campos y registros de longitud variable. Dispone de índices analíticos, controles terminológicos y amplios instrumentos de consulta (con operadores booleanos y ayudas para una recuperación probabilista). Su finalidad es la adquisición de conocimiento y la satisfacción de necesidades de información para el estudio, la investigación, etc.

Las características del **modelo textual** en el que se basa cualquier sistema de gestión documental (SGD) son las siguientes (entre paréntesis y en cursiva se indica el software existente relacionado).

- Modelo de registro irrestricto. No tiene restricciones al tipo de registros que se pueden manejar: desde esquemas totalmente abiertos (por ejemplo,

askSam) hasta modelos perfectamente articulados en campos y tipos de datos (por ejemplo, *Inmagic* o CDS */ISIS*), pasando por tipos intermedios con flexibilidad para trabajar con campos articulados, pero sin demasiadas complicaciones (por ejemplo, *FileMaker* o *Inmagic*).

- Capacidad monobase o multibase, de manera indistinta. Se puede abrir y utilizar una sola base de datos cada vez (*askSam*) o más de una a la vez (*CDS/ISIS*, *Inmagic* o *FileMaker*).

- Índice analítico (también denominado *fichero inverso* o *fichero invertido*). Se compone de todas las palabras (términos) que representan los asuntos que aparecen en los registros o documentos de la base de datos. Para cada término del índice, hay una sola entrada (lo que permite tiempos de respuesta muy bajos) y datos de contexto como la frecuencia (número de documentos en los que aparece), la posición absoluta (número de documento - número de campo - número de palabra), posibles sinónimos, etc. La estructura del índice permite la existencia de valores repetidos, la búsqueda rápida en documentos de texto completo y tareas de control terminológico.

- Herramientas de control terminológico como diccionarios de palabras vacías (palabras sin significado que no se utilizan para indexar los documentos), sinónimos, listas de descriptores o tesauros (que permiten establecer relaciones lógicas entre los términos y los descriptores).

- Lenguaje e interfaces de consulta orientados al usuario que dispone de herramientas para la conversión de una necesidad de información en una estrategia de consulta o para el mantenimiento y la gestión de operaciones de búsqueda complejas.

A continuación, hemos clasificado los SGD en dos tipologías principales: los sistemas de indexación y los sistemas de gestión de bases de datos documentales (SGBDD). Una variante de estos últimos para usuarios personales son los sistemas de gestión bibliográfica (SGB).

1) Los **SGBD** gestionan solo referencias de documentos. Los programas más implantados en el mercado español pueden responder a necesidades de alto nivel (*Inmagic DB/Text* y *CDS/ISIS*) o de nivel medio (*FileMaker* y *Knosys*). Sus elementos estructurales sirven para las tareas siguientes:

- Definición de los registros: permite definir campos, especificar su comportamiento, definir modelos de registro mediante agrupaciones de campos y aspectos relativos al tipo de dato y al control terminológico. Todas estas especificaciones para el diseño se suelen detallar en el diccionario de datos.

- Mantenimiento: operaciones de alta, baja y modificación de registros.

- Indexación y recuperación: funciones relativas al proceso de generación de los índices, las prestaciones de recuperación, las maneras de ofrecer los resultados, etc.

- Salida e intercambio: procesos que aseguran la difusión y el intercambio de datos con el exterior por medio de la salida de los registros (exportación) y la incorporación y adaptación de ficheros externos (importación).

- Administración de la base de datos: funciones y procesos relacionados con el control y la gestión –es decir, el sistema de seguridad (creación de grupos de usuarios, adscripción de privilegios, administración de nombres de usuario y contraseñas)–, así como con la programación y las modificaciones en la interfaz.

2) Los **sistemas de gestión bibliográfica** (SGB) tienen los grupos funcionales siguientes:

- Un conjunto de medios para la entrada de referencias bibliográficas que incluye la entrada manual a través de registros predefinidos para documentos de tipo académico (artículos, libros, ponencias, etc.) y la importación automática de referencias desde bases de datos o catálogos.

- Un sistema de búsqueda que permite crear grupos selectivos de referencias: páginas de resultados con referencias bibliográficas basadas en palabras clave, nombres de autor, títulos, etc.

- Un sistema para generar bibliografías configuradas según una norma o un formato (impreso o como archivo informático: ASCII, RTF, DOC, HTML, XML, etc.).

- Un sistema para introducir citaciones y, posteriormente, generar de manera automática la bibliografía, bien formateada y ordenada al final del documento, mediante búsquedas en la base de datos hechas desde el editor de texto.

Los SGB del mercado pueden ser de dos tipos.

- *Sistemas de escritorio* que se cargan desde el ordenador del usuario. Se diferencian según el usuario final al que van destinados: *ProCite* (usuario básico), *EndNote* (usuario intermedio) y *Reference Manager* (usuario experto).

- *Sistemas en línea* que se ejecutan a través del navegador. Encontramos aplicaciones comerciales con prestaciones de muy alto nivel (*RefWorks* y *EndNote*) y aplicaciones gratuitas que se caracterizan por su facilidad de uso y por la importación de información de páginas web (*Zotero*, *Connotea*, *CiteULike*, *Mendeley*).

3) Los **sistemas de indexación** o motores de búsqueda están orientados al tratamiento del texto completo de los documentos. No necesitan definir modelos de registro y tienen capacidad de generar índices analíticos del contenido de los documentos. Hemos descrito sus módulos funcionales de la manera siguiente:

- Administración del fondo documental. La colección está formada por dos tipos de datos: los ficheros con los documentos (almacenados en unidades locales o en servidores externos remotos) y los índices (generados por el sistema) que remiten a estos documentos. Los índices son punteros que permiten acceder de manera selectiva a los documentos a partir del contenido del texto completo. Con un apuntador, el documento se puede visualizar a través de la aplicación original con la que fue creado.

- Mantenimiento. La entrada de datos se hace a partir de los ficheros informáticos que contienen la información que hay que procesar y que pueden estar en varios formatos (HTML, texto, hoja de cálculo, gráfico, etc.).

- Indexación. Se indexa el texto completo de los documentos y, si los hay, los indicadores o los metadatos para delimitar las consultas a un campo determinado del registro.

- Recuperación. Aparte de utilizar el álgebra booleana y operadores complementarios, últimamente se ha introducido la búsqueda semántica, que amplía la consulta de un término a los que están relacionados con el mismo (por derivación morfológica, equivalencia lingüística, sinonimia, etc.), y la búsqueda por patrones, que basa el análisis en la apariencia física de los términos (su código binario), lo que permite comparar y recuperar información (como texto, sonido o imagen) que comparte una serie de características estructurales comunes.

- Ponderación de resultados. Ayuda a determinar los documentos más relevantes y minimiza el elevado número de resultados no deseados que se recuperan en la búsqueda semántica y por patrones.

Los principales sistemas de indexación del mercado español son *Apache Lucene*, *askSam*, *Autonomy*, *Google Search*, *Greenstone* y *Swish-e*. Sus aplicaciones son de dos tipos: buscadores de páginas web (como *Google*, *Yahoo*, etc.) y bases de datos de texto completo con unidad temática o de publicación (por ejemplo, académicas como *Ariadne* o fondos editoriales como *Ocenet*). La mayoría de los sistemas utilizan metadatos descriptivos del contenido (autor, título, fecha, etc.), y todos usan metadatos administrativos, estructurales o de derechos de propiedad.

En el segundo apartado del módulo, hemos visto las limitaciones de los primeros sistemas utilizados para el acceso a bases de datos documentales. La

consulta local obligaba a desplazarse al centro o a la unidad en la que la aplicación estaba disponible; la publicación de bibliografías impresas suponía altos costes de impresión, distribución y actualización; y el soporte óptico requería la incorporación del programa de recuperación o del módulo de consulta en el propio soporte.

Actualmente, el sistema más utilizado para la **distribución de bases de datos documentales** es la Web. A través del navegador, el usuario puede acceder a los registros de manera actualizada y disponer de las mismas prestaciones de consulta y explotación que los sistemas de gestión documental, sin necesidad de instalar ninguna versión cliente del gestor de la base de datos.

Los elementos que intervienen en la distribución son los siguientes: navegador, servidor HTTPD, programa CGI, interfaz de consulta y base de datos. El servidor HTTPD (*Apache*, *Internet Information Server*) y el programa CGI (*Knosys Internet*, *WebPublisher* o *WWWIsis*) tienen que estar instalados en un servidor. El programa CGI sirve para consultar bases de datos documentales creadas con el SGBD correspondiente (*Knosys*, *Inmagic* o *CDS /ISIS*, respectivamente).

1) El **programa CGI** (o interfaz de pasarela) permite la comunicación entre los registros de la base de datos –que no están codificados en HTML– y el navegador, que solo puede interpretar páginas HTML. El protocolo CGI establece una manera de enviar la petición de datos desde una página web (por medio de un formulario HTML) y de procesarlos mediante un fichero ejecutable.

2) La **interfaz de consulta** es una de las partes más importantes del proceso de distribución de una base de datos, y sirve para establecer la comunicación entre el usuario y el sistema de recuperación de la información. Está formada por un conjunto de páginas HTML que deben tener una serie de elementos y funcionalidades para facilitar la recuperación de la información. Sin embargo, todos estos elementos (como por ejemplo, la elección de la forma de presentación o del sistema de ordenación de los resultados y la navegación entre registros o entre páginas de la interfaz) no tienen la misma importancia y hay que ponderarlos según dos cuestiones: la jerarquización (algunos pesan más que otros) y la universalidad (algunos no son útiles para todo tipo de usuarios). Todos ellos los hemos agrupado en tres clases de página, que constituyen el núcleo fundamental de la interfaz de consulta a bases de datos.

- Consulta: formulario para la recogida de datos del usuario (con cuadros de texto para introducir los términos de búsqueda, operadores disponibles y botones de ejecución), niveles de consulta (simple y avanzada), elección de especificaciones de visualización, historial de búsquedas, acceso multilingüe, etc.

- Lista de resultados: descripción básica de los documentos, indicación del tipo de documento, agrupamiento por categorías, ordenación por relevan-

cia u otros criterios, reformulación de la búsqueda e información sobre errores.

- Visualización del documento completo: cambio del formato de visualización y de la resolución de la página, resaltado de los términos de búsqueda.

Otros tipos de páginas que complementan las anteriores son las siguientes:

- Descripción general del contenido: ámbito geográfico, temático y lingüístico de la base de datos, datos sobre su estructura, número de registros, etc.

- Ayudas: textos sobre el funcionamiento y mensajes de ayuda y error.

- Página de identificación: conexión/desconexión por *login* y *password*.

También hemos destacado cuatro tendencias actuales de las interfaces de consulta.

- El filtrado de los resultados por características o por tipos de documento. Ejemplos de esto son *Google* y el buscador de noticias de *El País*.

- La presentación gráfica de los resultados. Es el caso de *Newsmap* o de *WebBrain* (muestran relaciones de jerarquía entre contenidos de páginas).

- La agrupación temática de los resultados según categorías generadas de manera automática (*Clusty*) o según la coocurrencia de los términos y en jerarquías subdivididas (*IBoogie*).

- Las herramientas de descubrimiento que están incorporando los nuevos OPAC de biblioteca para la consulta conjunta en diferentes colecciones, el filtrado de resultados (por temática, autor, tipo de documento, etc.) y la visualización de portadas de libros. Por ejemplo, *AquaBrowser* y *Encore Synergy*.

El tercer apartado del módulo lo hemos dedicado a la metodología para la **creación de bases de datos documentales**, y hemos indicado que la dirección del proceso tiene que ir de lo conocido (sistema de conocimiento) a lo desconocido, de los aspectos lógicos (entidades que formarán parte de la base de datos) a los aspectos físicos, y de lo general (propósito de la base de datos) a lo específico.

Este proceso de creación se tiene que ajustar al ciclo de vida siguiente:

1) Análisis de la empresa o la organización y de los objetos candidatos a ser registrados.

2) Diseño del modelo conceptual y determinación del tratamiento documental.

3) Implantación del modelo conceptual en el soporte informático.

Cada una de estas tres fases se ha dividido en subfases. El proceso es lineal, pero también interactivo, puesto que integra el procedimiento de ensayo y error para refinar el producto y obliga a repensar aspectos de las fases previas. En todo caso, es importante llegar a la fase de implantación con un modelo conceptual sólido porque, a partir de esta fase, es difícil reconsiderar el proyecto.

Cada fase tiene unos objetivos, debe producir unos resultados concretos y debe utilizar unas herramientas determinadas, tal como resumimos a continuación.

1) El objetivo de la **fase de análisis** es conocer bien la parte del mundo real, denominada *sistema objeto*, que justifica y requiere la creación de la base de datos. El sistema objeto se divide en sistema de actividades humanas (SAH) y sistema de entidades registrables (SER). El primero se refiere a la organización o sistema social (que incluye al propietario del sistema y sus usuarios) y el segundo, a los objetos (cosas, personas o conceptos) que estarán representados en la base de datos.

El resultado de esta fase es el *informe de funciones*, que consiste en la clara identificación de los aspectos siguientes: propósitos y objetivos de la empresa (SAH), funciones y beneficios que se esperan de la futura base de datos, características de las entidades registrables (SER) y referencias de sistemas similares ya en funcionamiento. La herramienta principal para hacer el informe son las entrevistas con quien hace el encargo y con los futuros usuarios, así como el análisis de la documentación sobre la empresa que aporte una comprensión global del sistema.

2) El objetivo de la **fase de diseño** es obtener un modelo conceptual de la base de datos que contenga, al menos, los siguientes elementos:

- Objetivo y propósito de la base de datos con identificación de los usuarios.
- La definición de los ámbitos temáticos de la base de datos.
- La identificación de las entidades representadas (SER).
- El diccionario de datos, que consiste en la lista detallada de los campos de la base de datos con la especificación de los siguientes parámetros: etiqueta o nombre del campo, dominio (conjunto del que puede obtener los valores el campo), tipos de datos, indexación, tratamiento documental (lenguaje libre o controlado), lengua (del documento o del centro de documentación), obligatoriedad, repetibilidad, instrucciones para la entrada de datos y otros controles de validación. El diccionario de datos ayuda al diseñador a garantizar la calidad, fiabilidad, consistencia y coherencia de la información, y tiene que incluir campos de las categorías siguientes:

  - De control (gestión interna del registro: número del registro y fecha de entrada).
  - Descriptivos (características del documento o de la entidad: autor, título, etc.).
  - Temáticos (contenido del documento o de la entidad: resumen, descriptores, etc.).
  - De derechos (restricciones o derechos que limitan la utilización del documento).
  - De ubicación (localización física del documento o puntero informático que lo abre).

- Una descripción funcional, que tiene que detallar la clase de información que se tratará y cómo se introducirá en el sistema; los procesos documentales que se llevarán a cabo; los servicios y productos que generará el sistema y las aplicaciones que podrá soportar.

- Una propuesta de tratamiento documental con orientaciones sobre el proceso de descripción y de representación del contenido semántico de las entidades.

Las herramientas principales para producir el documento del modelo conceptual son el informe de funciones, el modelo E/R (entidad-interrelación) y el diccionario de datos. Por otro lado, las normas ISBD son una herramienta inmejorable de análisis de cualquier clase de documentos para establecer los atributos de algunas entidades y la manera conveniente de representarlos.

3) La **fase de implantación** del modelo conceptual puede seguir el proceso siguiente:

- Preselección del sistema informático a partir de criterios como el grado de compatibilidad con la plataforma informática de la empresa, el grado de satisfacción de los requerimientos establecidos en el diseño conceptual, la disponibilidad de herramientas de desarrollo y de parametrización, la valoración de otros clientes y usuarios, el coste económico, la utilización de estándares, y la compatibilidad con sistemas abiertos y normas.

- Elaboración del presupuesto y el calendario de implantación. Este paso incluye la instalación del programa, una primera implantación de la base de datos y la designación de un administrador de la base de datos.

- A continuación, se prosigue de manera ordenada con las pasos siguientes: implementación de controles terminológicos, pruebas con una pequeña colección de documentos, introducción de cambios según el resultado de las pruebas; automatización de procesos repetitivos, segunda carga de datos con otra colección de documentos, prueba de usabilidad, y diseño de vistas de usuario y de carátulas de inicio.

- Definición de los grupos de usuarios y de sus privilegios de uso: administrador responsable de la base de datos (designado más arriba), documentalistas y analistas que hacen resúmenes y asignan descriptores, operadores que hacen la carga de datos y usuarios finales que utilizan la información.

- Inicio de la carga de datos y explotación del sistema estableciendo rutinas de carga con personal formado, rutinas de copias de seguridad, controles de calidad periódicos y política de mantenimiento y explotación. Esta última incluye la edición del libro de estilo con la versión definitiva del modelo conceptual, la normativa de tratamiento documental y la formación al personal y a los usuarios.

- Finalmente, si procede, acciones de promoción.

Para establecer una metodología de **evaluación y análisis de la calidad de las bases de datos documentales**, en el último apartado del módulo hemos presentado una aproximación global e integradora (es decir, aplicable tanto a catálogos de biblioteca en línea como a bases de datos especializadas o motores de búsqueda) que se compone de dos partes. La primera recopila los principales indicadores o criterios de evaluación que hay que considerar, y la segunda describe las principales técnicas de recogida de datos que se refieren a aspectos subjetivos del usuario.

1) Los **indicadores o criterios de evaluación** se pueden medir mediante un sistema de cuantificación basado en análisis externos. Los hemos agrupado en tres ámbitos: contenido de la base de datos (calidad de la información, indexación, documentos elegidos, etc.), continente (interfaz de consulta y prestaciones del programa de recuperación) y servicios adicionales al usuario (gestión y administración de la base de datos).

a) Los aspectos de calidad del *contenido de la base de datos* son los siguientes:

- Exactitud y precisión: indicadores referidos a la imprecisión en la entrada de datos (errores gramaticales y mecanográficos), la ausencia o duplicación de información en los campos y la fiabilidad de los datos (correspondencia exacta del contenido de los registros con los documentos a los que representan).

- Alcance y cobertura: grado de cobertura temática (proporción de fuentes de información de alto interés para cada área temática), grado de cobertura geográfica y lingüística, grado de inclusión de determinados tipos de documentos, estructura (número de campos definidos y recuperables), tamaño (número de registros indexados) y nivel de crecimiento (número de registros nuevos por año).

- Actualización: grado de actualización (tiempo desde que un documento está disponible hasta su inclusión en la base de datos) y periodo de actualización (periodicidad con la que se actualizan los registros).

- Consistencia de la catalogación (coherencia en el análisis formal de los documentos y la elección de puntos de acceso) y consistencia en el análisis de contenido (coherencia en la asignación de términos de indexación y de códigos de clasificación para representar una temática concreta).

b) Los indicadores relativos al *sistema de recuperación* (proceso de búsqueda y visualización de los resultados) son los siguientes:

- Prestaciones del lenguaje de interrogación: operadores booleanos y de proximidad, búsqueda por campos, búsqueda en lenguaje natural, búsqueda semántica, consulta de tesauro, etc.

- Precisión: capacidad de proporcionar solo los documentos relevantes a la pregunta formulada.

- Exhaustividad: proporción de documentos relevantes suministrados respecto de todos los de la base de datos.

- Tiempo de respuesta: tiempo transcurrido desde que se formula una petición de información hasta la obtención de resultados.

- Utilidad: consistencia de los resultados, grado de actualización, presencia de registros duplicados, proporción de enlaces erróneos, etc.

- Formatos de visualización: selección de formatos, prestaciones de impresión, grabación o envío de los registros por correo electrónico, etc.

- Usabilidad de la interfaz: sencillez y claridad de presentación de los contenidos y prestaciones para hacer consultas.

c) Los indicadores referidos a la *gestión de la base de datos* miden la eficacia y el grado de calidad y la promoción de la distribución. Incluyen todo lo que se refiere a la documentación sobre la base de datos y al sistema de consulta, las facilidades otorgadas a los usuarios (cursos de formación, servicios de atención, etc.), los precios y los sistemas de facturación.

2) Las **técnicas de recogida de datos** sobre el usuario, que permiten conocer de manera directa o indirecta aspectos subjetivos como la utilidad de la base de datos y la satisfacción del usuario o seguir las acciones que este lleva a cabo, y que son las siguientes:

- Los *cuestionarios* y las *entrevistas* permiten recoger los datos directamente del usuario y conocer su satisfacción respecto del uso de la base de datos. A pesar de que las dos técnicas parten de un test, en el cuestionario las respuestas son escritas por la persona encuestada y, en la entrevista, por la persona que hace la encuesta. Estas técnicas, muy conocidas y utiliza-

das, permiten un conocimiento profundo del grado de satisfacción pero son lentas (no se pueden automatizar), caras (se tienen que pasar personalmente) y no se pueden aplicar a un conjunto elevado de usuarios (en especial, la entrevista).

- La *observación* efectúa una vigilancia del comportamiento del usuario y registra (normalmente en vídeo) sus acciones en el proceso de búsqueda de información. Se trata de una técnica más objetiva que las anteriores, que prácticamente no incomoda al usuario, complementa su percepción subjetiva y es aplicable a niños y personas con poca formación. Ahora bien, requiere mucha paciencia, presenta cierto grado de superficialidad y obliga a obtener el permiso del usuario.

- El *análisis de transacciones* hace un registro cronológico de las operaciones de procesamiento de datos llevadas a cabo por el usuario, en un fichero de registro que puede incluir lo siguiente: fecha y hora de inicio, identificador de usuario, expresión de búsqueda y duración de la consulta, número de documentos recuperados y acciones de impresión y exportación. Se trata de la técnica más simple para recoger datos a distancia, no requiere presencia humana externa, se puede aplicar a un número muy elevado de usuarios y no condiciona sus acciones. Por el contrario, resulta un poco superficial (está limitada al análisis de las acciones y el comportamiento del usuario) y genera ficheros muy voluminosos.

Bibliografía

Abad, F. (1997). *Investigación evaluativa en Documentación: aplicación a la documentación médica*. Valencia: Universidad de Valencia.

Abadal, E. (2002, septiembre-octubre). "Elementos para la evaluación de interfaces de consulta de bases de datos". *El profesional de la información* (vol. 11, núm. 5, pág. 349-360). Swets & Zeitlinger Publishers. [Fecha de consulta: 25 de mayo del 2010.] (http://www.elprofesionaldelainformacion.com/contenidos/2002/septiembre/3.pdf)

Abadal, E.; Codina, L. (2005). *Bases de datos documentales: características, funciones y método*. Madrid: Síntesis.

Blair, D.C. (1990). *Language and representation in information retrieval*. Amsterdam (etc.): Elsevier.

Checkland, P. B. (1981). *Systems thinking, systems practice*. Chichester: Wiley.

Codina, L. (1993). "Metodología de análisis de sistemas de información y diseño de bases de datos documentales: aspectos lógicos y funcionales". *Anuari SOCADI de documentació i informació* (p. 195-209). Barcelona: SOCADI.

Codina, L. (1993). *Sistemes d'informació documental: concepció, anàlisi i disseny de sistemes de gestió documental amb microordinadors*. Barcelona: Pòrtic.

Connolly, T.M.; Begg, C. E. (2005). *Sistemas de bases de datos: un enfoque práctico para diseño, implementación y gestión* (4.ª ed.). Madrid: Pearson.

Hearst, M. A. (1999)."User interfaces and visualization". En: Baeza-Yates, Ricardo; Ribeiro-Neto, B. *Modern information retrieval* (pág. 257-323). Nueva York: ACM; Harlow: Addison-Wesley.

Jacsó, P. (1997). "Content evaluation of databases". *ARIST* (vol. 32, pág. 231-267).

Jansen, B. J.; Pooch, U. (2001). "A review of web searching studies and a framework for future research". En: *JASIS* (vol. 52, núm. 3, pág. 235-246). Nueva York: John Wiley.

Marchionini, G. (1995). *Information seeking in electronic environments*. Cambridge: Cambridge University.

Miguel, A. de; Piattini, M. G. (1999). *Fundamentos y modelos de bases de datos* (2.ª ed.). Madrid: Ra-Ma.

Morville, P.; Rosenfeld, L. (2006). *Information architecture for the world wide web*. (3.ª ed.). Sebastopol (California) (etc.): O'Reilly.

Nielsen, J. (2000). *Usabilidad: diseño de sitios web*. Madrid (etc.): Prentice-Hall.

Nielsen, J.; Loranger, H. (2006). *Usabilidad: prioridad en el diseño web*. Madrid: Anaya Multimedia.

Peña, R. (2002). *Gestión digital de la información: de bits a bibliotecas digitales y la web*. Madrid: Ra-Ma.
Information Market Observatory (IMO) (1995). "The quality of electronic information products and services". *IMO Working Paper* (vol. 4, núm. 95).

Raya, F. (1987). *Database design for information retrieval: a conceptual approach*. Nueva York (etc.): John Wiley & Sons.

Rodríguez Yunta, L. (1998). "Evaluación e indicadores de calidad en bases de datos". *Revista española de documentación científica* (vol. 21, núm. 1, pág. 9-23). Madrid : Consejo Superior de Investigaciones Científicas: Instituto de Información y Documentación en Ciencia y Tecnología.

Rodríguez Yunta, L.; Tejada, C. (coord.) (2003). *Directorio español de software para la gestión bibliotecaria, documental y de contenidos*. Madrid: Consejo Superior de Investigaciones Científicas.

Shneiderman, B.; Byrd, D.; Croft, W. B. (1997, enero). "Clarifying search: a user-interface framework for text searches" [article en línea], *D-Lib Magazine*. (vol. 3, núm. 1) [Consulta: 25/05/2010.]
<www.dlib.org/dlib/january97/retrieval/01shneiderman.html>

Soergel, D. (1985). *Organising information principles of data base and retrieval systems*. San Diego [etc.]: Academic Press.

Tramullas, J.; Olvera, M. D. (2001). *Recuperación de la información en internet*. Madrid: RaMa.

Van Rijsbergen, C. J. (1975). *Information retrieval* [en línea]. London: Butterworths. [Consulta: 1/6/2010].
<http://www.dcs.gla.ac.uk/Keith/Preface.html>.

Villanueva, E. (1996, enero-junio). "Bases de datos y bibliotecología: como deshacer la innecesaria incomunicación". *Investigación bibliotecológica* (vol. 10, núm. 20, pág. 27-32). Mèxic, DF: Centro Universitario de Investigaciones Bibliotecológicas, UNAM.

Willitts, J. (1992). *Database design and construction: an open learning course for students and information managers*. Londres: Library Association.

Wilson, T. D. (1998). "EQUIP: a european survey of quality criteria for the evaluation of databases". *Journal of information science* (vol. 24, núm. 5, p. 345-357). Amsterdam: Elsevier.

Xie, M.; W., H.; Goh, T. N. (1998). "Quality dimensions of Internet search engines". *Journal of information science* (vol. 24, núm. 5, pág. 365-372). Amsterdam: Elsevier.