# Bases de datos documentales
## Características, funciones y métodos

Ernest Abadal, Lluís Codina

Forma recomendada de citación

Abadal, Ernest; Codina, Lluís (2005). *Bases de datos documentales: características, funciones y métodos*. Madrid: Síntesis.

[versión de los autores]

**Sumario**

## Presentación

Los autores empezamos nuestra aventura docente e investigadora en la Universidad Autónoma de Barcelona en 1987, ayudando a los alumnos de la facultad de Ciencias de la Comunicación a conocer las técnicas y procedimientos de la Documentación Periodística; ayudándoles a saber organizar la información de actualidad, publicada en revistas o prensa diaria, informaciones que eran eminentemente textuales pero con notables componentes gráficos, y que podían ser susceptibles de futura utilización para generar nuevas informaciones.

En aquellos momentos nuestro programa docente e investigador tenía dos ejes fundamentales. En primer lugar, el desarrollo y la aplicación de una metodología pensada especialmente para el diseño y creación de bases de datos documentales en un contexto en el cual tan sólo se aplicaba el modelo relacional, muy contrastado y fundamentado, aunque pensado para solventar otro tipo de situaciones. En segundo lugar, dedicamos muchos esfuerzos a caracterizar, buscar y seleccionar programas informáticos que se ajustaran a las necesidades antes descritas, es decir, que facilitaran la representación y recuperación de la clase de documentos complejos tan característica de la información de actualidad. Así fue como nos las vimos con *Archivist* (y posteriormente, con *Idealist*, su evolución), *CDS/ISIS*, *Knosys* e *Inmagic*, entre muchos otros. Los pusimos a prueba, nos entrevistamos con sus creadores o sus distribuidores, los analizamos y evaluamos, y elaboramos incluso tutoriales para facilitar su uso por parte de nuestro alumnado.

Con el paso de los años aparecieron otros ejes de interés complementarios a los antes comentados. La irrupción del Web, sin duda, fue uno de los más impactantes. El uso masivo de este canal de distribución de información atrajo el interés de los productores de bases de datos y generó la necesidad de ver "publicadas" o accesibles las bases de datos también en el Web a fin de poder llegar con facilidad a un número de usuarios inimaginables en la consulta local.

Finalmente, el último centro de interés del cual también nos hemos hecho eco es la preocupación por la calidad y la evaluación –ampliamente extendida en la mayoría de los productos y servicios documentales- y que, obviamente, que también afectó de lleno a las bases de datos.

Los conocimientos que fuimos adquiriendo como consecuencia de nuestra actividad académica pronto pudimos ponerlos en práctica y contrastarlos con la realidad ya que empezamos una línea de colaboración con distintas organizaciones públicas y privadas (medios de comunicación, etc.) que necesitaban diversos grados y diversos tipos de asesoramiento para organizar sus documentos y facilitar su recuperación. Este banco de pruebas resultó ser un complemento fundamental para nuestras tareas académicas. Por razones obvias, el contraste con la realidad nos ayudó a refinar, complementar o modificar nuestras metodologías y modelos conceptuales, buena parte de los cuales presentamos aquí.

Así pues, confiamos que esta breve presentación de nuestra trayectoria ayude a dejar bien claro que la recuperación de la información aplicada a entornos documentales ha

sido y sigue siendo nuestra gran preocupación científica y profesional. El libro que tiene el lector en sus manos es, por tanto, el fruto, casi la destilación, de 16 años de experiencia y estudio académico y profesional.

De hecho, las cuatro grandes líneas temáticas que nos han acompañado en nuestra trayectoria vital y sobre las cuales hemos podido reflexionar en el aula y en la empresa –diseño y creación de bases de datos, utilización de programas documentales, distribución a través del web, y evaluación de bases de datos– constituyen los ejes vertebradores del libro que presentamos. Así pues, en sus páginas se encuentran los elementos metodológicos necesarios para diseñar y crear una base de datos documental, se describen y analizan los programas informáticos que facilitan la producción de esta base de datos y también su distribución en el Web y, finalmente, se presentan los principales indicadores que han de servir para evaluar la calidad de la base de datos.

Se trata de un texto de carácter a la vez teórico y práctico. Es de carácter teórico porque presenta y discute modelos y conceptos, pero también es práctico porque se preocupa de los métodos y procedimientos. Por otro lado, no hemos olvidado el necesario componente aplicado que esta temática requiere y que hemos intentado transmitir siempre en nuestra actividad docente.

Para finalizar, una breve referencia a la audiencia potencial. El asunto central de este libro es materia de estudio en diversas titulaciones: Biblioteconomía y Documentación, Periodismo, Comunicación Audiovisual, Traducción e Interpretación, etc. así como en cursos de postgrado y extensión universitaria. Por otro lado, también es objeto de interés por parte de profesionales de especialidades diversas, ya sean periodistas que necesitan organizar los artículos de prensa de su especialidad, gestores de la cultura, maestros, técnicos de la administración, filólogos preocupados por la recopilación de críticas literarias, historiadores, arquitectos, etc. y, por supuesto, de manera muy especial, archiveros, bibliotecarios y documentalistas. Todos ellos son colectivos profesionales que tienen la necesidad estratégica de organizar documentos e informaciones científicas o culturales y de concebir o diseñar sistemas de información apoyados en bases de datos para facilitar la explotación de su contenido.

Dicho esto, tan sólo nos queda desear que la lectura de esta obra os sea útil y provechosa –aunque sea parcialmente– y se adecue al máximo a vuestros intereses.

Los autores: Ernest Abadal y Lluís Codina
Barcelona, diciembre de 2003

# 1 Introducción

## 1.1 La información en la sociedad post-industrial

Podríamos iniciar este capítulo recurriendo a una obviedad: la información y el conocimiento constituyen uno de los activos más importantes de la sociedad contemporánea. El motivo está claro. La información se encuentra en la base de los procesos de toma de decisiones. Toda organización necesita disponer de recursos informativos de la máxima calidad y fiabilidad antes de abordar la resolución de sus problemas.

Estos recursos informativos, no obstante, son de poca utilidad si no se encuentran organizados y formando parte de lo que se denominan *sistema de información*. Las bases de datos, por su parte, son uno de los componentes esenciales (tal vez **el** componente esencial) de los sistemas de información actuales. Presentar y discutir la naturaleza de las bases de datos, la metodología necesaria para diseñar sistemas de información apoyados en bases de datos, así como dar a conocer los instrumentos que se requieren para su producción y distribución son los ejes fundamentales sobre los que gravita este libro.

Pero, ¿por qué las bases de datos precisamente? A medida que la humanidad ha ido generando progresos en los diversos campos de la ciencia, ha ido necesitando cada vez mejores herramientas para gestionar este conocimiento. Este doble proceso: *generación* / *gestión* del conocimiento es inseparable, al menos desde los años 40 del siglo anterior, y sin duda continuará siéndolo en el futuro. Por tanto, podemos predecir que cada vez serán necesarios sistemas más eficientes de información, puesto que debemos confiar en que no se detendrán en el futuro los avances en el conocimiento...

En este sentido, las bases de datos son la mejor tecnología de que disponemos en la actualidad para gestionar información, ya que es el único sistema que permiten procesar la información de una forma que es, a la vez, segura, rápida y eficaz. De hecho, existen otras tecnologías basadas en ordenadores para gestionar información. Para mencionar algunas de ellas: 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 de manera selectiva, mostrarla de forma diferente a diferentes grupos de usuarios, explotarla de forma diferente si cambian los objetivos, etc., y, todo ello, en el marco de una relativa seguridad y confidencialidad tanto frente a accesos maliciosos como frente a errores involuntarios.

Se ha dicho que el destino de una tecnología que triunfa es volverse transparente. En realidad, todos estamos familiarizados con el uso de bases de datos, aunque no siempre somos conscientes de ello. En Internet, por ejemplo, muchos profesionales y ciudadanos consultan con frecuencia bases de datos de cine, de música o de libros como parte habitual de sus actividades profesionales o de ocio. De hecho, la mayor parte de las actividades que tienen alguna cosa que ver con la cultura, la investigación, la ciencia, la comunicación, la enseñanza o la I+D están vinculadas con la creación o con el uso de bases de datos de algún tipo y, sobre todo, con la creación y con el uso de bases de datos documentales.

Si echamos la vista atrás, podemos ver que, históricamente, las primeras bases de datos documentales aparecen a finales de los años 60 en los EUA, y aparecen vinculadas al mundo de la información periodística y de la información científico-técnica. Desde entonces no han dejado de extenderse a todos los terrenos y actividades sociales. Es por ello que en la actualidad se dispone de una abundante bibliografía científica en la que se hace referencia a las bases de datos desde perspectivas y contextos muy distintos.[1]

## 1.2 Una distinción básica

Después de resaltar la importancia de las bases de datos en la sociedad de la información, y con el objetivo de contribuir a fijar los conceptos básicos que vamos a manejar en este libro, introduciremos la dicotomía *base de datos* versus *Sistema de Gestión de Bases de Datos*, que no siempre se distingue con claridad, aunque tal distinción es de gran importancia.

En primer lugar, corresponde recordar que una *base de datos* es el conjunto o colección de datos, mientras que un *sistema de gestión de bases de datos* es el programa que permite la creación, el mantenimiento y la explotación de la base de datos. Aunque (a todos los efectos prácticos) no es cierto que existan bases de datos sin un sistema de gestión de bases de datos, lo contrario si es cierto, o sea una empresa puede poseer programa de gestión de bases de datos y no tener ninguna base de datos, por la misma razón que una cosa es adquirir un programa de edición de textos, como *Word*, y otra crear documentos con el mismo. Por eso, y porque son realidades conceptualmente distintas, conviene tener presente la diferencia entre ambos conceptos, que desarrollaremos en los siguientes apartados.

## 1.3 Base de datos

En lo que sigue, intentaremos realizar una aproximación al concepto de *base de datos* desde dos niveles diferenciados: en primer lugar, situándonos en un contexto conceptual, indicando qué representan las bases de datos en el mundo de los sistemas de información; en segundo lugar, describiendo directamente cuáles son sus características técnicas básicas.

En los textos académicos se pueden encontrar definiciones muy teóricas de las bases de datos, que ponen el énfasis no tanto en lo *que son* sino en lo *que representan*. Siguiendo esta línea, como primera aproximación partiremos de la idea que una base de datos es un intento de representación de una parte del mundo real:

> "Una base de datos es un almacén de datos de una parte seleccionada del mundo real para ser utilizado con propósitos particulares." (Fidel, 1987: 5)

[1] Hay que señalar, sin embargo, que tan sólo una pequeña proporción se dedica específicamente a las bases de datos documentales.

Podemos decir, por tanto, que una base de datos es una representación de alguna parte de la realidad; que esta representación ha sido realizada por una persona, empresa u organización con algún propósito determinado, en general, para dar servicio a un grupo de usuarios o para dar soporte a determinados procesos. Por ejemplo, la base de datos del catálogo de una biblioteca es una representación del fondo documental de la biblioteca. Esta base de datos estará producida por la propia biblioteca y su propósito será facilitar la consulta del fondo por parte de los usuarios, pero también realizar procesos de mantenimiento del mismo (préstamos, adquisiciones, etc.).

Si analizamos el concepto anterior de base de datos observamos que alude al menos a dos elementos básicos:

- *El objeto de la base de datos* (*empresa*, en la terminología de Fidel)
Como hemos señalado, la base de datos es una representación de una parte del mundo real. Esta parte del mundo real puede estar formada por un solo tipo de entidad o por diversos tipos de entidades. En este sentido, cuando se crea una base de datos se ha de definir con precisión cuál es la parte del mundo real a la que se va a hacer referencia.

- *La finalidad o el sujeto de la base de datos* (*entorno*, en la terminología de Fidel)
El entorno es el contexto en el cual se crea la base de datos, es decir, se trata de una determinada organización, con unos usuarios concretos y unas necesidades de información más o menos precisas. La finalidad de la base de datos no es otra que solucionar necesidades de información. Ahora bien, las características y necesidades propias del entorno condicionarán en cada caso de forma directa la estructura y organización de las bases de datos.

Como se verá más adelante (v. 5), cuando se aborde la metodología de diseño y creación de bases de datos, las características del objeto o parte de la realidad que se representa (la *empresa*) y de la finalidad para la que se utilizará (el *entorno*, el contexto) constituyen los dos condicionantes básicos que hay que tener presentes cuando se va a diseñar una base de datos.

Ahora bien, también son abundantes las definiciones descriptivas de las bases de datos, es decir, aquellas que se limitan a presentar las características diferenciales, más o menos técnicas, que deben estar presentes en este concepto. A continuación, indicamos una de estas definiciones que puede servirnos de modelo:

> [Una base de datos es] “una colección de datos almacenada en archivos de ordenador que es accesible a diversos usuarios y diversos programas.” (Willits, 1992: 9)

Vemos ahora que, desde este punto de vista, se sitúa el énfasis en el hecho de que una base de datos es un conjunto de datos estructurados de forma sistemática. Si realizamos un análisis de esta definición y de otras similares, especialmente presentes en manuales de informática, comprobaremos como, en la mayoría de ellas, aparecen las siguientes características propias de las bases de datos:

- *Los datos están interrelacionados y estructurados siguiendo un modelo*
Los datos han de poseer alguna estructuración interna, es a decir, no se puede tratar de un mero depósito o almacén de información. Para ello hay que recurrir a diversos modelos –que se tratarán más adelante (v. 5)- que ayudan a estructurar e interrelacionar

los datos para facilitar la recuperación de la información que contienen con la máxima eficacia.

- *Los datos están almacenados en un soporte informático*
Este es otro aspecto fundamental: el contenido de una base de datos ha de estar grabado en un soporte digital. De otra forma nos encontramos, por ejemplo, frente a un listado impreso.

- *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 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. El SGBD (programa informático) es diferente a la base de datos (el conjunto de datos e información) y es (relativamente) independiente de los datos. Esto quiere decir que una misma base de datos podría ser gestionada por programas diferentes y que un mismo programa podría gestionar bases de datos distintas.

- *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 datos de una base de datos de recursos humanos servirá, 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.

### 1.3.1 Estructura

Una *base de datos* es un conjunto de informaciones sobre algún ámbito o dominio del conocimiento. A diferencia de otras estructuras de información, en una base de datos estas informaciones están tratadas de manera uniforme y sistemática, de modo que su explotación puede realizarse de forma óptima. Por ejemplo, en el momento de la búsqueda, puede encontrarse información de manera rápida y selectiva, aunque el universo de búsqueda esté compuesto por centenares de miles o por millones de documentos, ya que antes esos documentos han sido tratados de manera uniforme y sistemática, en el momento de la entrada de la información.

Es este tratamiento uniforme y sistemático el que proporciona un gran valor de explotación a esta clase de sistemas de información. Por un lado, la uniformidad es una garantía de calidad: si un gran conjunto de informaciones está tratada de manera homogénea entonces su tratamiento será mucho más cómodo para el usuario final. Compare el lector la facilidad de uso de un catálogo típico de biblioteca respecto del explorador del sistema operativo que permite ver los documentos en un típico disco duro: en un catálogo de biblioteca (y en cualquier base de datos en general) los datos están representados siempre igual y son explícitos: el título del documento primero, por ejemplo, el autor después, etc. De este modo es fácil comparar y buscar documentos e informaciones.

Por otro lado, la *sistematicidad* propia de las bases de datos asegura que cada documento de su fondo tenga un mismo conjunto de datos (*metadatos*, en este caso), por ejemplo, la fecha de creación, el nombre de autor, los descriptores sobre su contenido, etc., lo que no sucede con otras clases de fondos o de colecciones de documentos. Si un atributo de un documento o de una información es clave en una organización, por ejemplo, el nombre del autor, la fecha de creación o el tema, solamente mediante el uso de una base de datos podremos estar seguros que todos los documentos contendrán esos atributos bien representados.

Como resultado final, solamente cuando los atributos de una información compleja, como es el caso de algunos documentos científicos o culturales, están sistematizados se pueden realizar procesos de explotación que sean, a la vez, inteligentes y seguros: por ejemplo, podremos saber qué documentos sobre el tema *X* desde el punto de vista *Z*, se han publicado en un rango de fechas determinado y, a la vez, podremos estar seguros que la respuesta será (relativamente) fiable.

Ahora bien, los dos componentes de una base de datos responsables de la uniformidad, sistematicidad y seguridad que hemos señalado son los siguientes: registros, y campos. Es mediante las estructuras formadas por *registros* y *campos* (conceptos que se tratan a continuación) que se pueden sistematizar y tratar de manera uniforme las informaciones de una base de datos. De aquí la importancia que consideremos estas dos estructuras con cierto detalle.

Los registros son, a la vez, la unidad mayor y la unidad principal de trabajo de una base de datos. Pero, ¿qué es un registro? Un registro es una representación de una entidad. Aclaremos este extremo. Hemos dicho que una base de datos es una representación de un aspecto de la realidad. Las cosas que representamos en una base de datos se denominan *entidades* y sus representaciones se denominan *registros*.

Una entidad, a su vez, es cualquier objeto, físico o conceptual, real o imaginario, que está descrito o representado en la base de datos. Por ejemplo, en una base de datos de films, la entidad son films. En una base de datos de fotografías, la entidad son fotografías, etc.

En una misma base de datos pueden estar representados uno o más tipos de entidades. En una base de datos de cinematografía, por ejemplo, en *AllMovie Guide* (allmovie.com), las entidades representadas en la base de datos son films y también cineastas. Esto significa que en *AllMovie Guide* encontramos registros que describen a cineastas, así como encontramos registros que describen películas de cine. La relación, en todo caso, siempre es la misma: una entidad, un registro. Si *AllMovie Guide* contiene datos sobre más de 250.000 films, esto significa que contiene más de 250.000 registros (de films) y si contiene datos sobre, por ejemplo, 900.000 cineastas, esto significa que contiene 900.000 registros de este otro tipo (de cineastas).

Para consultar bases de datos basta conocer la relación entre registros e entidades. Ahora bien, para *diseñar* bases de datos es importante distinguir entre *tipos* de registro y *ocurrencias* de registro. Sigamos con el ejemplo de *AllMovie Guide:* podemos encontrar, por un lado, información sobre el film *2001* y, por otro, información sobre su realizador, Stanley Kubrick. Cada uno de estos registros es una *ocurrencia*. Y cada uno de estos registros forma parte de un *tipo de registro.* Así pues, en *AllMovie Guide* hay al

menos dos tipos de registros: el de *films* y el de *cineastas*. Como se puede ver, cada *tipo de registro* sirve para representar a un *tipo de entidad.* El tipo de registro "Films" sirve para representar las diversas ocurrencias del tipo de entidad films, por ejemplo, la conocida obra de Kubrick *2001: una odisea del espacio*. En síntesis: el modelo de registro que sirve para representar films, cualquier film en general, es un *tipo de registro*. El registro concreto que representa a un film concreto, por ejemplo, a *2001,* es una *ocurrencia de registro.*

Cuando diseñamos bases de datos, necesitamos acostumbrarnos a pensar en tipos de entidades, en primer lugar, para pasar después a pensar en cómo serán los tipos de registros que habrán de servir para representar a los tipos de entidades que hemos determinado.

Como hemos dicho, el registro es la unidad principal de trabajo, tanto para el usuario de la base de datos como para el profesional de la información o el documentalista que la diseña. Ahora bien, para este último existe una estructura adicional de enorme importancia: los *campos.* En concreto, para el profesional, muchas veces la unidad de trabajo más significativa es el *campo*, aún más si cabe que el propio registro. El motivo es que muchas de las decisiones que harán que una base de datos cumpla su función como sistema de información dependerán de las opciones que tome el diseñador de la base de datos a propósito de los campos.

Pero, ¿qué es un campo? Un campo es una zona de un registro. En términos lógicos, un campo representa un atributo de una entidad. Si los films tienen un título, una fecha de distribución, un realizador, una sinopsis, unos actores, un director de fotografía, etc., entonces decimos que cada una de estos elementos mencionados (título, fecha, realizador...) son atributos del film. Por tanto, en el tipo de entidad que sirve para representar films todos y cada uno de esos atributos deberán estar representados, y si respecto a las entidades hablamos de atributos, en los registros hablamos de campos.

Ya es hora de ilustrar estos conceptos. La figura siguiente representa a un registro que, a su vez, representa a un film (seguimos el modelo de *AllMovie Guide*, aunque obviamos algunos campos).

**Tabla 1.1: La representación de un film en una típica base de datos documental (datos de AllMovie)**

| | |
|---|---|
| **Movie** | **2001: A Space Odyssey** |
| **Year** | 1968 |
| **Nationality** | UK/USA |
| **Time** | 139 min. |
| **Type** | Feature |
| **Color** | Color |
| **Rating** | ***** |
| **Director** | Stanley Kubrick |
| **Genre** | Space Adventure, Science Fiction |
| **Keywords** | Computer, Exploration, Kill, Space, Technology |
| **Plot Lines** | Computer on rampage, Exploration of uncharted territory, Exploration of Space, Technology on rampage, Machines on |

| | |
|---|---|
| | rampage |
| **...** | |

La tabla anterior es un registro, y cada una de las filas que van de izquierda a derecha de la tabla y que tienen una etiqueta o rótulo del estilo *Movie, Year*, *Nationality*, etc., son cada uno los campos o zonas de información que componen el modelo de registro. En el caso del modelo de *AllMovie*, el registro completo tiene unos 26 campos en total.

### 1.3. Sistemas de Gestión de Bases de Datos (SGBD)

En el anterior subapartado hemos aludido a los sistemas de gestión de bases de datos (SGBD), un término directamente relacionado con la base de datos y sobre el que hemos puesto especial interés en diferenciar. Para profundizar en este concepto, y yendo un poco más allá de su consideración como el instrumento que permite la creación y explotación de bases de datos, podemos partir del análisis de 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: 38)

En este caso, el eje de la caracterización se sitúa en considerar al SGBD como el instrumento que facilita el acceso a los datos por parte de diferentes tipos de usuarios, y esto es posible porque el SGBD permite representar y almacenar los datos de una forma al mismo tiempo estructurada y desagregada, de modo que el mismo conjunto de datos puede dar servicio a distintos tipos de usuarios, cada uno de los cuales utilizará las vistas (grupos de características de los datos) más convenientes para sus tareas. Por ejemplo, los administradores de la base de datos del catálogo de la biblioteca verán características de los documentos (p.e., el precio de adquisición y otros datos técnicos de procesamiento documental) distintas de los usuarios de la biblioteca (p.e., título, autor, tema, etc.)

Ahora bien, no todos los SGBD son iguales en funciones y objetivos, por lo cual, de hecho, existen al menos dos grandes tipos de SGBD: aquellos que podemos denominar de tipo *administrativo* (o relacional) y aquellos que podemos denominar de tipo *documental* (o textual). Veamos sus diferencias:

- *Sistemas de gestión de bases de datos administrativos.*
Suelen utilizar un modelo lógico de datos denominado *relacional.* Por este motivo, acostumbran a denominarse también *Sistemas de Gestión de Bases de Datos Relacionales* (SGBDR). Son programas especialmente adecuados para la gestión de información muy estructurada (datos propiamente dichos, por ejemplo, volumen de ventas, sueldos o existencias de almacén). En la concepción informática clásica, de hecho, es el único tipo de sistema de gestión de base 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).

- *Sistemas de gestión de bases de datos documentales*
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: artículos de revistas, páginas web o reportajes fotográficos, para mencionar tres ejemplos muy dispares. Por ese motivo, se suelen denominar *Sistemas de Gestión Documental* (SGD). Mientras que 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 y los índices analíticos, tal y como se verá más adelante (v. 3).

La siguiente tabla ofrece un resumen de los rasgos diferenciales fundamentales de los dos grandes tipos de sistemas de gestión de bases de datos considerados. En los capítulos siguientes nos centraremos en los sistemas de gestión documental, que son los utilizados para la creación y explotación de las bases de datos documentales.

**Tabla 1.2: SGBDR versus SGD**

| ***Tipo de sistema*** | ***Contexto*** | ***Tipo de datos*** | ***Finalidad*** |
|---|---|---|---|
| **SGBDR** | 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. |
| **SGD** | 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. |

Después de esta breve caracterización y diferenciación de los conceptos básicos que se van a utilizar en el libro, los capítulos siguientes se dedican a tratar de forma detallada las metodologías para el diseño y creación de bases de datos documentales, los programas informáticos para su producción y distribución y, finalmente, los elementos fundamentales para su evaluación como producto documental. Antes de ello, no obstante, se presenta un capítulo dedicado a la Recuperación de información, que pretende dibujar el marco teórico general en el que se inscriben las bases de datos documentales.

**Bibliografia**

Blair, D.C. *Language and representation in information retrieval*. Amsterdam [etc] : Elsevier, 1990. 335 p.

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

Fidel, Raya. *Database design for information retrieval: a conceptual approach*. New York [etc.]: John Wiley & Sons, 1987. 232 p.

Miguel, Adoración de; Piattini, Mario. *Fundamentos y modelos de bases de datos*. Madrid: Ra-Ma, 1997.

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

Willitts, John. *Database design and construction: an open learning course for students and information managers*. London: Library Association, 1992. XXII, 425 p.

# 2 Recuperación de Información

## 2.1 Definición y contexto

Recuperar significa volver a tener. Recuperar información significa volver a tener una información que alguna vez, hace unos minutos o hace unos años, había sido producida por alguien, bien por nosotros mismos o bien por terceras personas.

Ahora bien, para volver a tener algo, hace falta un mínimo de organización. Al menos hace falta un sitio donde guardar ese algo y si el sitio es grande y tenemos muchas cosas juntas, necesitaremos algún procedimiento para saber dónde está exactamente lo que necesitamos en cada caso.

La *Recuperación de información* (RI, a partir de ahora) es la disciplina que estudia la representación, la organización y el acceso a la información. Esta disciplina considera la información como un recurso y su objetivo es explotar ese recurso de la forma más eficiente para tomar mejores decisiones, para refinar el conocimiento existente o para crear conocimiento nuevo.

Aunque es frecuente presentar a la información como si fuera algo inmaterial o intangible, lo cierto es que la información siempre requiere de un soporte material para ser explotada, es decir, para poder rendir su máxima utilidad. Al igual que algunas formas de energía son más útiles que otras según la forma en la que se presenten, las informaciones más útiles desde el punto de vista de su explotación económica y cultural son aquellas que están registradas en documentos. La razón es simple: por un lado, sin documentos todo el conocimiento de la humanidad se limitaría al que pudiera almacenarse en el cerebro humano. Por otro lado, todo el saber que podría adquirir una persona se limitaría al que pudieran transmitirle oralmente sus vecinos. En cambio, los documentos permiten que la información se convierta en un recurso social de enorme potencia, al mismo tiempo que permiten que supere las barreras del tiempo y del espacio. Una información que no queda registrada en alguna clase de soporte material, por muy valiosa que sea, se pierde de manera tan definitiva como la energía que se pierde por rozamiento en una máquina. Ni la una (la información) ni la otra (la energía) podrán volver a ser utilizadas. Hasta que Darwin no publicó sus conocimientos de nada le sirvieron a la sociedad (ni a él mismo, en realidad). Sin ser dados a conocer en forma de su famosa serie de artículos, de nada hubieran servido las ideas de Einstein, etc.

Por tanto, de las operaciones propias de la RI, sin duda la más característica consiste en la *selección* de documentos, bien a partir de las características de su *contenido* (los temas tratados en los documentos) bien a partir de características de su *contexto* (p.e. la fecha de publicación) bien a partir de alguna combinación de ambas cosas (p.e: "documentos sobre desarrollo humano publicados por la UNESCO entre 2003 y 2005"). El objetivo siempre es obtener información, pero como ya hemos argumentado que la mayor parte de la información útil reside en documentos, el paso intermedio siempre implica seleccionar documentos. Por tanto, la RI puede verse como un sistema de comunicación *asíncrono* que pone en relación a los *productores* de la información (autores de los documentos) con los *consumidores* de la información (usuarios de sistemas de información)

Ahora bien, para que la RI tenga sentido se presupone un entorno en el cual no es trivial, precisamente, el hecho de acceder a los documentos por su contenido. Este contexto lo genera, típicamente, cualquier fondo documental a partir del momento que contenga unos centenares o unos miles de documentos. Ejecutivos, abogados, químicos o ingenieros que necesitan encontrar una información en fondos internos o externos es un ejemplo. Universitarios e investigadores que necesitan consultar bases de datos bibliográficas o de patentes para asegurarse de que no reinventan la rueda es otro. Finalmente, la Web que, en realidad, es un enorme sistema de información documental con varios miles de millones de documentos, es el ejemplo extremo de contexto característico de RI.

Los sistemas de RI no son los únicos sistemas de información que existen. En tal sentido, en relación a otros métodos de procesamiento de la información, la RI presenta algunos rasgos bien definidos que presentamos a continuación:

- *Primero*, aunque la RI utiliza ordenadores, como casi cualquier otro sistema de información, la intervención de los mismos varía mucho, yendo desde sistemas de RI *totalmente basados* en ordenador a sistemas de RI *asistidos* por ordenador.
- *Segundo*, gestiona información de cualquier tipo, desde textos hasta vídeos, pasando por reproducciones de arte o fotografías, pero siempre mediante la utilización intensiva de información *textual*.
- *Tercero*, tiene lugar en lo que aquí llamaremos un *contexto de descubrimiento*.

El significado detallado de los tres rasgos precedentes es el siguiente:

> *1. Uso de ordenadores (automatización).* La RI se caracteriza por el uso de ordenadores y, por tanto, por el uso de bases de datos u otros sistemas automáticos o semi automáticos de procesamiento de la información, tales como hipertextos. Aunque es lógicamente posible desarrollar sistemas de RI exclusivamente manuales, la teoría (y la práctica) de la RI nació de hecho con las primeras bases de datos y la mayoría de sus procedimientos o algoritmos sólo tienen sentido en un medio automatizado.
>
> *2. Uso de información textual.* La RI gestiona información textual de tipo narrativo o discursivo, en lugar de, por ejemplo, datos numéricos o alfanuméricos muy estructurados, como hacen otros sistemas de información, por ejemplo, los sistemas administrativos (Salton; McGill, 1983: viii). Sin embargo, cuando la RI gestiona documentos u objetos no textuales, como imágenes, fotografías, vídeo, etc., lo hace también a través de descripciones textuales (p.e., descripciones de las imágenes) y/o de conjuntos de palabras que expresan el contenido y el contexto de las imágenes (p.e. palabras clave).
>
> *3. Contexto de descubrimiento*. La RI se caracteriza por tener lugar en un contexto en el cual los usuarios del sistema de información tienen la necesidad de descubrir qué entidades cumplen una o más condiciones, por ejemplo, qué documentos contienen información relevante para interpretar, desde el punto de vista *x*, el tema *y*. En otros sistemas de información, en cambio, los usuarios, partiendo de una entidad previamente conocida, quieren saber algo más de ella.

> La diferencia entre *descubrir cosas* y *ampliar datos* es esencial para entender la naturaleza de la RI.

Algunos desarrollos en sistemas de información son ineficaces porque sus diseñadores no entendieron esta última característica. Por ejemplo, un sistema de información documental automatizado mediante el uso de una base de datos relacional probablemente no podrá satisfacer la necesidad de *descubrir*, aunque solucione muy bien la necesidad de *ampliar*.

En concreto, como sistema documental su utilidad probablemente será parcial, porque las preguntas de descubrimiento, las que son del estilo ¿qué documentos contienen información relevante sobre los temas *x* e *y* (p.e.: "documentos sobre museos y turismo")?, no podrá contestarlas de manera eficiente. Sólo dará un buen rendimiento ante preguntas de *ampliación de datos*, de la forma ¿cuál es el valor del parámetro *a* en el registro *X* (por ejemplo: "cuál es el teléfono del Museo del Prado")? Naturalmente, de un buen sistema de RI se espera que pueda satisfacer preguntas de ampliación de datos como la anterior pero, sobre todo, se espera que pueda responder a preguntas de descubrimiento.

Para profundizar un poco más en esta idea, cabe señalar que la RI está relacionada con la gestión de documentos que contienen informaciones culturales, científicas y técnicas y, más concretamente, con el problema de cómo explotar el conocimiento que incluye esta clase de publicaciones.

Entendemos por información científica el resultado de aplicar el método científico, que es hipotético-deductivo, a un problema de conocimiento, y su expresión en forma de proposiciones contrastables, argumentos, explicaciones, etc. La técnica es ciencia aplicada, y entendemos por información técnica el resultado de aplicar alguna rama de la ciencia a un rango de problemas concretos. Por otro lado, el concepto de información cultural es mucho más amplio. Un artículo de opinión puede contener conocimientos muy valiosos, y formar parte, por tanto, de la alta cultura, pero no es ni científico ni técnico. Algo parecido podría decirse de un buen ensayo, un reportaje periodístico, etc.

Para referirnos a esta triple clase de documentos (científicos, técnicos, culturales), y siguiendo a Van Slype (1988: 1-3), utilizaremos, en adelante, el término *información cognitiva* en lugar de la expresión habitual de *información científica y técnica*. Además de ser un término más económico, hace más justicia a la clase de información que constituye el objeto de estudio y de tratamiento de la Documentación. Así pues, el término cognitivo subsume no sólo a la información de tipo científico y técnico, sino, en general, a toda forma de producción cultural.

Muchos documentos cognitivos son, en sí mismos, narraciones textuales, aunque también contienen partes no textuales, tales como gráficos e ilustraciones, como es común en la información que publica la prensa escrita y muchas revistas científicas.

Por otro lado, los sistemas de RI utilizan descripciones textuales para gestionar también documentos no textuales, tales como fotografías o filmaciones audiovisuales. De esta forma, la manipulación de información textual es típica de la RI.

A los documentos *cognitivos* se oponen los *administrativos*. Para advertir la diferencia esencial entre ambas clases de documentos, basta con practicar un sencillo experimento mental: piense el lector en una enciclopedia. Esto es información cognitiva. Piense ahora en una factura. Esto es información administrativa. Son dos casos extremos, pero nos ayudan a visualizar las diferencias.

La *información cognitiva* es útil, por ejemplo, para aumentar nuestros conocimientos sobre algún aspecto de la naturaleza o, simplemente, para que la humanidad no se vea obligada a reinventar la rueda en cada generación. La *información administrativa*, en cambio, es necesaria para la gestión de cualquier institución y para administrar de forma eficiente sus recursos propios y poder realizar de forma adecuada sus actividades de explotación.

Como es obvio, ambas clases de información son absolutamente necesarias y es evidente que no existe jerarquía entre ellas (p.e., la información cognitiva no es más importante que la administrativa, etc.), pero su naturaleza, ciclo de vida, forma de consumo y propiedades semánticas son distintos y, por tanto, su tratamiento debe serlo también. Muchos sistemas de información fracasan por no advertir esa diferencia: el error más común consiste en gestionar la información cognitiva como si fuera administrativa (aunque también se da el error contrario).

En realidad, los típicos errores mencionados consisten en no observar que la información cognitiva es de carácter probabilístico, y la información administrativa, determinista. En efecto, nunca podremos saber, de entrada, cuáles de los atributos de contenido o de los atributos de contexto de un documento cognitivo pueden hacer de éste una respuesta adecuada a un futuro problema de información. Para peor, no existe ningún algoritmo que sea eficiente al cien por cien para determinar cuáles son, de hecho, los atributos semánticos relevantes de un documento cognitivo (Blair, 1990: 1-23; Blair, 2002).

Naturalmente, el *objetivo* de las operaciones de RI, como ya hemos señalado, consiste en intentar solucionar los problemas de información que requieren información cognitiva. Esta clase de necesidades de información la experimentan, en realidad, todos los seres humanos, puesto que todos ellos, al menos en alguna etapa de su vida, necesitan descubrir, estudiar, aprender o investigar. Ahora bien, para muchos esta necesidad pasa desapercibida o se vuelve transparente: una buena biblioteca, la orientación de un buen experto, etc., contribuyen a esa transparencia. Acceden a los documentos adecuados sin observar que, en las bambalinas, está funcionando alguna clase de sistema de RI.

Sin embargo, para otras personas esta necesidad es una cuestión crítica. Tales personas pueden ser profesionales embarcados en un proyecto de I+D; periodistas realizando un reportaje de investigación; alumnos de doctorado preparando su tesis; ejecutivos de empresa buscando nuevas oportunidades de mercado; médicos de un hospital obteniendo información sobre nuevas terapias; investigadores de un laboratorio que persiguen una nueva patente; profesores de universidad ampliando las fronteras de sus disciplinas, etc. En todos los casos señalados, la satisfacción de la necesidad de información pasará necesariamente por el uso de alguna clase de sistema que contenga información cognitiva.

La información textual es central en casi todos los procesos de RI, aún en el caso de que el fondo documental esté compuesto por objetos no textuales, como fotografías (tal como ya hemos señalado). La razón es que las operaciones básicas de la RI en fondos icónicos (como fototecas o videotecas), a saber, la descripción y la recuperación, se realizan en base a textos que, o bien describen las características de las imágenes o bien describen la necesidad de información. Es por ese motivo que, aunque los documentos icónicos (gráficos, ilustraciones, fotografías, imagen animada, etc.) también forman parte del contexto propio de la RI, no alteran el carácter predominantemente textual de la RI.

Hemos insistido ya que la selección de documentos a partir de su contenido como un aspecto muy significativo de la RI. Ahora bien, para poder seleccionar documentos por su contenido, es necesario, previamente, (1) identificar y (2) representar ese contenido. Esta doble operación se denomina *indización*. La indización puede realizarse, bien de forma intelectual ("a mano") o bien de forma automática (mediante ordenadores).

En este sentido, hay dos interpretaciones distintas, en general de tipo implícito, sobre la naturaleza de la RI. Según la primera, se entiende de manera implícita, como decimos, que los sistemas de RI son exclusivamente automáticos, es decir, realizan los procesos anteriores sin ningún tipo de intervención humana significativa. Esta visión de la RI es la que acostumbra a encontrarse en la bibliografía científica de orientación informática. Es habitual referirse a esta orientación como *RI algorítmica*, puesto que se centra en los programas o algoritmos que pueden automatizar los procesos de RI: típicamente, la indización y la presentación de la información.

En cambio, según una segunda interpretación, los sistemas de RI abarcan en realidad diversos grados de utilización de ordenadores y, por tanto, diversos grados de combinación de operaciones intelectuales y automáticas en un mismo sistema. Podemos hablar entonces de indización automática *versus* indización asistida por ordenador. Para esta segunda interpretación, los sistemas exclusivamente automáticos son solamente un caso particular dentro de la gran variedad existente de sistemas de RI. Esta visión de la RI es la que suele predominar en la bibliografía científica de las ciencias de la documentación. En esta visión se acepta que los procesos de análisis e indización intelectual y la creación y el uso de los lenguajes documentales asociados a estos procesos, como los tesauros y las clasificaciones, forman parte de la RI, siempre que tales procesos involucren el uso de ordenadores. Es habitual referirse a este enfoque como *RI cognitiva*, puesto que sitúa el énfasis no tanto en los algoritmos concretos, sino en los aspectos propios o cercanos a las ciencias cognitivas: lenguaje, semántica documental, psicología, interacción persona-ordenador, percepción, etc.

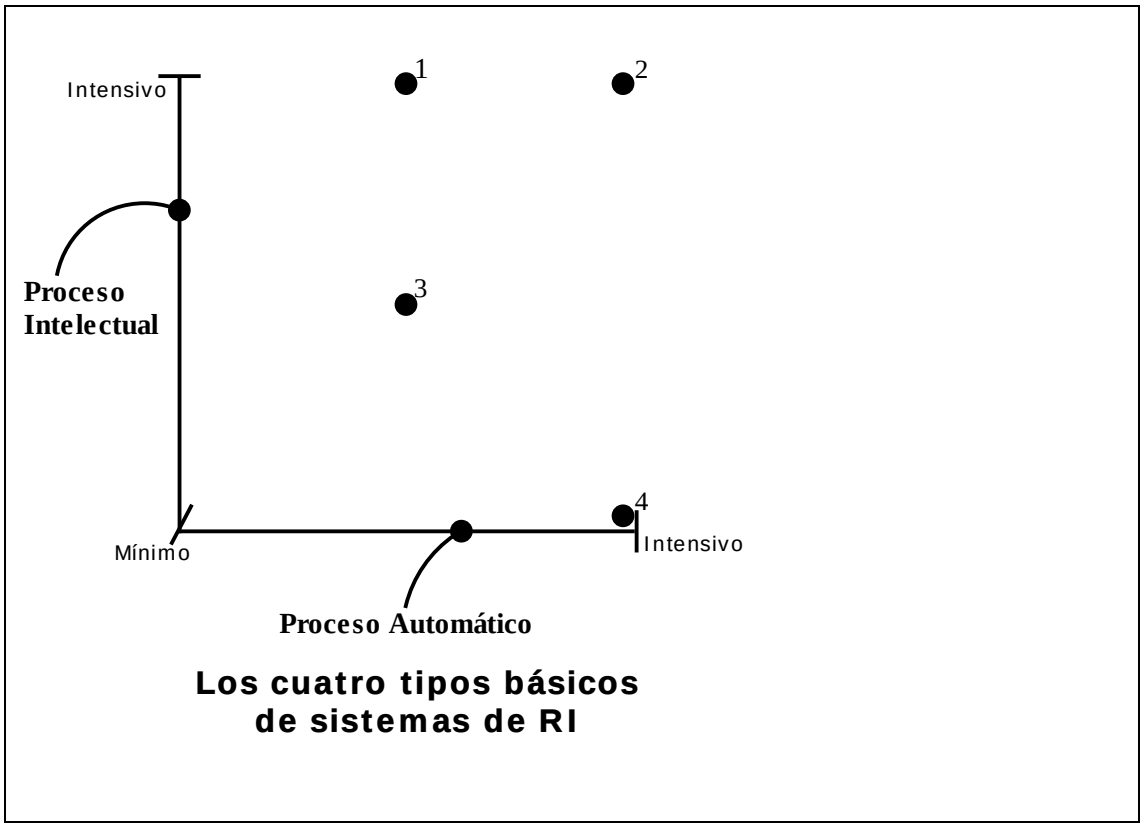
Si obviamos el hecho anecdótico que la RI algorítmica suele ignorar la existencia de sistemas con participación intelectual, ambas corrientes de la RI, en realidad, son igualmente necesarias dado que son perfectamente complementarias. La razón es que, para automatizar un proceso, primero es necesario comprenderlo y modelarlo de manera conceptual. Los estudios de la RI cognitiva aportan los materiales para ese modelado. Por tanto, la RI algorítmica en realidad presupone a la RI cognitiva. Por otro lado, no se ha conseguido automatizar todos los procesos propios de la gestión documental. Hasta que no se consiga (si es que se logra alguna vez), la RI cognitiva es lo único que tenemos en esos terrenos.

De hecho, mientras la RI algorítmica no suele incluir a la cognitiva, lo contrario no es cierto, ya que la RI cognitiva incluye a la algorítmica, aunque sea, como suele decirse, en forma de visión a mil metros de altura…

Es por eso que, en esta obra, nosotros optamos preferentemente por el enfoque cognitivo, ya que es el que, de facto, coincide con la situación real de los sistemas de gestión documental, donde encontramos una mezcla de procesos de tipo algorítmico (donde el profesional de la documentación no suele intervenir) con operaciones de tipo cognitivo (donde el profesional de la documentación debe orientar la mayor parte de sus decisiones). A partir de ahora, por tanto, siempre que nos refiramos a la RI lo haremos en el contexto de la RI cognitiva.

En este sentido, si desplegamos las diversas posibilidades de combinación de procedimientos intelectuales y automáticos en un eje de coordenadas de dos dimensiones, con el componente intelectual a la izquierda y el automático a la derecha, obtendremos el diagrama que nos muestra la figura siguiente donde podemos ver que existen cuatro grandes posibilidades lógicas de combinación:

**Figura 2.1: Sistemas de RI**




De este modo, en el diagrama anterior vemos representadas los cuatro tipos básicos de sistemas de RI:
- **1** (arriba y en el centro): sistemas que son intensivos en procedimientos intelectuales y semi intensivos en procedimientos automáticos;
- **2** (arriba y a la derecha): sistemas que son intensivos tanto en procedimientos intelectuales como automáticos;

- **3** (en el centro): sistemas semi intensivos tanto en procedimientos automáticos como intelectuales;
- **4** (abajo y a la derecha): sistemas que son intensivos en procedimiento automáticos y que no utilizan procedimientos intelectuales.

Obviamente, la clasificación anterior es de grano grueso: existen muchas otras posibilidades que no mostramos en el diagrama. La razón, además de la deseable claridad, es que o bien no tienen mayor interés por tratarse de simples variaciones de uno de los cuatro casos básicos, o bien tales variaciones, sencillamente, no se dan en la realidad. Por tanto, el rango de sistemas significativos que incluye la RI desde el punto de vista de la combinación de procedimientos intelectuales y automáticos puede reducirse de modo conveniente a los cuatro mostrados, según recoge también la tabla 1, donde aportamos una presentación más detallada:

**Tabla 2.1: Tipos básicos de sistemas de RI**

| **Sistema** | **Procesos Intelectuales** | **Procesos Automáticos** | **Explicación y ejemplos** |
|---|---|---|---|
| **Tipo 1** | Intensivos | Semi intensivos | Bases de datos referenciales con uso de herramientas complejas de indización intelectual y sin indización de texto completo (Ej: *LISA, ERIC*). |
| **Tipo 2** | Intensivos | Intensivos | Bases de datos con indización de texto completo y con uso de herramientas complejas de indización intelectual (Ej.: *Sosig*). |
| **Tipo 3** | Semi-intensivos | Semi-intensivos | Bases de datos referenciales con uso de herramientas simples de indización (p.e. listas de descriptores). Como ejemplo, se puede decir que es un caso típico de muchas bases de datos de uso corporativo. |
| **Tipo 4** | Ninguno | Intensivos | Motores de búsqueda (*Google*), programas de indización de texto completo (*ZyLab*), bases de datos con indización exclusivamente automática (*FindArticles*). |

Como puede verse, en los actuales sistemas de RI pueden existir casos en los que no haya intervención intelectual (tipo 4), pero no existen casos en los que no intervengan ordenadores. Cabe señalar también que, en la explicación del tipo 3, no hemos indicado ningún ejemplo de base de datos que pueda consultarse de forma externa. La razón es que se trata del tipo de sistema de RI que es frecuente en el uso privado (personal o corporativo) de bases de datos.

## 2.2. Disciplina

### 2.2.1. Inicios y desarrollo

Como campo de estudio, la RI recibe el nombre de *Teoría de Recuperación de información* (Teoría de RI, a partir de ahora). Se trata de un ámbito (relativamente) interdisciplinar al que contribuyen especialistas procedentes de disciplinas diversas, pero sobre todo de la Informática y de las Ciencias de la Documentación. Obtiene aportaciones valiosas y frecuentes de la Lingüística y la Terminología, así como (aunque en menor medida) de la Psicología y las Matemáticas.

El antecedente más remoto de la Teoría de la RI se sitúa entre los años 30 y 40 del pasado siglo y consisten en los trabajos del estudioso del lenguaje George Kingsley Zipf (1902-1950), descubridor de una ley que lleva su nombre (también llamada "distribución de Zipf") según la cual la frecuencia de las palabras de un corpus representativo de una lengua obedece a la siguiente relación:

| Frecuencia x Rango = Constante |
|---|

*Frecuencia* es el número de veces que aparece una palabra y *Rango* es el número de orden de la palabra listadas en orden decreciente de frecuencias, de manera que la primera palabra es la más frecuente y la última es la menos frecuente.

Por tanto, la ley de Zipf indica, entre otras cosas, que si tomamos una muestra suficientemente grande de textos de una lengua, observaremos que habrá miles de palabras que tendrán valores de ocurrencias muy bajos, por ejemplo entre 1 y 10 veces, en cambio habrá unos centenares de palabras que tendrán valores de ocurrencias muy altos, por ejemplo entre 10.000 y 100.000 veces. De este modo, en una distribución de Zipf ideal, el número total de palabras distintas es igual al número de veces que aparece la palabra más frecuente. Si la colección documental tiene un total de 100.000 palabras distintas, la palabra más frecuente sucederá 100.000 veces y la última palabra en orden decreciente de frecuencia, o sea la palabra número 100.000, ocurrirá 1 vez.

En colecciones reales de documentos, la distribución obtenida no tiene por qué ser idéntica a la distribución de Zipf, pero se ha comprobado que se aproximan de manera suficiente al ideal como para poder realizar predicciones útiles.

En síntesis, los trabajos de Zipf demostraron que era posible detectar regularidades de tipo estadístico en grandes masas de informaciones textuales y que tales regularidades, debido a su carácter estructural, eran susceptibles de ser usadas con fines de planificación de procesos de análisis y de indización de documentos.

En los años 50, un investigador de la empresa IBM, Hans Peter Luhn (1896-1964), postuló la creación automática de índices utilizando tales regularidades. Entre otras cosas, propuso el concepto de "poder de resolución" de un término. El poder de resolución es la capacidad que posee una palabra para identificar de manera no ambigua el tema de un documento. Este poder de resolución está relacionado con la frecuencia del término en un conjunto de documentos.

La idea es extremadamente simple: para seleccionar de manera automática las palabras que deben formar parte de un índice deben evitarse las palabras que son muy frecuentes en el conjunto de los documentos, de lo contrario el índice sería muy poco útil ya que casi todos los documentos tenderán a poseer esas palabras.

Por tanto, se dice de tales términos muy frecuentes que tienen escaso poder de resolución, o poca "capacidad de discriminación". Visto de otro modo: si se indizan documentos utilizando términos de baja capacidad de discriminación, todos los documentos tienden a parecerse, sin que sea posible crear grupos separados. En ese contexto, seleccionar un documento entre otros en base a su contenido es imposible.

En cambio, los términos con baja y mediana frecuencia en el conjunto de los documentos son los que poseen mayor capacidad de discriminación a la hora de construir índices. Si se indizan los documentos con tales palabras, se crean grupos temáticos bien definidos, muy separados entre ellos. Encontrar así documentos en base a su perfil temático es mucho más fácil.

La teoría de RI evolucionó de manera progresiva hasta que dió un salto cualitativo muy importante con autores como Gerard Salton (1927-1995), y C. J. van Rijsbergen.Salton sistematizó los principios de la teoría de RI de tipo algorítmico en un importante trabajo de 1983 (escrito en colaboración con M. J. McGill) que sigue siendo uno de los mejores sobre el campo. Continuó desarrollando su trabajo en su libro de 1989 (esta vez en solitario) y en numerosos artículos que fue publicando hasta bien entrados los años 90. En los dos libros indicados (1983, 1989), Salton proporcionó una visión sólida y unificada de la disciplina y presentó los procedimientos y conceptos más importantes, sobre todo de la RI algorítmica. Rijsbergen, por su parte, enriqueció la Teoría de RI con estudios de tipo lógico y estadístico.

Posteriormente, numerosos autores han contribuido a la disciplina desde enfoques diversos. Para los interesados en profundizar en la RI, posiblemente, los autores actuales más prestigiosos sean (por orden de "antigüedad"): W. F. Lancaster, Edward A. Fox, Gary Marchionini, David C. Blair, Ricardo Baeza-Yates, Richard K. Belew y Gobinda Chowdhury. En nuestro país, también podemos encontrar grupos de investigación que están haciendo avanzar el campo, tanto desde la especialidad de la RI algorítmica, como la RI cognitiva y se están consolidando reuniones científicas, como sería el caso, entre otras, de las *Jornadas sobre Organización, Tratamiento y Recuperación de Información* (JOTRI) y de las *Jornadas Españolas de Documentación*, que reúnen a los principales expertos españoles de esta especialidad.

### 2.2.2. Operaciones de RI

Como ya hemos señalado, el objetivo final de la RI es el estudio y desarrollo de los métodos, bien algorítmicos (preferentemente) o bien intelectuales (cuando no es posible su automatización), que faciliten al máximo el siguiente grupo de operaciones:

1. *Indización.* La indización puede aplicarse a los documentos y a la necesidad de información. Podemos hablar, por tanto, de indización de documentos y de indización de preguntas. En ambos casos, el resultado es un conjunto de descriptores. En el caso de la necesidad de información, los descriptores de la

pregunta pueden estar relacionados entre sí con operadores lógicos (operadores booleanos). Esta operación, en particular cuando se realiza en modo intelectual, se divide en realidad en otras dos:

1.1. *Análisis:* identificación de los temas o conceptos más relevantes del documento o de la pregunta.

1.2. *Normalización*: transformación de los conceptos que expresan el contenido del documento (o de la pregunta) en los términos de indización (descriptores) más adecuados. A veces, esta segunda fase recibe también el nombre de indización, obviando o dando por supuesto a la primera.

2. *Selección*: identificación del conjunto de documentos más relevante para una necesidad de información dada. También se denomina *recuperación* (en este caso, debido a que es la parte más significativa del proceso, a menudo sirve para dar nombre al todo).

3. *Ordenación:* determinación del orden más adecuado de presentación al usuario de los documentos seleccionados o recuperados (en caso que sean más de uno, claro). La idea es ofrecer la lista de los documentos en orden decreciente (el más relevante primero) de probabilidad de satisfacer la necesidad de información. También se denomina *ranking*.

4. *Interconexión:* establecimiento de relaciones hipertextuales, caminos y, en general, estructuras de navegación entre secciones del mismo documento o entre documentos distintos.

5. *Categorización:* asignación de cada documento a un grupo, clase o subclase de un cuadro de clasificación, taxonomía u ontología.

6. *Abstracción*: producción de resúmenes de documentos que, en algunas circunstancias, puedan sustituir la lectura del documento completo.

7. *Visualización:* representación en forma gráfica de informaciones no necesariamente icónicas, así como de conceptos o procesos.

De los siete procesos anteriores, todos están automatizados en algún grado, pero ninguno lo está en modo óptimo. Ante ello, nos podemos preguntar si se podrán automatizar algún día al completo tales tareas. Ignoramos la respuesta, pero sí podemos indicar algunas cosas al respecto: las tareas indicadas están relacionadas con las habilidades más complejas de la condición humana, aquellas que se vinculan con la cognición y el lenguaje. Si, en algún momento del futuro, pudieran automatizarse tales tareas de manera completa y satisfactoria, seguramente esto significaría que las máquinas pueden pensar, en el sentido más profundo y completo del término.

En estos momentos, en tal sentido, lo cierto es que las posturas entre los investigadores están divididas: la mayor parte de la comunidad de ingenieros informáticos afirma, en la más pura tradición de Alan Turing (1912-1954), el gran pionero de la inteligencia artificial, que la inteligencia es una cuestión de conducta observable: si un sistema *se comporta como si fuera* inteligente, entonces *es* inteligente. En cambio, otros

investigadores procedentes de la lingüística y las ciencias cognitivas, como por ejemplo, John Searle o de la física, por ejemplo, Roger Penrose, niegan la posibilidad de que algún día las máquinas puedan pensar, al menos en el sentido del término *pensar* que atribuimos a la especie humana.

A las Ciencias de la Documentación como actividad profesional y como campo de investigación le conviene apurar los límites e intentar aprovechar al máximo las posibilidades de la RI algorítmica. Además de razones de índole social que ya justifican por sí solas esta meta, tales como el poner de forma más eficiente el conocimiento al alcance de la sociedad, existe un motivo egoísta: en la Documentación se cumple también el principio según el cual, cada vez que la informática automatiza una determinada tarea, lo que hace en realidad es liberar a los profesionales de la Documentación de una tarea repetitiva y tediosa y les proporciona recursos (es decir, tiempo, entre otras cosas) para ocuparse de aspectos mucho más creativos de su profesión. Por consiguiente, a los profesionales y académicos de la Documentación, nos conviene no solamente seguir muy de cerca, sino promover y contribuir a los avances de la RI.

## 2.3. Modelos básicos en RI

Un modelo en RI es una representación simplificada que sirve para alcanzar una comprensión global de un sistema, sin necesidad de descender a los detalles concretos. La simplificación puede realizarse por abstracción o por generalización. La abstracción prescinde de detalles accidentales y selecciona solamente los aspectos fundamentales del objeto modelado. La generalización elige representar sólo los aspectos comunes a los diversos objetos modelados. Por ejemplo, un diagrama de flujo de datos es una representación por abstracción de alguna actividad o función.

Muchos modelos se generan por medio de ambos mecanismos combinados: la abstracción y la generalización. Por ejemplo, una base de datos documental es un modelo de una parte de la realidad que combina abstracción (sólo algunos atributos de las entidades reales se representan en la base de datos) y generalización (todas las entidades similares se representan en un único modelo de registro).

En general, la función de un modelo es la de facilitar la comprensión global de un objeto o de un fenómeno complejo, así como la de facilitar el intercambio de ideas entre los componentes de un equipo de personas que trabajan con un objetivo común y que, de este modo, pueden estar seguras de que todos utilizan un lenguaje y un aparato conceptual común. Manejar un modelo es más fácil que manejar la realidad, de este modo, un modelo cumple una tercera función sumamente valiosa: ayudar a crear y a desarrollar objetos.

Un modelo conceptual de un sistema de RI describe la estructura general, las funciones básicas y los aspectos lógicos de una determinada forma de representar la información y de seleccionar documentos relevantes. Estos modelos conceptuales se encarnan en tecnologías concretas o en programas concretos, cada uno de los cuales puede presentar pequeñas o grandes variaciones de implementación del modelo general. En RI se suelen utilizar diversos modelos sobre los cuales se pueden realizar, después, desarrollos concretos.

Autores como Belkin, Croft, Risjbergen y Salton (entre otros) contribuyeron a generar lo que denominamos *Modelo Universal de un Sistema de RI* y que vamos a exponer a continuación. Decimos que es universal porque, en principio, todos y cualquier sistema de RI se ajusta a este modelo. Es universal también, porque como puede suponerse, es de un gran nivel de abstracción. En la práctica, por tanto, cada sistema concreto de RI lo desarrolla de una forma distinta. Ahora bien, gracias a este modelo podemos comprender y estudiar mejor la naturaleza de los sistemas de RI.

Una forma muy abstracta (pero muy comprensiva) de presentar este modelo es la que recoge el siguiente diagrama:

**Figura 2.2. Modelo Universal de un SRI**

NI -> *Análisis* -> RNI -> *Comparación* <- RTEXTO <- *Análisis* <- TEXTO

**NI**: necesidad de información (la pregunta del usuario)
**Análisis:** puede ser totalmente automático, totalmente intelectual, o una combinación de ambos
**RNI:** representación de la necesidad de información (representación de la pregunta)
**Comparación:** función que genera los resultados mediante la comparación de la representación de la pregunta con la representación del texto,
**RTexto**: representación del texto (representación del documento)
**TEXTO**: documento

**Observación:** en redonda, las entidades; en cursiva, los procesos

Fuente: adaptado de Belkin y Croft (1987)

En los siguientes apartados, tendremos ocasión de discutir con detalle los componentes que hemos presentado aquí de forma tan abstracta y concisa.

Los sistemas que desarrollan el sistema básico anterior se denominan *Sistemas de Recuperación de Información* (SRI a partir de ahora). Estos sistemas pueden consistir en programas informáticos o conjuntos de rutinas aislados o bien pueden estar integrados en el seno de un *Sistema de Gestión de Bases de Datos Documentales* (el caso que consideramos en los próximos capítulos).

En cualquier caso, los componentes principales de un SRI son los siguientes (ver la figura 2.2.):

- La entidad *necesidad de información* (1), también llamada *pregunta*.
- La entidad *texto –documento*-- (2) que, a su vez, forma parte de un fondo documental (3) más amplio.
- *Representaciones*, tanto de los documentos (4a) como de las necesidades de información (4b).
- Una función de *comparación* (5) entre la representación de la necesidad de información y la representación de los documentos del fondo documental, que tiene

por objeto determinar cuáles son los documentos más *relevantes* (6) para cada necesidad de información.

Finalmente, los documentos se muestran en uno o más *formatos de visualización* (7) y el proceso tiene lugar en lo que hemos denominado con anterioridad como un *contexto de descubrimiento* (8). A continuación, examinaremos con más detalle los ocho ítems señalados como característicos de un SRI.

### 2.3.1. Necesidades de información

Una necesidad de información es, por definición, una entidad inobservable, ya que consiste en un estado mental. Se supone que este estado mental o psicológico es el auténtico disparador de todo el proceso de RI, ya que, a partir de experimentar una necesidad de información, el individuo puede comenzar una conducta de búsqueda de información que, a su vez, puede tener su punto de inicio en la selección de la fuente de información que considere más adecuada.

Si la fuente de información consiste en un fondo documental de gran volumen, el individuo deberá desplegar alguna estrategia de examen del fondo que sea rentable, sobre todo, en términos de tiempo y que pueda conducirle, eventualmente, a encontrar información susceptible de solucionar su necesidad de información. En concreto, evitará realizar una exploración secuencial. Por ejemplo, si el fondo documental es una biblioteca, consultará su catálogo en lugar de mirar secuencialmente, uno a uno, los libros de las estanterías.

### 2.3.2. Documentos

Un documento es una información que está registrada, es decir, una información que está escrita, impresa, grabada, etc. en un soporte material. En el contexto de la RI se presupone que se trata de documentos de los denominados *cognitivos*, es decir, documentos que contienen obras culturales, técnicas o científicas.

Dicho de otra forma, los SRI no suelen aplicarse, porque tendría muy poco sentido, a la gestión de datos e informaciones administrativas, como los que intervienen en el sistema de contabilidad o de facturación de una empresa.

Por el contrario, en el contexto de la RI se da por supuesto que el problema a tratar tiene que ver con obras de creación sobre algún aspecto de la ciencia, la técnica, la cultura, etc. Es decir, la clase de documentos con un contenido mínimamente complejo que justifica la aplicación de procesos de RI.

### 2.3.3. Fondo documental

Los procesos de RI solamente tienen sentido en el contexto de un fondo documental no trivial. Buscar información en el seno de una colección compuesta por unas decenas de documentos no requiere un sistema de RI, ya que se puede explorar todo el fondo con una exploración secuencial.

La RI comienza a tener sentido cuando deben gestionarse colecciones de, al menos, varios cientos de documentos. En el límite, uno o varios sistemas cooperativos de RI deberían poder aplicarse al conjunto de toda la información producida por la humanidad; perspectiva no tan fantástica como podría parecer a primera vista si reflexionamos sobre las posibilidades futuras de la World Wide Web.

### 2.3.4. Representaciones de documentos y representaciones de necesidades de información

En un sistema de RI no podemos (o no resulta conveniente) intentar comparar directamente documentos y necesidades de información. Lo que se compara, en su lugar, son representaciones de cada una de las dos entidades mencionadas. La razón es que no es eficiente comparar dos elementos de naturaleza heterogénea: una necesidad de información es un estado mental, inobservable por definición, y los documentos son conjuntos de informaciones de morfología variable registrados en un soporte material. Para que sea posible comparar cosas tan dispares en su naturaleza, es necesario convertir ambas a una representación formada por elementos homogéneos.

La representación del documento puede consistir en un típico registro bibliográfico articulado en zonas como las que prescribe la norma ISBD más una descripción del contenido del documento formada por descriptores.

Ahora bien, desde el momento en que un documento textual o icónico se representa de este modo, entonces, desde el punto de vista de un ordenador (de una base de datos, en este caso) este registro es un conjunto de palabras o, más exactamente, un conjunto de términos de indización.

Si formalizamos esta idea, por tanto, en un sistema de RI un documento es un conjunto, *D*, los elementos del cual son términos de indización según este modelo general:

$$D = \{t1, t2, ... tn\}$$

Por tanto, en el modelo anterior, t1, t2,... t*n* son, o bien palabras simples (p.e., "economía", "Barcelona") o bien compuestas (p.e., "economía política", "Ciudad Condal") que expresan las propiedades semánticas del documento $D_i$.

Por ejemplo, supongamos, para simplificar, que el documento $D_i$ trata sobre 5 temas diferentes que identificamos de manera abstracta como tema 1 (o t1), tema 2 (o t2), etc. Entonces, la representación de $D_i$ en un sistema de RI seria la siguiente:

$$D_i = \{t1, t2, t3, t4, t5\}$$

Supongamos que el documento en cuestión trata de "legislación sobre economía y trabajo en España y Europa", entonces t1, t2, t3, t4, t5 corresponderían respectivamente a:

Economía (t1)
España (t2)
Europa (t3)
Legislación (t4)
Trabajo (t5)

Por tanto, el documento se podría representar así:

$D_i$ = {Economía, España, Europa, Legislación, Trabajo}

La cuestión interesante aquí es que las necesidades de información se pueden representar también, como ya sabemos, por términos de indización, según el mismo modelo general:

$P_j$ = {t1, t2, ... t*n*}

donde $P_j$ es una necesidad de información, o pregunta, y t1, t2, etc. son cada uno de los términos de indización que representan esta necesidad de información. En concreto, supongamos que $P_j$ representa la siguiente necesidad de información: "legislación sobre mujer y trabajo en España".

La representación de la pregunta $P_j$ en base a palabras o términos de indización sería la siguiente:

$P_j$ = {España, Legislación, Mujer, Trabajo}

De este modo, conseguimos que dos cosas muy diferentes en su estado natural, necesidades de información y documentos, adquieran una naturaleza homogénea a través de un modo de representación similar:

$D_i$ = {Economía, España, Europa, Legislación, Trabajo}

$P_j$ = {España, Legislación, Mujer, Trabajo}

Se observa, a simple vista, que ahora ya resulta posible comparar ambas cosas y, dado un grupo de documentos, determinar cuál de ellos se parece más a una necesidad de información. Discutiremos este proceso de comparación en el siguiente punto.

### 2.3.5. Proceso de comparación

Como ya sabemos, uno de los dogmas centrales de la RI es que, dado un grupo de documentos, el que más se asemeje a la necesidad de información, será el documento más relevante. Tal como hemos visto, a partir de la forma que adquieren $D_i$ y $P_j$ en el punto anterior, podemos concluir que tales conjuntos poseen tres elementos en común.

Supongamos que en el fondo documental hay otros dos documentos con elementos comunes a la necesidad de información que estamos considerando ($P_j$). Por ejemplo, sean los documentos $D_h$ y $D_g$. Supongamos que $D_h$ tiene dos elementos en común (es decir, dos términos de indización en común) y que $D_g$ tiene, en cambio, cuatro elementos en común (es decir, cuatro términos de indización en común) con el conjunto de la necesidad de información $P_j$.

A partir de aquí, el subsistema de comparación del sistema de RI podría presentar los documentos por orden decreciente de semejanza con la necesidad de información, de este modo:

1: Documento $D_g$
2: Documento $D_i$
3: Documento $D_h$

La anterior es una ordenación de los documentos en base al grado de probabilidad que presenta cada uno de ellos de satisfacer la necesidad de información. Esto es, tenemos una forma de medir la relevancia de cada documento y de ordenarlos de acuerdo a la misma. Vemos que, en este caso, la relevancia ha sido estimada en función del número de elementos en común entre cada uno de los documentos y la necesidad de información.

Es un modelo muy simple, pero está en la base de la mayor parte de los sistemas de RI que podemos encontrar en el mercado, si bien es cierto que la mayoría también presentan importantes modificaciones.

### 2.3.6. Relevancia

La relevancia es una de las propiedades más interesantes de los documentos y, al mismo tiempo, una de las más difíciles de definir. Intuitivamente, podemos afirmar que un documento es tanto más relevante cuanto mejor pueda solucionar una necesidad de información. Ara bien, definida de esta forma, se ve que la relevancia no es una propiedad exclusiva de los documentos, sino, en realidad una coproducción triangular entre las características del documento, las características de la necesidad de información y las características de la persona que hace la pregunta.

Por otro lado, la relevancia tiene grados, ya que un documento no se limita a ser relevante o a no serlo, sino que la relevancia de un documento (como hemos visto en el ejemplo anterior) puede situarse en cualquier punto de un continuo de entre, por ejemplo, 0 y 1, en el cual el 0 representa la ausencia total de relevancia y el 1 la relevancia absoluta. Entre esos puntos, un documento muy semejante respecto a la pregunta podría tener una relevancia del 0.8, mientras que otro menos similar podría tener una relevancia del 0.5, etc. Naturalmente nada impide utilizar escalas de 0 a 10 en lugar de 0 a 1 o de tantos por ciento para representar el grado de relevancia de cara al usuario.

El punto importante aquí es que, si diferentes documentos tienen un grado de relevancia diferente respecto a una pregunta, entonces no tiene mucho sentido entregar los documentos de una forma aleatoria o bajo un orden tan poco eficiente como el título o la

fecha de entrada en el fondo documental. Sin embargo, mientras este principio está muy asumido en los motores de búsqueda de Internet, todavía es ignorado en algunos sistemas corporativos de gestión documental.

En contraste, una vez aceptado el principio teórico de la relevancia, lo que hacen los mejores sistemas de RI es intentar determinarla de la forma más eficiente posible. De hecho, en grandes fondos documentales la eficiencia del método de determinación de relevancia es un factor crítico que puede condicionar la calidad total del sistema.

Si la respuesta a una pregunta incluye una lista de mil documentos y los documentos relevantes están distribuidos de manera aleatoria entre estos mil, el usuario no sabrá cuando debe detener su búsqueda ya que la información más útil podría estar, precisamente el último documento de la lista. En teoría, en lugar de limitarse a examinar los primeros diez o veinte documentos, debería examinar mil pero tal cosa distaría de ser eficiente. Los usuarios lo considerarán un sistema inviable. En otras palabras: dejarán de usarlo.

### 2.3.7. Descubrimiento

Ya hemos señalado anteriormente, que resulta difícil apreciar correctamente la naturaleza de la RI sin entender la siguiente cuestión: la RI no sirve exclusivamente para saber más cosas de una entidad previamente conocida, sino para descubrir qué entidades cumplen una condición o una serie de condiciones.

Sin comprender esta diferencia no se puede entender cuál es, entonces, la aportación específica de un programa documental comparado con un programa ofimático estándar. En concreto, es imposible distinguir entre un sistema de gestión de bases de datos documental y un sistema de gestión de bases de datos relacional. Otra forma de enfocar esto mismo consiste en señalar que el entorno de trabajo típico de los programas ofimáticos es de tipo determinista, es decir, se sabe siempre lo que se quiere y se sabe que tales acciones producirán siempre tales resultados. En cambio, en el entorno típico de la RI no siempre se sabe lo que se quiere, ni tan sólo se sabe si habrá entidades que puedan satisfacer las condiciones indicadas en la petición de información.

La petición de información típica de un entorno ofimático sigue este tipo de modelo general: “¿qué valor asume la variable *V* de la entidad *E*, previamente conocida?”, Por ejemplo, “¿cuál es el importe total de las ventas del mes de abril de la delegación de París?”. El valor que se quiere saber es “el importe total”; la variable de la que se quiere saber este valor es “las ventas del mes de abril”, y la entidad, previamente conocida, es “la delegación de París”. Aquí tenemos un entorno determinista: ante esta clase de pregunta, tiene que haber siempre una respuesta y tan sólo una única respuesta.

La petición de información típica de un entorno de RI sigue, en cambio, este otro modelo general: “¿qué entidades --desconocidas por definición--, son susceptibles de satisfacer la condición *C* o el complejo de condiciones $C_1$, $C_2$... $C_n$?”. Por ejemplo, “¿qué documentos son más útiles para satisfacer una necesidad de información sobre la relación entre psicología y cine?”. Las entidades desconocidas por definición son los hipotéticos documentos relevantes, y el complejo de condiciones que han de satisfacer los documentos para ser considerados relevantes son, en este caso, tres: tratar de

psicología (1), tratar de cine (2) y que la relación lógica entre (1) y (2) sea la que se expresa con un AND booleano (3).

Aquí tenemos un típico entorno probabilístico: puede existir, o no, una respuesta, y, en caso de que exista, no tan sólo no tiene por qué ser única, sino que, lo más habitual, es que haya una colección de documentos (respuestas) diferentes, cada uno de ellos con un grado de relevancia diferente. Finalmente, aunque el sistema sea capaz de suministrar documentos relevantes, esto puede significar que, en vez de solucionar de manera definitiva la necesidad de información, se le abran al usuario nuevos interrogantes, por tanto, nuevos "estados anómalos de conocimientos", la necesidad de hacer nuevas operaciones de RI, etc.

### 2.3.8. Ordenación y visualización de la información

Una vez seleccionados los documentos, cabe decidir su forma de ordenación de cara a su presentación al usuario. En sistemas de RI simples, la ordenación no es significativa: se suministrarán por orden de número de registro, por ejemplo.

En sistemas de RI avanzados, se presentarán por orden de relevancia, de modo que los documentos juzgados más útiles estarán situados en primer lugar. En algunos sistemas, es posible elegir el tipo de relevancia, o conmutar entre distintos tipos de ordenación: relevancia, fecha de publicación, orden alfabético, etc.

Una vez ordenados los documentos por su grado de relevancia o por cualquier otro procedimiento, el sistema de RI puede tener uno o más tipos de presentaciones de los documentos individuales o de los grupos de documentos, denominadas habitualmente "vistas" o "formatos".

Cada vista puede representar los intereses o las necesidades de diversos grupos de usuarios, o diversos estilos de visualización. Por ejemplo, en el primer sentido, es habitual que haya un formato para los administradores del sistema, otra para usuarios finales, etc.

Algunos motores de búsqueda de Internet, como por ejemplo *HotBot* (http://www.hotbot.com) permiten escoger entre respuestas resumidas o detalladas. En bases de datos como *Special Collections* de NL Search (http://www.nlsearch.com), se presentan tres vistas diferentes de los documentos, según la fase de la búsqueda, siendo más detallada cada vez hasta llegar al documento completo en la última fase. (En el apartado 4.3 se profundiza en estas cuestiones).

Algunos bancos de imágenes también permiten escoger el formato de visualización de las imágenes recuperadas, ni que sea para poder seleccionar entre las dimensiones y el número de imágenes que tiene que presentar el sistema de manera simultánea (véase, por ejemplo, Corbis, http://www.corbis.com).

Por su parte, las técnicas de visualización de la información consisten en mostrar de forma gráfica informaciones que no son necesariamente icónicas. Por ejemplo, la empresa *Cartia* (www.cartia.com) ha desarrollado un sistema para representar en forma de mapa espacial los temas de cualquier grupo de documentos y lo han aplicado a

diversos ámbitos, uno de los cuales es la información de prensa (http://www.newsmaps). La empresa *Inxight* (www.inxight.com) ha producido una interfaz de visualización, denominada *Hiperbolic*, que se puede aplicar a fondos documentales. Se puede ver una demostración aplicada a la base de datos de fuentes de información de *Lexis-Nexis* (www.lexis-nexis.com/lncc/hyperbolic/).

Una vez discutidas algunas de las características generales más importantes de todos (o casi todos) los sistemas de RI, pasaremos a describir tres modelos muy habituales y que, por tanto, están presentes, de una forma u otra, en un gran número de sistemas de gestión de bases de datos y de motores de búsqueda de la Web: el modelo booleano, el modelo vectorial y un modelo mixto que combina aspectos de ambos.

### 2.3.9. Modelo booleano puro

En un sistema de RI booleano, una vez indizados los documentos, y ante la pregunta de un usuario, existen dos objetos, la entidad documento que se representa mediante un conjunto de términos de indización {t1, t2, ... tn} que expresan los diversos temas contenidos en el documento (tema *a*, tema *b*, ... etc.) y la pregunta, que se representa también por otro conjunto de términos de indización {t1, t2... tn}, pero en este caso, combinándose entre sí mediante operadores booleanos (AND, OR, NOT).

La representación de la pregunta, en un sistema booleano, tiene esta forma general:

| |
|---|
| T1 [Operador] T2 |

T1 y T2 pueden ser palabras simples o compuestas (p.e. "Economía", "Gestión cultural"), y [Operador] puede ser cualquier operador booleano (típicamente, AND, OR, NOT).

Por ejemplo:

| |
|---|
| Economía AND Gestión cultural |

Toda expresión de la forma general, T1 [Operador] T2, se denomina ecuación de búsqueda. Se supone que el otro extremo de la ecuación contiene el conjunto de los documentos verdaderos, DV, o documentos que satisfacen la ecuación, según este modelo:

| |
|---|
| T1 [Operador] T2 = {dv}<br><br>{dv}: Conjunto de los documentos que satisfacen la ecuación |

Naturalmente, pueden darse ecuaciones booleanas con más de dos términos y más de un operador, si conviene, con el uso de paréntesis para debilitar el alcance de cada operador, por ejemplo:

| |
|---|
| (Economía OR Financiación) AND (Gestión cultural OR Museos) |

El resultado de una ecuación de búsqueda booleana es un conjunto que contiene los documentos relevantes (en ocasiones, este conjunto puede ser vacío). Éstos se seleccionan, naturalmente, siguiendo la lógica booleana, según la cual un documento es verdadero (es decir, satisface la ecuación) cuando contiene uno o más de los términos de la pregunta (en el caso del operador OR); cuando contiene todos los términos de la pregunta (en el caso del operador AND) o cuando no contiene alguno de los términos de la pregunta (en el caso del operador NOT), respectivamente.

Ahora bien, según la lógica booleana, las variables solo pueden ser verdaderas o falsas, con lo cual los sistemas RI booleanos únicamente pueden crear conjuntos de documentos relevantes o no relevantes, pero sin establecer grados de relevancia entre los documentos relevantes. En concreto, si una operación de recuperación obtiene *n* documentos, digamos 100 documentos, para el sistema booleanos tan relevante es el documento primero como el número cien. Sin embargo, la simple experiencia demuestra que, para el usuario, tal cosa no responde a la realidad, ya que unos documentos le serán más útiles (más relevantes) que otros, además con grandes diferencias entre ellos. Sin embargo, al entregarse los documentos al usuario de forma aleatoria, casi nunca los más relevantes estarán en los primeros lugares de la lista de documentos recuperados. La consecuencia es una lastimosa pérdida de tiempo, pues los *n* documentos examinados hasta llegar al documento realmente relevante representan un tiempo inútil.

Esto ha generado con frecuencias muchas críticas a tal modelo y aquí es donde intervienen los sistemas vectoriales y los sistemas mixtos (booleanos/vectoriales), que son capaces de ordenar los documentos por grado de relevancia.

Otra crítica habitual a los sistemas booleanos es que resultan poco intuitivos. En particular, los usuarios no habituados a la lógica booleana con objetivos de recuperación suelen confundir el OR booleano que siempre es inclusivo en RI, con la conjunción *O* del lenguaje que, a veces es inclusiva y a veces es exclusiva. Por ejemplo, cuando alguien dice "iré esta noche al cine **o** al teatro", obviamente usa un O exclusivo: o bien irá al teatro o bien irá al cine, pero no a ambos a la vez, al menos no aquella noche. En cambio, una búsqueda mediante la ecuación Teatro OR Cine, seleccionará documentos que o bien tengan la palabra Teatro, o bien tengan la palabra Cine o bien tengan ambas palabras, y el sistema considerará válidas a las tres clases de documentos por igual. Otras veces, usamos en el lenguaje la conjunción *Y* con el sentido que en lógica booleana usaríamos un AND. Por ejemplo, alguien puede decir, "necesito información sobre congresos y festivales de cine y televisión". Para representar la anterior necesidad de información, muchos usuarios estarán tentados en transformar cada uno de los *Y* anteriores en AND booleanos; sin embargo, si lo hace así seguramente no encontrará nada, ya que la ecuación correcta sería: (Congresos OR Festivales) AND (Cine OR Televisión)

### 2.3.10. Modelo vectorial puro

Dados *n* únicos términos de indización, tanto los documentos como las preguntas pueden concebirse como vectores formados por uno de los dos valores posibles que puede adquirir cada uno de los términos: 1 si está presente en el documento o en la pregunta, y 0 si no lo está (Salton y McGill, 1983). Como recordará el lector, en un vector la posición de cada elemento es significativa, y su número de elementos es fijo.

De esta forma, si en un sistema de RI se utilizaran sólo seis únicos términos de indización: t1, t2, t3, t4, t5, y t6, un documento D1 que poseyera los términos t1, t3, t4, t5, se representaría con el siguiente vector: (1,0,1,1,1,0), mientras que un documento D2 que poseyera los términos t2 y t6 se representaría como: (0,1,0,0,0,1). Por su parte, una pregunta P1 que se supone representada con los términos t1, t4, t5, se representaría como (1,0,0,1,1,0).

La función de comparación, entonces, se realiza situando los documentos en un espacio vectorial de *n* dimensiones, en nuestro ejemplo, en un espacio vectorial de 6 dimensiones. La situación de cada vector en el espacio vendrá determinada por sus respectivos valores respecto a cada uno de los seis ejes del espacio, y así los documentos más parecidos entre ellos tenderán a situarse próximos en dicho espacio vectorial.

Cuando se representa como un vector, la pregunta "caerá" en algún lugar del espacio vectorial. Cuanto mayor sea la proximidad de un documento respecto al lugar donde ha caído la pregunta, más relevante será el documento. Como los documentos similares tienden a formar grupos (*clusters),* todos los documentos que formen parte de un clúster próximo a la pregunta tenderán a ser relevantes. Así, se puede establecer un umbral de semejanza por debajo del cual un documento se considerará no relevante. Todos los documentos que superen el umbral serán relevantes, pero no en el mismo grado, de manera que, gracias a las propiedades diferentes de cada vector de cada documento, podrán entregarse al usuario ordenados por su capacidad de satisfacer la pregunta del usuario.

El umbral de relevancia hará innecesario, en principio, el uso de operadores booleanos aunque la pregunta contenga dos o más términos. Bastará con colocar en la pregunta todos los términos de la necesidad de información. Como solamente se entregarán al usuario los más relevantes, de hecho, quedarán excluidos los documentos que, por ejemplo, traten de uno solo de los términos de la pregunta en caso de haber documentos que traten de **todos** los términos de la pregunta.

Por otro lado, en teoría se evitan los problemas de la validez o no validez de tipo "todo o nada" (binaria) propia de los sistemas booleanos. En un sistema vectorial, podemos situar el umbral en un nivel muy bajo de modo que, si usamos cinco términos en la pregunta, la lista de respuesta contenga primero los documentos con los cinco términos, pero no por ello excluya los que tienen cuatro o tres de los términos, etc.

El modelo vectorial puro goza de un estatus existencial parecido al de los algoritmos de indización automática. Véase lo que se dirá más adelante sobre ellos, ya que puede serle

enteramente aplicado. Los interesados en los modelos teóricos vectoriales pueden seguir la excelente obra de Salton o de Baeza-Yates.

### 2.3.11. Modelo booleano/vectorial

El modelo precedente, pese a su impecable base lógica, apoyada en espacios vectoriales y teorías de *clusters*, resulta poco implementado en la práctica, seguramente por el uso intensivo de recursos de cómputo que requiere y los problemas de recálculo del espacio vectorial cada vez que se añaden nuevos documentos. Pensemos que, en una base de datos con *n* términos distintos, se requeriría un espacio vectorial de *n* dimensiones. Una base de datos de unos pocos miles de documentos puede fácilmente generar 100.000 términos distintos; por ello, se necesitaría un espacio vectorial de 100.000 dimensiones.

El modelo vectorial, al menos, ha servido para inspirar otras formas en las cuales podrían funcionar los sistemas de RI, así como ha sido una fecunda fuente de ideas para mejorar el funcionamiento de los sistemas booleanos; gracias a estas ideas muchos sistemas de RI aunque tienen un sistema de filtro o de selección de tipo booleano, ordenan después los documentos por relevancia en lugar de considerar que la relevancia es una propiedad binaria.

En cualquier caso, el modelo mixto actúa de la siguiente forma, los documentos y las preguntas se representan como vectores, pero en vez de calcular su similitud en base a clusters y espacios vectoriales, se calculan estimando cuántos elementos en común presentan los vectores respectivos de preguntas y documentos (Frakes y Baeza-Yates, 1992). Por ejemplo, dada la pregunta P1, puede calcularse que el documento D1 exhibe un mayor grado de semejanza que D2 si, por ejemplo, el vector P1 (vector de la pregunta) tiene la siguiente composición (1, 1, 1, 1, 1, 1, 1); el vector D1 (documento 1) tiene la siguiente: (1, 1, 1, 1, 0, 1) y el vector D2 (documento 2) la siguiente: (1, 0, 1, 1, 0, 1). Es fácil ver que D1 tiene cinco elementos en común (todos menos el quinto), mientras que D2 tiene solo cuatro.

El aspecto booleano de este sistema radica en que la selección de documentos se realiza de acuerdo con el álgebra de Boole, pero una vez creado el subconjunto de documentos, éstos se ordenan mediante el método anterior. Otras formas de ordenación pueden incluir ponderación de cada elemento del vector, de manera que cada vector puede multiplicarse por el peso del término en cada documento.

Por ejemplo, supongamos que un usuario desea obtener documentos sobre los ordenadores aplicados a la gestión de documentación periodística. La pregunta se podría representar mediante los conceptos: ordenadores, documentación, periodismo. Combinados tales conceptos con un OR booleano, el sistema podría recuperar *n* documentos, cada uno de los cuales tendría uno o más de los términos de la pregunta.

Supongamos que el documento D1 posee los tres términos y el documento D2 posee sólo dos de ellos. En el sistema no ponderado, el documento D1 es el más relevante, pero en un sistema ponderado podría no ser así. Veamos: supongamos que el documento D1 presenta el siguiente vector (2, 1, 1), que se debe leer así: el término primero aparece dos veces en el documento (2*1=2), el término segundo y tercero aparece una vez (1*1=1).

La suma total de los valores del vector del documento D1 es igual a 4 (2+1+1+1). Supongamos que el vector del documento D2 tiene la siguiente composición (0, 3, 2), lo cual significa que el término 1 no aparece (el término ordenadores), pero en cambio el término segundo (documentación) aparece tres veces, y el término cuarto (periodismo), dos veces. El sumatorio da un valor de 5, por lo tanto, superior al valor del documento D1.

En la práctica puede suceder que, pese a todo, el documento D1 sea más relevante, ya que el usuario puede estar más interesado en documentación automatizada aplicada al periodismo, que no en documentación periodística a secas, y el segundo documento, que solo trata de documentación y periodismo, puede no hacer ninguna mención a sistemas automatizados. De ser así (y es así muchas veces) ello demostraría que los sistemas de comparación vectoriales basados en propiedades estadísticas no proporcionan un 100% de aciertos.

Ante ello, la respuesta es la siguiente: en primer lugar, una ordenación parcialmente eficaz es mejor que ausencia de ordenación. Los sistemas que realizan ordenaciones nunca sitúan al final de todos los documentos más relevantes, sino que siempre quedan situados en los primeros lugares, así que no es tan importante si el primer lugar del ranking debería ocuparlo el documento 2 o el documento 3, en lugar del documento 1. En cambio, en un sistema sin cálculo de relevancia, es perfectamente frecuente que el documento más relevante esté situado en los últimos lugares de la lista recuperada.

En segundo lugar, el ranking puede efectuarse también en base a los documentos recuperados después de una operación booleana con el uso del operador AND, con lo cual el usuario se asegura que todos los documentos recuperados independientemente de su situación en el ranking tratan los tres temas de su interés. Finalmente, el usuario puede ponderar también el vector de la pregunta, e indicar así que, para él, la presencia del primer término debe tener el 60 por ciento del peso, y los otros dos el 20 y el 20 por ciento respectivamente, con lo cual el documento D1, por seguir con nuestro ejemplo hubiera obtenido el siguiente vector: (1.2, 0.2, 0.2), ya que 2*0.6=1.2 y 1*0.2=0.2; por tanto, 1.2+0.2+0.2=1,6; mientras que el segundo vector obtendría un valor de 1.0, ya que, 0+0.6+0.4=1.0.

## 2.4. Representación de la información

### 2.4.1. Clasificar

Hasta que se utilizaron ordenadores en los centros de documentación y bibliotecas, la idea de ordenar un fondo documental de cara a su posterior recuperación se limitó con frecuencia a la asignación a cada documento de una categoría o clase de una clasificación en aplicación del viejo ideal de "un lugar para cada cosa y cada cosa en su lugar".

La idea de que es posible construir clasificaciones que sean perfectas desde el punto de vista lógico posee una fuerza enorme, por tanto, no es extraño que siga apareciendo espontáneamente en la cabeza de las personas que se ocupan de estas tareas por primera vez. Sin embargo, cualquiera que haya acometido la tarea de desarrollar un cuadro de clasificación para acomodar de manera unívoca objetos de una mínima complejidad,

como artículos de revistas científicas, por ejemplo, habrá comprobado la imposibilidad práctica de aplicar el anterior principio, porque cada cosa puede estar en más de un lugar, y no siempre los lugares prefijados sirven para acomodar a todas las cosas.

De hecho, si la analizamos con atención, la gestión documental basada en clasificaciones presenta estas características:

*1. Limitación de puntos de acceso:* el número de categorías a las que puede ser asignado un documento es, a veces por razones pragmáticas y a veces por razones intrínsecas y, más frecuentemente, por ambas razones, extremadamente limitado. De hecho, el número de categorías a las que se asigna un documento suele oscilar entre uno y tres.

2. *Limitación ontológica:* toda clasificación implica una concepción del mundo. Sin embargo, las formas de ver el mundo, los puntos de vista, los intereses, etc., de los autores de los documentos no tienen por qué coincidir con la forma de entender el mundo, los intereses, etc., de los autores de los cuadros de clasificación. Igual sucede con los usuarios: su punto de vista no tiene por qué coincidir ni con el de los autores ni con el de los documentalistas. Por tanto, si el autor trata sobre un aspecto de la realidad que no estaba contemplado cuando se concibió la clasificación, o lo aborda de un punto de vista ajeno a la concepción de la clasificación, ni el aspecto ni el punto de vista podrá ser representado. Por su parte, si el usuario no "piensa" en términos del sistema, no podrá encontrar la información.

3. *La limitación sintáctica*: en el argot de los lenguajes documentales se dice que las clasificaciones son lenguajes precoordinados. La razón es la siguiente, si una clasificación, por ejemplo, contempla el concepto de la maquinaria para usos de excavación y perforación en minería, encontraremos una entrada como ésta (el ejemplo está tomado de la CDU):

622.23.05 Minería, trabajos de excavación. Maquinaria

En este caso, se dice que se trata de un lenguaje precoordinado porque la relación entre los términos Minería (622), trabajos de excavación (23) y maquinaria (05) se ha establecido *a priori*, antes, e independientemente, de las consultas de los usuarios.

Otra forma de contemplar la precoordinación es la siguiente: en las clasificaciones, se parte de ámbitos muy generales que constituyen las categorías o clases principales y se va descendiendo a subclases o subcategorías más específicas. Por ejemplo, si deseamos acceder a información sobre "lámparas eléctricas", debemos empezar en la clase “6 Ciencias Aplicadas”, descender a la subclase “62 Ingeniería”, seguir bajando por el árbol lógico de la CDU a la subclase “621 Ingeniería mecánica en general”, seguir bajando hasta “621.3 Ingeniería eléctrica” hasta llegar, finalmente, a la sección “621.32 Lámparas eléctricas”. De este modo, vemos que el concepto “Lámparas eléctricas” está precoordinado con el concepto de “Ciencias Aplicadas” en una relación, en este caso, de tipo jerárquico.

### 2.4.2. Indizar

Se produjo un gran avance en la gestión documental cuando se aplicó un principio totalmente distinto, y muy característico de la RI: en lugar de intentar encajar cada documento en una única categoría *a priori*, lo que se hace es lo siguiente: primero, se determina cuál es el conjunto de características semánticas específicas y representativas de cada documento; segundo, se representa cada documento en base a todas y cada una de esas características, sin necesidad de precoordinarlas de forma alguna. En general, el conjunto de características adopta la forma de un conjunto de términos, incluso en el caso de documentos no textuales.

La operación anterior, como ya sabemos, se denomina *indización*. La razón de este término es la siguiente: cada una de las palabras que se utilizan para indicar sobre *qué* trata un documento es una entrada de un *índice* que facilita la consulta y la recuperación de los documentos. Observemos que la base lógica de esta operación es la misma tanto si se realiza de modo automático como intelectual. En ambos casos se trata de generar un conjunto de palabras que representan de qué trata un documento.

Imaginemos, por ejemplo, un documento, al que denominaremos Doc 1, con un texto como el siguiente:

**Cuadro 2.1. Documento Doc 1**

| |
|---|
| Una nutrición sana y el ejercicio habitual, en particular comer fruta y realizar actividades deportivas, bien sea en algún recinto o al aire libre, es muy importante tanto en la infancia como en la adolescencia. De este modo, además, se previenen una de las causas de retraso en el rendimiento escolar: la falta de salud y vigor físico. |

En la aproximación clásica, basada en la idea de las clasificaciones, el documento Doc 1 hubiera debido ser colocado en una categoría *a priori* de un cuadro de clasificación. En la operación de clasificar el documento, cualquier solución hubiera comportado, al mismo tiempo una creación y una destrucción de orden. Por ejemplo, si se hubiera colocado en “Alimentos”, el documento no aparecería por cualquiera de los otros temas para los cuales es relevante, a saber: “Infancia”, “Educación física”, “Rendimiento escolar”, etc. Tenemos aquí una muestra de las limitaciones propias de las clasificaciones que ya hemos discutido.

Sin embargo, con el método de indización, desaparecen tales limitaciones (aunque aparecen otras). En concreto, desaparece:

1. *La limitación de los puntos de acceso*. Mediante indización automática, por simple eliminación de las palabras más frecuentes, un algoritmo de ordenador derivaría los siguientes términos como candidatos para representar el contenido del documento (mostrados en orden alfabético): actividad, adolescencia, aire, comer, deportivo, ejercicio, escolar, físico, fruta, habitual, infancia, libre, nutrición, previenen, recinto, rendimiento, retraso, salud, sana, vigor.

   Mediante una clasificación es virtualmente imposible que podamos hacer lo mismo. En primer lugar, con toda probabilidad no dispondremos de todos los

términos o clases equivalentes en la clasificación. En segundo lugar, aunque dispusiéramos de tales entradas, por razones pragmáticas solamente se podrá asignar un pequeño número de categorías y necesariamente algunas serán demasiado genéricas.

2. *La limitación sintáctica.* El método de indización no requiere precoordinar los términos entre ellos, de modo que cualquier combinación de los 20 términos anteriores, dos a dos, por ejemplo: {infancia, nutrición}, {fruta, rendimiento}; tres a tres, por ejemplo: {infancia, nutrición, rendimiento}, etc., serían otros tantos puntos de acceso válidos. Igual pasaría con cualquier otra combinación cuatro a cuatro, etc. Es cierto que algunos lenguajes documentales de indización, como las listas de encabezamientos, contienen entradas precoordinadas, pero ello es debido, sobre todo, a su origen histórico. Las primeras listas de encabezamientos se utilizaron en sistemas manuales, de forma que no podían multiplicarse fácilmente las entradas. Esto indujo a preferir un sistema mixto en el cual se utilizaran entradas precoordinadas. Posteriormente, se comprobó que las listas de encabezamiento son una buena forma de proporcionar un sistema de exploración (o *browsing*) en sistemas informáticos.

3. *La limitación ontológica.* No hay un marco a priori que marque un límite o un modo de concebir los temas que pueden representarse mediante indización en un sistema documental. Si un aspecto de la realidad o un tema está presente en el documento, ese tema o ese aspecto de la realidad quedarán representados en el índice de la base de datos a través de las palabras correspondientes del autor del documento.

En total, el método de indización mediante términos que no están precoordinados entre sí proporciona hasta $n^2$ puntos teóricos de acceso al documento, siendo $n$ el número total de palabras o términos de indización asignados a cada documento (la razón es que cada una de las palabras sería un punto de acceso, pero cada combinación de palabras, dos a dos, tres a tres, etc. son otros tantos puntos teóricos de acceso. En total, la fórmula aritmética que nos dice cuántas combinaciones distintas de n términos son posibles es: $n^2$). En nuestro ejemplo, la fórmula anterior nos proporciona 400 maneras teóricas distintas de acceder al documento, contra las tres o cuatro formas de acceso que proporciona el sistema clásico basado en clasificaciones a priori. Es evidente que esas 400 formas de acceso incluyen combinaciones imposibles de prever por ninguna clasificación a priori.

Comparado con la clasificación, sin embargo, la indización también presenta limitaciones, en particular la indización automática del tipo que hemos simulado aquí. En primer lugar, los documentos carecen de un contexto que ayude a tomar decisiones a un usuario que aún no sabe exactamente lo que quiere. En segundo lugar, este tipo de indización no reconoce conceptos, sino cadenas de caracteres ante lo cual, aunque el documento anterior trata de educación, el índice generado de forma automática no incluye esa palabra. Estas limitaciones pueden paliarse en parte o totalmente, pero para ello se requieren otros métodos de indización y sistemas adicionales de visualización y representación de información que no están exentos de costes y, por tanto, no siempre son viables.

Por supuesto, otra forma de superar estas limitaciones es combinando la indización automática (como la que hemos visto o aún más sofisticada) con la indización intelectual. Sin embargo, en este apartado, examinaremos únicamente el procedimiento de indización automática. En síntesis, este tipo de indización automática se basa en estas tres ideas:

1. La representación de la información contenida en los documentos puede ser realizada de forma eficiente mediante conjuntos de palabras (términos de indización del documento), y no necesariamente por la asignación de cada documento a una clase o subclase predefinida de un cuadro de clasificación.

2. Las necesidades de información de los usuarios también se pueden representar mediante conjuntos de palabras (términos de indización de la pregunta).

3. Los documentos más relevantes son los que tienen los conjuntos de palabras más parecidos al conjunto de palabras de la necesidad de información.

### 2.5. Evaluación de sistemas de RI

Antes de entrar en consideraciones sobre la indización automática es necesario que dediquemos un tiempo a describir como se evalúa el rendimiento de los sistemas de RI. Las dos medidas más utilizadas acostumbran a ser el índice de exhaustividad (*recall*) y el índice de precisión (*precision*). Las fórmulas para calcular estos dos índices son las siguientes:

$$\text{Exhaustividad} = \frac{\text{Número de documentos relevantes recuperados}}{\text{Número total de documentos relevantes presentes en el fondo documental}} \times 100$$

$$\text{Precisión} = \frac{\text{Número de documentos relevantes recuperados}}{\text{Número total de documentos recuperados}} \times 100$$

*Ejemplo para el índice de exhaustividad*
Supongamos que en una colección hay 10 documentos relevantes sobre el tema *X*, y que, como consecuencia de una operación de recuperación de información sobre la materia *X* se obtienen tan sólo 6 documentos. Entonces, la fórmula anterior nos dice que el índice de exhaustividad de esa búsqueda ha sido del 60%.

*Ejemplo para el índice de precisión*
Supongamos que, en respuesta a una operación de recuperación de información, se han obtenido 10 documentos, pero que 5 de ellos no corresponden en realidad al tema solicitado, o sea, no son relevantes. Entonces, el índice de precisión para ese resultado ha sido del 50%.

Mientras que el índice de exhaustividad proporciona una medida de la habilidad del sistema para recuperar documentos relevantes, el índice de precisión proporciona una medida de la habilidad del sistema para evitar el ruido.

Naturalmente, el objetivo consiste en diseñar sistemas que proporcionen al mismo tiempo un 100% de exhaustividad y un 100% de precisión, es decir, sistemas que recuperen todos los documentos relevantes y tan sólo los documentos relevantes, pero, en la práctica, estos dos indicadores se comportan de manera antagónica, ya que las medidas para incrementar la exhaustividad tienden a disminuir la precisión y al revés.

La razón es la siguiente, si queremos asegurar la precisión del sistema adoptaremos medidas tendentes a aumentar la especificidad de la indización. Por ejemplo, si un documento trata sobre "gladiolos" entonces, diseñaremos un sistema de indización que tienda a indizar el documento con el descriptor "gladiolos", y no con el descriptor "flores" y mucho menos con el descriptor "plantas" o "jardines", etc. De esta manera tendremos un sistema muy preciso, aunque, sin duda, cuando alguien solicite documentos sobre "flores" dejará de recuperar documentos relevantes sobre el tema general "flores".

En general, podemos observar que algunos motores de búsqueda generalistas que funcionan en Internet, como Google o AltaVista, proporcionan buenas tasas de exhaustividad, es decir, tienden a recuperar muchos de los documentos relevantes del fondo (en este caso, el fondo es la Web) pero, como es fácil comprobar, el índice de precisión es bajo, ya que sólo una pequeña parte de los documentos recuperados son relevantes. Esta falta de precisión pasa desapercibida en algunas búsquedas debido a la calidad actual de los procedimientos de ordenación de estos motores. Por ejemplo, en Google, si limitamos el análisis a los diez o veinte primeros documentos recuperados, es posible que el índice de relevancia parezca muy alto. El problema, en este caso, es que puede haber otros documentos relevantes (incluso más que los primeros) en posiciones muy alejadas del principio y que nunca examinaremos por motivos prácticos.

En cambio, los sistemas muy especializados, como las agencias de selección y evaluación de recursos digitales tales como BUBL (www.bubl.ac.uk), ADAM (www.adam.ac.uk) o Cercador (www.cercador.com) que suelen combinar procedimientos de indización automáticos con otros de intelectuales, ofrecen mayor precisión, aunque a costa de la exhaustividad. A cada petición de información proporcionan menos recursos y, probablemente, por tanto, índices de exhaustividad más bajos, pero la tasa de precisión se aproxima al 100%.

También resulta útil, para discutir los problemas de evaluación de los sistemas de RI, utilizar los conceptos, adoptados de la teoría estadística, de los falsos positivos y de los falsos negativos. Un documento es un falso positivo cuando se recupera, pero no es relevante, es decir, se ha recuperado de facto, pero no tendría que haberse recuperado, ya que no es realmente relevante. Un documento es un falso negativo cuando, aunque es relevante, no se recupera. Es decir, no ha sido entregado al usuario a pesar de ser un documento relevante.

Los motivos de los rendimientos inadecuados en los índices de exhaustividad y de precisión, y por tanto, el fenómeno de los falsos positivos y de los falsos negativos son

diversos, pero se pueden señalar cuatro factores, los tres primeros propios de entornos donde se realiza una indización de tipo intelectual o mixta y el cuarto, de entornos de indización automática pura. Son los siguientes:

*a) Deficiente indización del documento*
Por ejemplo, el documento trataba del asunto *X* pero, en cambio, por error, no se ha asignado este descriptor. El documento no se recuperará cuando se solicite información sobre *X*. El caso contrario: un documento en realidad no trata del tema *Y*, pero le ha sido asignado el descriptor *Y*, por tanto, proporcionará ruido cuando alguien solicite información sobre *Y*.

*b) Deficiente indización de la necesidad de información*
La indización de las necesidades de información presenta el mismo problema. Tal vez el usuario desconoce que el tema por el cual está buscando información se representa con el descriptor *X*, por lo cual utiliza un término menos adecuado, por ejemplo, más general, esto le proporcionará un índice muy bajo tanto de precisión como de exhaustividad, etc.

*c) Grado insuficiente de especificidad del lenguaje documental*
El lenguaje documental utilizado en la representación de los documentos puede ser inadecuado. Por ejemplo, podrían existir diversos documentos en el fondo documental sobre "gladiolos", "rosas", "amapolas", etc., en cambio, el lenguaje documental tan sólo contempla el descriptor "flores", o peor aún, "plantas", con lo cual los documentos no quedan representados en su adecuado nivel de especificidad.

*d) Deficiente algoritmo de relevancia*
Cuando el sistema debe entregar muchos documentos como respuesta a la pregunta, entonces, el rendimiento final de la calidad del sistema vendrá determinado por el acierto en el cálculo de relevancia. En general, casi siempre que el sistema entregue varias decenas de documentos la relevancia adquirirá un factor esencial. La razón es que, en promedio, casi ningún usuario examina con atención más de allá de los veinte o treinta primeros documentos. Por ejemplo, supongamos que se han utilizado los términos *X*, *Y*, para indizar la pregunta, y supongamos que el cálculo de la relevancia otorga un gran peso, es decir, un valor positivo, a los documentos en los que aparecen muchas veces cualquiera de los dos términos, sin discriminar si solamente aparece uno o ambos términos. El documento más relevante para el usuario podría tener pocas ocurrencias de *X* y pocas ocurrencias de *Y*, por ejemplo, debido a la creatividad del autor que, tal vez, posee un rico vocabulario. Como resultado, el sistema podría desplazar el documento más relevante para el usuario a las últimas posiciones de la lista y privilegiar a documentos en los que tan sólo *X* (pero no *Y*), aparece muchas veces. Este, por ejemplo, es uno de los síndromes habituales de algunos motores de búsqueda de Internet, aunque cada vez sucede menos. En particular, tanto Google como AltaVista, por ejemplo, otorgan mayor valor a los documentos que poseen todos los términos de la pregunta. De hecho, suelen utilizar un primer filtro en el cual únicamente seleccionan los documentos que responden a un AND booleano con todos los términos de la pregunta (del estilo Término1 AND Término2).

## 2.6. Algoritmos básicos de RI

Como es sabido, los sistemas informáticos ni entienden ni pueden interpretar el significado de los textos y, a pesar de esto, los sistemas informáticos de RI desarrollan tareas que simulan inteligencia o, al menos, algún grado de comprensión del significado de la información textual. Esto es posible porque, en general, la capacidad de los ordenadores para resolver cualquier tarea o cualquier problema, desde lo más simple hasta lo más complejo, está basada en lo mismo: la determinación de un procedimiento que permita descomponer los pasos necesarios para la resolución de la tarea en un número finito de suboperaciones, cada una de las cuales no requiere inteligencia ni, por tanto, ninguna capacidad de comprensión o de interpretación de nada, ni de la información textual ni de la información de cualquier otra clase. A partir de aquí, la inteligencia aparente es un comportamiento que emerge de la totalidad del sistema.

Obviamente, donde sí hay inteligencia, y mucha, es en la persona o en el equipo de personas que han sabido descomponer la resolución de un problema en este número finito de pasos al que nos referimos y que, en matemáticas y en ciencias de la computación, tiene un nombre concreto, como es sabido: algoritmo (de aquí, por supuesto, la idea de una RI algorítmica).

Así pues, podemos definir un algoritmo como un método de resolución de problemas que consta de un número finito de pasos bien enunciados. En matemáticas, el procedimiento para resolver una suma, una raíz cuadrada o una división, son ejemplos de algoritmos. En informática, todo programa de ordenador consiste en uno o más algoritmos, codificados en un lenguaje de programación que pueda ser leído por un ordenador. Por tanto, antes que un programador pueda escribir un programa, hace falta que alguien, este mismo programador u otro, haya encontrado el algoritmo para resolver el problema que la aplicación informática tratará de solucionar.

En RI existen un buen número de algoritmos que se han ido descubriendo y refinando desde hace años. Estos algoritmos suelen presentarse bajo su forma lógica más abstracta, es decir, en forma independiente de su implementación en lenguajes de programación concretos, y así es como los presentaremos aquí también. Más adelante, examinaremos algoritmos para la indización automática de documentos y para el cálculo de relevancia. Ahora bien, el lector ha de entender que tal y como se presentarán estos algoritmos, no se podrían implementar en ningún ordenador, sino que, antes de esto haría falta traducirlos a alguno de los lenguajes de programación existentes, ya sea *C*, *Visual Basic*, *Java*, etc.

### 2.6.1. Indización automática

El objetivo de los procedimientos de indización automática es imitar lo mejor posible la capacidad de la indización intelectual (indización humana) de operar con conceptos, pero sin los errores e inconsistencias propios de la subjetividad humana y sin los altos costes económicos derivados de un trabajo que es, al mismo tiempo, intensivo en tiempo y muy especializado. Sin embargo, mientras la indización intelectual se caracteriza por permitir el trabajo con los conceptos, la indización automática trabaja, en principio, únicamente con cadenas de caracteres.

Para un indizador humano, las expresiones (1), "aumento de precios en un periodo determinado", (2) "índice de carestía", (3) "incremento periódico de precios" significan lo mismo, al menos desde el punto de vista de la indización documental y, por tanto, un indizador humano no tiene ningún problema para establecer una equivalencia entre los tres términos anteriores {(1), (2), (3)} y el término (4) "inflación". Por tanto, para un indizador humano, la relación entre los términos anteriores es una cadena de igualdades del tipo:
(1) = (2) = (3) = (4), en virtud de la cual, cualquiera de los términos -el (4), por ejemplo- puede ser declarado término preferente y, por tanto, descriptor autorizado para representar este concepto.

A partir de este momento, la aparición de las expresiones (1), (2), (3), u otras semánticamente equivalentes, en un documento, permite al indizador humano realizar la inferencia válida de que el documento tiene que indizarse con el descriptor (4) "inflación", aunque esta palabra "inflación" (es decir, esta cadena de caracteres, desde la lógica del ordenador) no aparezca en el documento.

En cambio, para un ordenador, lo que es significativo son las cadenas de caracteres, por tanto, la relación entre (1), (2), (3), (4) es la de una desigualdad simétrica entre todos ellos.

Partiremos de un documento-ejemplo sencillo, que llamaremos Doc2 y de un ejemplo de indización intelectual de este documento, para discutir el posible rendimiento de los diversos procedimientos de indización automática más habituales actualmente.

## 2.2 Cuadro 2. 2: Documento Doc2

### 2.3 La información como propiedad

La información no es una sustancia ni un objeto, sino una propiedad de los mensajes bien formados, a saber, la propiedad de dar a conocer algún aspecto de la realidad.

En este sentido, estamos de acuerdo con la teoría de la información de Alfred Dretske, según la cual, en realidad, una información falsa no es una información, en el mismo sentido que un pato de madera no es un pato.

Es por este motivo que podemos decir también que, en el contexto de la teoría de los símbolos, los mensajes son una clase de sistemas de información.

Estadísticas del documento
- Número total de palabras: 101
- Número total palabras distintas: 51 (términos únicos)

A partir de un hipotético documento como éste, una indización intelectual típica para representar el documento sería como la que recoge el cuadro 2.3.

**Cuadro 2.3: Descriptores asignados a Doc2 con indización intelectual**

1. Información
2. Mensajes
3. Teoría de la información
4. Semiótica
5. Sistemas de información
6. Alfred Dretske

Para un indizador humano, o al menos, para un indizador entrenado, es trivial identificar tanto los descriptores simples como los compuestos ("información" *versus* "sistemas de información"), así como asignar un descriptor por inferencia, y no por mera transcripción de palabras ("semiótica", como resultado de la expresión "teoría de los símbolos"); finalmente, el indizador humano no se deja engañar y no asigna el descriptor "patos", aunque el término aparece dos veces en el texto del documento.

En conclusión, un indizador humano (en el caso ideal), de manera rutinaria:
*a)* detecta tanto descriptores simples como compuestos;
*b)* asigna descriptores, aunque la palabra no esté presente en el documento;
*c)* no asigna descriptores, aunque la palabra esté presente en el documento.
En cambio, para un ordenador, conseguir *a), b)* y *c)* es una auténtica proeza. A pesar de todo, veremos más adelante como los ordenadores pueden aproximarse bastante a esto.

La indización que realizaría una máquina podría ser de tres tipos básicos, cada una de ellos según algoritmos sucesivamente más sofisticados, que veremos a continuación. En primer lugar, examinaremos un algoritmo que realiza una indexación simple, y que queda representado en el siguiente cuadro:

**Cuadro 2.4. Algoritmo 1: modelo de indización simple**

1. Identificar las cadenas de caracteres del documento.
2. Agrupar las cadenas únicas.
3. Considerar cada una de las cadenas únicas del documento como un término de indización del documento.

Cabe aclarar que cada una de las palabras diferentes de un documento o de una base de datos recibe el nombre de palabras únicas o términos únicos. En este caso, hablamos de cadenas de caracteres únicas.

El algoritmo precedente es de una gran simplicidad conceptual, pero su implementación no es tan simple como puede parecer. En primer lugar, hemos obviado algunas cuestiones, rutinarias en programación, como son prever como se iniciará y cómo finalizará el proceso, indicar cuál será la entrada de la información y cuáles serán las salidas, etc.

En segundo lugar, habrá que especificar en el programa informático qué se considerará una cadena de caracteres y lo que no se considerará una cadena de caracteres. Por ejemplo:

a) ¿La expresión "sistema de información" es una, son dos o son tres cadenas de caracteres?
b) ¿Los espacios en blanco y los signos de puntuación son siempre separadores de cadenas de caracteres? Por ejemplo, el punto (.), la barra (/), el guión (-), ¿son siempre separadores de cadenas de caracteres? Si es así, expresiones como "E.U." serán dos cadenas de caracteres; y ¿qué pasará con fechas expresadas como en "01-10-2004", o con expresiones como "importación/exportación"? etc.
c) Habrá que especificar qué es una cadena única de caracteres. En el caso más simple son cadenas o términos únicos las cadenas idénticas. "Información", por ejemplo, aparece diversas veces en el texto; se trata de una misma cadena y, por tanto, es un término único, pero, ¿qué pasaría con "información" e "informaciones"? ¿son uno o dos términos únicos?

Por tanto, aunque no sea evidente a primera vista, incluso un algoritmo conceptualmente tan simple como el Algoritmo 1 requiere de un cierto análisis, debido a que, como ya hemos indicado antes, se trata de que una máquina que no puede interpretar las palabras sea capaz, en cambio, de identificarlas en un texto en base a instrucciones simples.

En cualquier caso, la indización que produciría un algoritmo simple de indización coincidiría con el resultado del cuadro 2.7, es decir, los términos de indización asignados coincidirían con la lista de palabras únicas del documento, tal como recoge el siguiente cuadro.

**Cuadro 2.5: Resultado de la indización de Doc 2 con un algoritmo simple (términos únicos del documento)**

| | | |
|---|---|---|
| a<br>acuerdo<br>Alfred<br>algún<br>aspecto<br>bien<br>clase<br>2.4 como<br>con<br>conocer<br>contexto<br>cual<br>dar<br>de<br>decir<br>Dretske<br>el<br>en | es<br>estamos<br>este<br>falsa<br>formados<br>información<br>la<br>los<br>madera<br>mensajes<br>mismo<br>motivo<br>ni<br>no<br>objeto<br>pato<br>podemos | por<br>propiedad<br>que<br>realidad<br>saber<br>según<br>sentido<br>símbolos<br>sino<br>sistemas<br>son<br>sustancia<br>también<br>teoría<br>un<br>una |

A continuación, vamos a realizar cuatro comentarios sobre esta clase de indización.

En primer lugar, se ha multiplicado el número de términos de indización asignados al documento. Hemos pasado de los 7 términos de la indización intelectual, a 51 con indización automática simple.

En segundo lugar, y como consecuencia directa del anterior, este documento tendrá muchas más posibilidades de ser recuperado, pero en muchas de estas posibilidades, este documento será un falso positivo, es decir, proporcionará ruido. El caso más evidente, será si alguna vez este documento es recuperado a partir de una pregunta sobre patos.

En tercer lugar, y en contraste con el anterior, este documento será un falso negativo cada vez que algún usuario solicite documentos sobre "semiótica", ya que este término no aparece en el texto y, por tanto, el sistema automático de indización no ha podido identificar este concepto.

En cuarto lugar, debido al algoritmo utilizado, se ha perdido mucha información, ya que este algoritmo tan sólo es capaz de identificar palabras simples, como "información", pero no cadenas como "sistema de información" o como "Alfred Dretske".

Aunque, como decíamos, este algoritmo parezca muy simple e, incluso, dé resultados muy limitados, es uno de los más utilizados todavía actualmente. Es el que usan algunos motores de búsqueda en la Web, así como el que aún está presente en buena parte de los sistemas de gestión documental de las empresas.

También hay que señalar que, a menudo, este algoritmo de indización automática se complementa con una indización intelectual, con lo que el resultado final es, en realidad, una combinación de los términos de indización de los cuadros 2.3 y 2.5. A pesar de todo, esta no es la práctica mayoritaria en las empresas, sino más bien en el seno de centros de documentación y bibliotecas. Por tanto, en muchas empresas, el rendimiento máximo de sus sistemas de RI es el que ofrece el algoritmo que hemos discutido aquí.

Un tipo de programa que utilizan este algoritmo son los sistemas de gestión de bases de datos *FileMaker* (www.filemaker.com), *Idealist* (www.bekon.com), o Knosys (www.micronet.es) (véase apartado 3.2.6), muy populares como solución departamental, también en pequeñas y medianas empresas y en algunos centros de documentación. En todos ellos, además, se pueden filtrar las palabras consideradas vacías (como los artículos y preposiciones) de modo que el sistema de indización las descarte de entrada como candidatas a términos de indización. En el caso de programas de gestión documental más avanzados, como *Inmagic DB/Text* (www.inmagic.com) o Winisis (www.unesco.org/), es posible configurar el programa para que sea capaz de identificar cadenas compuestas como "Alfred Dretske" o "sistema de Información". Estos pasos, los veremos en las siguientes versiones del algoritmo.

El algoritmo 2, que discutiremos a continuación, presenta una importante mejora en relación al anterior, y en el cuadro siguiente indicamos sus características (seguimos, sobretodo, el modelo de Gerard Salton).

**Cuadro 2.6. Algoritmo 2: modelo de indización avanzada**

1. Identificación de las cadenas de caracteres, para determinar la primera lista de candidatos a términos de indización.
2. Eliminación de las palabras vacías de esta lista, es decir, de los términos muy frecuentes.
3. Creación de raíces con las cadenas de caracteres.
4. Combinación de términos sinónimos.
5. Cálculo de frecuencias absolutas.
6. Cálculo del peso o importancia de los términos en cada documento.
7. Eliminación, como candidatos a descriptores, de los términos con un índice de discriminación que quede por debajo de un umbral determinado.
8. Asignación de los descriptores ponderados a cada documento.

En este algoritmo, el primer paso es idéntico al anterior y los problemas a resolver en su implementación son exactamente los mismos, a saber, habrá que especificar algún procedimiento eficiente para determinar de manera correcta qué es y qué no es una cadena de caracteres válida, etc. En el segundo paso, en cambio, ya encontramos la operación nueva de la eliminación de las denominadas palabras vacías (*stopwords*) por un método automático.

Las palabras vacías son palabras con una frecuencia tan alta que en teoría no tienen ninguna capacidad para discriminar documentos y, por tanto, es mejor retirarlas de entrada de la lista de candidatos a descriptores. Determinar qué son las palabras vacías en cada caso se puede hacer de dos formas diferentes: a priori, a posteriori y, cómo no, con una combinación de los dos métodos.

En el método a priori, un operador humano introduce en el sistema una lista, denominada a veces diccionario de palabras vacías, que contiene todas aquellas partes de una lengua que tienen una función gramatical, pero un pobre significado semántico independiente, por ejemplo, pronombres, artículos, adverbios, etc. Para muchas lenguas, incluyendo, el castellano, el inglés o el catalán, acostumbran a salir al menos unas 300 palabras de este tipo.

Con el método a posteriori, las palabras vacías se determinan por cálculo de frecuencia. De esta manera, se retiran de la lista de candidatos todas aquellas palabras que aparecen, por ejemplo, en más del 80% de los documentos. De esta manera se detectan palabras vacías que, de otra forma pasan desapercibidas. Por ejemplo, en un fondo documental sobre economía, el término "economía" probablemente convendrá considerarlo una palabra vacía.

Según Salton, de esta manera la lista inicial de términos candidatos queda reducida típicamente en un 40% o un 50%. En nuestro caso, de 51 palabras pasamos a 30, es decir, efectivamente se ha producido una reducción de un poco más del 40%, como podemos ver en el siguiente cuadro.

**Cuadro 2.7. Primer grupo de candidatos a descriptores: resultado de la eliminación de las palabras vacías de la lista inicial de Doc2**

| | | |
|---|---|---|
| acuerdo | estamos | podemos |
| Alfred | falsa | propiedad |
| aspecto | formados | realidad |
| bien | información | saber |
| clase | madera | sentido |
| conocer | mensajes | símbolos |
| contexto | mismo | sistemas |
| dar | motivo | sustancia |
| decir | objeto | también |
| Dretske | pato | teoría |

El tercer paso consiste en fusionar los términos que tienen las mismas raíces. De esta manera si, por ejemplo, en el documento hubiera palabras como "información" e "informaciones", quedarían reducidas a una sola forma: "informacion*" (donde el asterisco indica un truncamiento).

El cuarto paso consiste en detectar posibles sinónimos. Por ejemplo, si en el documento tuviéramos dos palabras como "ordenador" y "computadora", en este paso quedarían fusionadas en una única palabra a efectos del cálculo de frecuencia del que hablaremos seguidamente. Es decir, se consideraría que, en vez de dos palabras, habría un mismo término con dos ocurrencias. Este paso se debería resolver con el uso de un tesauro o con una lista previa de sinónimos. En la práctica, muchos de los sistemas de indización automáticos actuales obvian este paso dadas sus dificultades de realización práctica.

En el quinto paso, se realiza el cálculo de las frecuencias absolutas de cada uno de los términos de la lista resultante. Este es un paso previo al cálculo del peso o índice discriminatorio de cada término.

Según este índice, los diversos términos de un documento pueden tener una capacidad discriminatoria diferente, que indica la posible utilidad de cada término como descriptor. Un término es tanto mejor descriptor cuanto mejor sirve para discriminar grupos de documentos. Por ejemplo, un término como "sistema" probablemente es un mal descriptor en casi cualquier contexto, ya que debe estar presente en un gran número de documentos y, por tanto, tiene un índice de discriminación muy bajo. En cambio, probablemente, el término "teoría de sistemas" tiene un índice de discriminación más alto.

En el sexto paso, se calcula, por tanto, el índice de discriminación o peso de cada término de la lista de descriptores. El cálculo que propone Salton, y que siguen bastantes sistemas de indización automática, es el siguiente:

FT x FID = índice de discriminación del término

FT = Frecuencia absoluta del término en el documento
FID = Frecuencia inversa del documento

La frecuencia absoluta (FT) es el número de veces que aparece el término en el documento. Por ejemplo, en nuestro caso, la lista de frecuencias absolutas es la siguiente:

**Cuadro 2.8. Frecuencias absolutas de los términos candidatos a descriptores de Doc2**

| | | |
|---|---|---|
| acuerdo 1 | estamos 1 | podemos 1 |
| Alfred 1 | falsa 1 | propiedad 3 |
| aspecto 1 | formados 1 | realidad 2 |
| bien 1 | información 6 | saber 1 |
| clase 1 | madera 1 | sentido 2 |
| conocer 1 | mensajes 2 | símbolos 1 |
| contexto 1 | mismo 1 | sistemas 1 |
| dar 1 | motivo 1 | sustancia 1 |
| decir 1 | objeto 1 | también 1 |
| Dretske 1 | pato 2 | teoría 2 |

Tan sólo con una rápida mirada a esta lista, ya se puede ver que los términos más frecuentes corresponden bastante bien al tema del documento y, por tanto, si adoptásemos como descriptores todos los términos de frecuencia superior a 1, por ejemplo, no nos quedaría una mala representación del documento como se puede ver en el cuadro siguiente, con la salvedad del candidato a descriptor "pato" que no sería un buen descriptor para este documento.

**Cuadro 2.9. Descriptores con frecuencia de aparición superior a 1**

| |
|---|
| información 6 |
| propiedad 3 |
| pato 2 |
| mensajes 2 |
| realidad 2 |
| sentido 2 |
| teoría 2 |

Ahora bien, el sexto paso no se limita a adoptar la frecuencia absoluta como indicador de la bondad de un término como descriptor, sino que, como hemos visto por la fórmula anterior, relaciona esta frecuencia con la denominada “Frecuencia inversa del documento” (FID). Esta se calcula así:

$$FID_j = \frac{\text{número total de documentos en el fondo documental}}{\text{número total de documentos que contienen el término } j}$$

donde, FID*j* significa que la frecuencia inversa del documento para el término *j* (por ejemplo, "economía") se obtiene dividiendo el número total de documentos de la base de datos, por el número de documentos que tienen el término *j*.

La FID de un término sirve para indicar su peso relativo, ya que relaciona su frecuencia en todo el fondo documental con el número total de documentos. Multiplicando el factor FID de cada término (que es una medida global) con la frecuencia absoluta (FT) en el documento (que es una medida local) se pretende lo siguiente: otorgar más peso a los términos que tienen una alta presencia local y una baja presencia global. Por ejemplo, si el término "información" tiene una presencia muy alta en el documento, pero también tiene una frecuencia muy alta en todo el fondo documental, podría obtener un peso relativo más bajo que el término "propiedad", el término "mensajes", el término "Dretske" o (en este caso, por desgracia) el término "pato".

En el paso número 7, los candidatos a descriptor con un índice de discriminación por debajo de un determinado umbral, quedarían eliminados. Este índice tiene que establecerse de manera empírica según las características de cada fondo. Podemos suponer que, de la lista de los 29 descriptores, probablemente, una tercera parte de ellos quedarían excluidos como candidatos a descriptores.

A partir de aquí (paso nº 8) es imposible saber de modo anticipado como quedaría esta lista, ya que el cálculo dependerá en cada momento de las características concretas del fondo del que formase parte, pero, podemos especular con que, en un momento determinado, podría parecerse a algo como esto:

**Cuadro 2.10. Lista (hipotética) de descriptores de Doc2, con el algoritmo 2**

| |
|---|
| información |
| propiedad |
| pato |
| mensajes |
| realidad |
| sentido |
| teoría |

Finalmente, además, cada descriptor quedaría asignado al documento con un índice numérico de su peso o capacidad discriminatoria como tal y esto se podría utilizar después en el cálculo de la relevancia del documento. Este índice, resultado del cálculo del paso nº 6, podría ser un número entre 0 y 1, de manera que, por ejemplo, el descriptor "información" podría tener un índice de 0,4 mientras que el descriptor "mensaje" podría tener un índice de 0,5, etc.

Se trata, por tanto, de un resultado bastante mejor que el que daba el modelo simple de indización automática, pero no es mejor aún que la indización intelectual (suponiendo, por otro lado, un indizador humano ideal).

Persisten problemas similares: este procedimiento no reconoce unidades superiores a la palabra (no reconoce "teoría de la información") y, probablemente, el término "pato" se asignaría como descriptor a este documento que, por supuesto, no trata en absoluto de patos.

Numerosos motores de búsqueda de Internet parecen aplicar un algoritmo como éste, o muy parecido, en su procedimiento de análisis e indización automática, aunque nunca es posible estar del todo seguros desde el momento que las empresas que administran estos motores no proporcionen los detalles exactos de sus algoritmos.

Ahora bien, existe la posibilidad de añadir aún algunos pasos más en el algoritmo 2 que estamos examinando ahora y que aún podría mejorar el resultado. En concreto, en algunas ocasiones, Salton y otros autores han presentado un modelo de indización automática que incorpora los pasos señalado aquí como *5a* y *6a* y que destacamos en cursiva):

**Cuadro 2.11. Algoritmo 2a: modelo de indización avanzada con variaciones**

1. Identificación de las cadenas de caracteres para determinar la primera lista de candidatos a términos de indización.
2. Eliminación de las palabras vacías de esta lista, es decir, de los términos muy frecuentes.
3. Creación de raíces con las cadenas de caracteres para crear los términos de indización.
4. Combinación de términos sinónimos.
5. Cálculo de frecuencias absolutas.

*5a. Eliminación de términos muy poco frecuentes en la colección.*

6. Cálculo del peso o importancia de los términos en cada documento.

*6a. Formación de frases (descriptores compuestos) con términos muy frecuentes*

7. Eliminación, como candidatos a descriptores, de los términos con un índice de discriminación que quede por debajo de un umbral determinado.
8. Asignación de los descriptores ponderados a cada documento.

Se supone que, gracias al paso *5a*, se eliminarían de los candidatos a descriptores un término como "patos". Ahora bien, esto sería cierto siempre que nos moviéramos en un fondo documental especializado y en el cual, por tanto, términos ajenos a la especialidad del fondo no aparecieran con frecuencia. Si suponemos que estamos hablando de un fondo especializado en información y comunicación, entonces es plausible suponer que el término "pato" sería muy infrecuente y quedaría, por tanto, eliminado. Ahora bien, esto tan sólo es una hipótesis que, en todo caso, en un fondo indiscriminado como el que existe en la Web no funcionaría bien.

Por otro lado, gracias al paso *6a*, se supone que, también en condiciones ideales, saldrían descriptores compuestos como "sistemas de información". Ahora bien, igual que en el caso anterior, esto tan sólo es una hipótesis que, a veces se cumple, según las características del fondo, y otras veces no, y en todo caso no siempre se cumple al 100%.

Sea como sea, en el caso más favorable, ahora el resultado que tendríamos, si aplicásemos el algoritmo 2a, podría ser el siguiente:

**Cuadro 2.11 Lista hipotética de descriptores de Doc2, con el algoritmo 2a**

| |
|---|
| información<br>propiedad<br>mensajes<br>realidad<br>sistemas de información<br>teoría de la información |

Los comentarios que podemos hacer a este resultado son los siguientes: en primer lugar, se aprecia una mejora en el sentido que se han eliminado algunos términos inadecuados, como el famoso "pato" (pero, recordemos que esto tan sólo es una hipótesis). En segundo lugar, se han añadido dos términos compuestos, como "sistemas de información" y "teoría de la información" que, sin duda, mejoran la indización. Ahora bien, por los mismos principios según los cuales han desaparecido algunos descriptores inadecuados, también podrían desaparecer los descriptores "Alfred" y "Dretske". Finalmente, no es plausible, al menos sin el concurso de un tesauro externo, que el descriptor "semiótica" quedase asignado al documento.

Para que la indización automática consiga un mejor rendimiento, faltaría incluir, al procedimiento avanzado, algunas operaciones y perfeccionamientos que pudiesen conducir a una indización no ya avanzada, sino inteligente.

Ahora bien, todo lo que se dirá a partir de ahora existe tan sólo o bien en sistemas propietarios que, por alguna razón, no han llegado al mercado como soluciones estandarizadas, o bien en productos de tipo experimental.

La mejora de los procedimientos de análisis e indización documental parece que tendría que provenir de combinar diversos instrumentos en este tipo de procesos:

- Instrumentos de análisis lingüístico
- Sistemas expertos
- Tesauros

Los instrumentos de análisis lingüístico permitirían detectar candidatos a descriptores con más fundamento que los simples datos estadísticos de los términos, aunque éstos continuarían siendo útiles. Por ejemplo, con técnicas de lingüística computacional y terminología, se podrían detectar candidatos a descriptores formados no tan sólo por palabras simples, como "información", sino también por palabras compuestas, como "sistemas de información", a partir de la determinación de las características sintácticas, semánticas y morfológicas de los textos y de reglas de formación de expresiones gramaticalmente válidas, y no tan sólo en base a propiedades estadísticas de los textos.

Por su parte, un sistema experto podría aplicar reglas de producción, del estilo "si... entonces...", para asignar descriptores de un tesauro o identificar sinónimos con la

ayuda también de un tesauro. Por ejemplo, una regla de producción del sistema experto podría servir para deducir que:

*si* <el término "diafragma" aparece en un contexto próximo al término "óptica">, *entonces*, <el documento se puede indizar con el término "diafragmas ópticos">.

En caso necesario, el uso de un tesauro como parte integrante del sistema experto ayudaría a formar clases de sinonimia y a escoger, en cada caso, el término preferido como descriptor, así como ayudaría a escoger el término más adecuado según el nivel de especificidad, etc.

O bien, podría aplicar reglas que determinasen que "Alfred Dretske" es un nombre propio que identifica a un autor y que este autor es suficientemente relevante para ser utilizado como descriptor. Por ejemplo, una regla según la cual:

*si* <dos cadenas conexas comienzan con mayúscula> *y si* <van precedidas de la expresión "según">, *entonces*, <se trata de un nombre propio y el documento se puede indizar con este nombre propio>.

### 2.7. La indización y la recuperación de información en la Web

Hasta ahora, hemos presentado diversos aspectos de la RI relacionados con entornos homogéneos, como los que se dan en el seno de una misma empresa, centro de documentación o biblioteca. Sin embargo, la Web presenta unas características propias, desde el punto de la RI, que deben ser consideradas también.

En primer lugar, la libertad de publicación en la Web hace que su contenido sea totalmente *heterogéneo* en todos los sentidos de la palabra, es decir, no solamente en las temáticas y géneros tratados, sino también en sus enormes variaciones de calidad. En segundo lugar, su enorme tamaño --se calcula que la Web visible contenía más de cinco millones de documentos a principios del 2004—hacen de la Web el medio unificado de información más grande del que ha dispuesto jamás la humanidad.

En tercer lugar, en la Web los documentos forman parte de una red de alcance mundial (como indica bien el término). Esta red consiste en enlaces hipertextuales que relacionan o bien documentos distintos entre sí, o bien secciones distintas del mismo documento (y frecuentemente ambas cosas a la vez). Este sistema de enlaces otorga a la web la característica de la navegación como forma de acceder a la información.

La heterogeneidad en temas y sobre todo en calidad de los contenidos, hacen que las fórmulas clásicas de la RI que hemos considerado, y que se desenvuelven con aceptable eficacia en entornos homogéneos, sean menos eficientes a la hora de “indizar” la Web. De hecho, se han desarrollado tres estrategias básicas para este fin: los directorios, los motores de búsqueda y las bases de datos.

Por tanto, para considerar aspectos de búsqueda de información en la web, hemos de comenzar por destacar la diferencia esencial que existe entre los directorios y los motores de búsqueda, ya que los directorios utilizan la *navegación* como acceso a la información mientras que los motores utilizan la *interrogación*.

Mientras la *navegación* es una forma de acceso a la información que consiste en realizar elecciones sucesivas a partir de un cuadro de clasificación o de una estructura de categorías previamente dada, la *interrogación* consiste en expresar una necesidad de información mediante uno o más términos. De hecho, hemos presentado la tabla siguiente de manera que esa diferencia quede enfatizada.

Las diferencias entre los directorios y los motores de búsqueda son tan grandes que sorprende que incluso la bibliografía técnica los trate a veces como si fueran equivalentes. En consecuencia, algunos usuarios de Internet tienden a creer que no hay una diferencia esencial entre hacer una búsqueda en un directorio como Yahoo, mediante navegación, o hacerla en un buscador como Google, mediante interrogación.

En primer lugar, un directorio consiste en una estructura jerárquica formada por clases y subclases. Cada una de estas clases contiene a un número determinado de recursos. El acceso a la información se realiza por navegación o desplazamientos sucesivos entre las clases y los niveles de la jerarquía.

Los recursos se seleccionan, analizan y clasifican de forma intelectual y, por este motivo, solo contienen a una parte pequeña de la Red. Los recursos clasificados por los directorios se cuentan por centenares de miles, cuando los recursos o documentos publicados en Internet se cuentan por miles de millones.

### 2.7.1. Motores de búsqueda

Los motores de búsqueda, por su parte, proporcionan la consulta de índices analíticos como los que hemos discutido más arriba. Estos índices representan el contenido de los sitios y páginas web publicados en Internet.

De este modo, como ya sabemos, cada recurso o documento se representa mediante un conjunto de palabras o frases, llamados términos de indización porque forman parte del mencionado índice.

Cuando un usuario busca una determinada clase de recursos, expresa su necesidad de información utilizando palabras con la esperanza de que estén presentes únicamente en los documentos relevantes. El motor de búsqueda, o mejor dicho, una parte especializada del sistema, compara entonces los términos de la pregunta con los que figuran en el índice y selecciona de este modo todos los documentos o recursos que coinciden, totalmente o en parte con dicha expresión de búsqueda.

La hipótesis que subyace en este procedimiento es doble. En el lado del motor de búsqueda, la hipótesis consiste en que el motor es capaz de extraer adecuadamente los términos que representan el contenido de los documentos. En el lado del usuario que plantea la pregunta, la hipótesis consiste en que los términos que utiliza estarán presentes en los documentos relevantes y no lo estarán en los documentos no relevantes.

Parecen hipótesis modestas, pero los hechos demuestran que solamente se cumplen parcialmente. En primer lugar, los motores de búsqueda son capaces de identificar cadenas de caracteres, pero no conceptos, con lo cual toda la vaguedad, ambigüedad,

etc., del lenguaje natural se traspasa a los índices. Además, los motores de búsqueda no pueden distinguir ni el género ni la calidad de los documentos.

Los índices de los motores de búsqueda, por tanto, se construyen detectando todas y cada una de las cadenas de caracteres que forman parte de los documentos. En algunos casos, todas las cadenas van a parar al índice y se convierten así en puntos de acceso al documento.

#### 2.7.1.1. Cálculo de relevancia en la Web

El caso más característico de la RI aplicada a la Web lo tenemos en la forma en que los motores de búsqueda calculan la relevancia para entregar sus resultados. En los primeros años de la Web, justo hasta la aparición de Google, este cálculo se realizaba con los algoritmos que hemos indicado más arriba. De hecho, la mayor parte de los motores los siguen usando hoy en día.

Sin embargo, Google aportó a finales de los noventa una importante novedad: además de los criterios intrínsecos de cada sitio o página web (p.e. la frecuencia absoluta y relativa de los términos de búsqueda), consideró el siguiente criterio externo a la web: el número y la calidad de los enlaces que recibe la página o el sitio web considerado.

A medida, denominada *PageRank*, consiste, consiste, en términos conceptuales en lo siguiente: para Google, un documento (esto es, una página web) es tanto más relevante (a igualdad de otros factores) cuantos más enlaces recibe de otras páginas web que, a su vez, reciban muchos enlaces.

Con variaciones, la mayor parte de los motores de búsqueda han incluido el número de enlaces recibidos (*popularidad*, en argot técnico) en una forma de ponderar la relevancia de un documento en la Web. En síntesis, del análisis de los principales motores de búsqueda (Google, AlltheWeb, AltaVista) se deduce que el cálculo de relevancia de un sitio o de una página web se obtiene combinando de alguna forma los factores que se indican en las dos tablas siguientes (el orden no es necesariamente significativo):

**Tabla 2.2: Criterios internos a la página web**

| ***Para una sola palabra clave*** | ***Para dos o más palabras clave*** |
|---|---|
| **1. Frecuencia absoluta**<br>Número de veces (sumatorio) que aparece el término de búsqueda. Cuanto mayor es la frecuencia, más relevante es la página. | **1. Frecuencia absoluta**<br>(Ver explicación columna izquierda) |
| **2. Ubicación**<br>Lugar donde aparece el término de búsqueda. Una página web donde el término aparezca en el título es más importante que si aparece solamente en el cuerpo. | **2. Variedad**<br>Número de términos de la pregunta presentes en el documento. |

| | |
|---|---|
| **3. Emergencia**<br>Número de orden de la palabra. Si el término aparece al inicio del título es más importante que si aparece al final. | **3. Ubicación**<br>(Ver explicación columna izquierda) |
| **4. Frecuencia relativa**<br>La frecuencia absoluta dividida por el número de palabras de la página. Cuanto mayor es la frecuencia relativa (o densidad) mayor es la relevancia, siempre que esta frecuencia relativa se mantenga en unos márgenes estadísticos. Por ejemplo, los motores de búsqueda pueden penalizar frecuencias relativas muy altas. | **4. Proximidad**<br>Número de palabras entre los términos de búsqueda. En general, cuantas menos palabras separen a los términos de búsqueda en el documento, mayor es su relevancia. |
| | **5. Emergencia**<br>(Ver explicación columna izquierda) |
| | **6. Frecuencia relativa**<br>(Ver explicación columna izquierda) |

**Tabla 2.3: Criterios externos a la página web**

| |
|---|
| ***Criterios basados en enlaces*** |
| **1. Número de enlaces recibidos por la página**<br>Un sitio web será más relevante (a igualdad de otros factores) cuantos más enlaces recibe de otras páginas web. |
| **2. Calidad de los enlaces recibidos por la página**<br>No todos los enlaces otorgan el mismo valor para calcular la relevancia. Las páginas que a su vez son muy enlazadas, otorgan más valor que las páginas poco enlazadas. Dicho de otro modo: el enlace procedente de una página personal otorga menos valor que el enlace procedente de Yahoo, por ejemplo. |
| **3. Texto de los enlaces recibidos por la página**<br>Algunos buscadores, notablemente, Google, consideran el texto que sirve de anclaje al enlace externo hacia otra página como una pista o una inferencia que utilizan para calcular la relevancia de la página así enlazada. En casos extremos, si muchas páginas web contienen un enlace con el texto "*x*" (p.e. el texto "biología") hacia un mismo sitio web, es posible que ese sitio web sea muy relevante para Google, incluso aunque el sitio web considerado no contenga el término "*x*" (o sea, aunque no contenga el término "biología"). |
| ***Criterios basados en tráfico*** |
| **1. Número de visitas que recibe la página**<br>Esta medida se refiere al número de visitantes que tiene un sitio web, y la ha desarrollado, entre otras empresas, *Alexa* (www.alexa.com) bajo la denominación *Traffic Rank* (ver más adelante). |
| **2. Número de páginas visitadas**<br>Además del número de visitas, se suele considerar también el número de páginas vistas en un sitio. Forma parte también del indicador *Traffic Rank* (ver más adelante). |

### 2.7.2. Interfaces de búsqueda en motores

Los motores de búsqueda en la Web representaron un claro paso adelante en relación a un aspecto de la RI (el cálculo de relevancia), pero al mismo tiempo representaron un claro paso atrás en otro aspecto: en relación a interfases de consulta.

Actualmente las interfaces de consulta de los motores están claramente por debajo de las posibilidades de las bases de datos que podríamos denominar *clásicas* (pre-web) y de los sistemas de gestión de bases de datos que son de uso común en empresas.

La mayor parte de los motores de búsqueda ofrecen una búsqueda simple, que consiste en una caja donde se pueden entrar términos sin ninguna relación explícita entre ellos (sin ninguna sintaxis ni ningún tipo de operador), y una búsqueda avanzada, donde se puede utilizar, en general de forma limitada, algunos operadores booleanos combinando con alguna forma también muy limitada de búsqueda por campos.

Carecen, en general, de opciones para ayudar en la búsqueda que en cambio están presentes en cualquier sistema de gestión de bases de datos: como guardar consultas, combinar resultados de búsquedas anteriores, consultar índices, etc.

Las actuales interfaces de consulta de los motores de búsqueda están orientadas a promover en el usuario la idea (falsa) de que la búsqueda de información en la web es simple. Buscar en Internet, igual que buscar en otros espacios mucho más reducidos y homogéneos, puede resultar muy simple o muy complicado en función de la necesidad de información. Es evidente que si deseamos encontrar sitios web con información turística sobre Londres (o sobre cualquier ciudad medianamente importante), la operación no reviste ninguna dificultad. Pero, si lo que necesitamos es encontrar información para desarrollar un proyecto que combine el tema *X*, desde el punto de vista *Z*, pero teniendo en cuenta el aspecto *M*, entonces la cuestión no es tan simple (por ejemplo: “web semántica y Dublin Core, desde el punto de vista de las bibliotecas digitales”)

Cualquier lector acostumbrado a realizar este tipo de búsqueda sabe que, en esos casos, los motores de búsqueda deberían proporcionar interfases de consulta preparadas para que los usuarios desarrollaran una búsqueda que podría requerir de varias sesiones de trabajo, así como de varios procesos de ensayo y error.

Hay un aspecto, sin embargo, donde los motores de búsqueda han aportado una mejora en la interfase de búsqueda. Casi todos han incorporado algún sistema automático de agrupación de resultados por temas (categorización) o de asociación de términos relacionados que, o bien ayudan al usuario a transformar su búsqueda en una operación parcial de navegación o bien le proporcionan nuevos términos de búsqueda para refinar su consulta. Dado que estas operaciones de categorización o de sugerencia de nuevos términos de búsqueda están soportadas por algoritmos automáticos que, a su vez, tienen una escasa base lingüística y un fuerte componente estadístico, su utilidad es muy aleatoria. En algunas consultas serán de gran utilidad, pero en otras será decepcionante. Probablemente, darán un buen rendimiento para búsquedas simples basadas en términos muy generales, pero muy mal rendimiento en búsquedas especializadas.

Por ejemplo, una búsqueda por el término Londres en AltaVista (www.av.com) proporciona, además de la lista de resultados, esta lista de términos relacionados con el objetivo de ayuda a precisar el sentido de la consulta:

Hoteles De Londres
Petit Palace Londres
Bed Breakfast
Gran Bretaña
London Hotel
London Stoc Exchange
National Gallery
Post Production
Reino Unido
River Thames

Una simple revisión de la misma sirve para convencernos de la posible utilidad de disponer de una sugerencia de búsqueda como “Hoteles” o como “Gran Bretaña”. Sin embargo, una búsqueda por un tema más especializado, como “Arquitectura de la Información” o “Recuperación de Información”, las sugerencias no siempre son tan afortunadas.

### 2.7.3. Posicionamiento Web

En cualquier caso, la utilidad de los motores de búsqueda es indiscutible (e insustituible). Reiteradas encuestas y diversos estudios del tráfico en la Web muestran que una gran parte de los internautas acceden a los sitios web a través de motores de búsqueda (y a través de Google concretamente en su inmensa mayoría). Es decir, cada vez más, los internautas realizan una búsqueda previa en un motor como Google para elegir el sitio al que visitar.

Por ejemplo, si alguien planea adquirir un automóvil a través de Internet o simplemente desea obtener información sobre el mercado de automóviles, hay una probabilidad muy alta de que acceda a Google o un motor similar (AlltheWeb, AltaVista, etc.) y: (1) escriba una expresión como “venta de automóviles”, (2) examine los resultados y a partir de ellos (3) haga clic en uno o más de los sitios ofrecidos por el motor como respuesta.

Una secuencia similar tendrá lugar en un elevado porcentaje de ocasiones si se trata de un investigador o un estudiante que prepara un trabajo académico o si se trata del ejecutivo de una empresa que está preparando un estudio de mercado.

Los datos relativos al inmenso número de consultas que procesan diariamente los motores de búsqueda dan buena fe de lo anterior. La cuestión es que, por otro lado, también es sabido que casi nadie lee más allá de la tercera o cuarta página de resultados que entregan los motores de búsqueda; como suelen listar 10 resultados por página, esto implica que casi nadie revisa los resultados que aparecen más allá de la posición número 30 o 40.

Así pues, para las empresas y organizaciones resulta cada vez más importantes que sus sitios y páginas web queden bien situados entre las listas de resultados que entregan los motores de búsqueda. Para conseguirlo existe toda una batería de procedimientos que pueden resultar relativamente eficaces. Estos procedimientos han dado lugar no solo a

una rama profesional muy especializada, sino a toda una industria dentro de las diversas economías relacionadas con Internet denominada posicionamiento web.

Por su lado, los ingenieros de los motores de búsqueda preferirían que las buenas posiciones en los resultados de las respuestas que ofrecen sus motores obedecieran exclusivamente a la calidad y relevancia propias de cada recurso. Por tanto, existe aquí algo levemente parecido a una carrera de armamentos: los expertos en posicionamiento intentan colocar en altas posiciones las páginas o los sitios de sus clientes, *independientemente* de la calidad intrínseca, mientras los responsables de los motores intentan que solo escalen posiciones los recursos en forma *dependiente* de la calidad intrínseca de los mismos.

Lo cierto, en todo caso, es que una página o un sitio web de gran calidad puede pasar totalmente desapercibido para los motores de búsqueda durante meses o años si los responsables del sitio no realizan al menos algunas acciones básicas de posicionamiento web.

Sea como sea, hace tiempo que los estudios y técnicas de posicionamiento web han adquirido madurez suficiente como para presentar una pequeña revisión sobre sus conceptos básicos y sobre los procedimientos más habituales.

Para presentar este estudio se ha seguido el siguiente procedimiento: en primer lugar, se realizó un análisis y revisión en profundidad de la producción técnica y científica más reciente sobre este nuevo campo de conocimiento. En segundo lugar, se utilizó la propia Internet para analizar las características del mercado, a través del análisis de algunas empresas que proporcionan servicios de posicionamiento web, tanto a escala nacional como internacional (en este último caso se utilizó, entre otros, el término "*web positioning*"). Por último, se realizaron diversos tests y pruebas de posicionamiento web a través de dos sedes web vinculadas con el autor (www.hipertext.net y www.documentaciondigital.org) y de dos campañas básicas de posicionamiento desarrolladas entre los meses de enero y noviembre de 2003.

En la tabla siguiente presentamos un grupo (discrecional) de conceptos básicos relacionados con el posicionamiento web. Hemos seleccionado los conceptos que parecen ser más comunes a la especialidad.

**Tabla 2.4: Definiciones sobre posicionamiento web**

| ***Término*** | ***Definición*** |
|---|---|
| **Posicionar** | Colocar una cosa en su lugar óptimo |
| **Posicionamiento web** | *Def1:* Colocar un sitio o una página web en un lugar óptimo entre los resultados proporcionados por un motor de búsqueda<br><br>*Def2:* Optimizar una página web de cara a los resultados proporcionados por los motores de búsqueda<br><br>*Def3:* Conjunto de procedimientos y técnicas para dotar a un sitio o una página web de la máxima visibilidad en Internet |
| **Palabra clave** | Término respecto al cual se persigue la optimización de una página web. |

| | |
|---|---|
| **Metadatos** | Datos sobre datos. En el contexto del posicionamiento web son datos sobre sitios o sobre páginas web que ayudan a su posicionamiento. Los metadatos adoptan, al menos estas cinco formas en el seno de páginas web:<br><br>1. Etiqueta <title> en la sección <head><br>2. Etiquetas <meta> en la sección <head><br>3. Atributos title en etiquetas de anclaje <a><br>4. Atributos title en etiquetas de imágenes <img><br>5. Atributos alt en etiquetas de imágenes <img><br><br>Por tanto, a efectos de posicionamiento web, también se consideran metadatos (o tienen un efecto similar) etiquetas y atributos distintos de las etiquetas tipo <meta> |
| **Popularidad** | Un término acuñado por el motor de búsqueda Google. La popularidad de un sitio está relacionada con el número de enlaces que recibe de otros sitios. Una alta popularidad es equivalente a un gran número de enlaces. En Google, además de considerar el número de enlaces que recibe una página se considera también la calidad de los mismos. |
| ***Page Rank*** | Una medida de la popularidad o número y calidad de los enlaces que recibe una página web. Esta medida es debida a Google. Una página web tiene mayor *pagerank* cuantos más enlaces recibe de páginas web que, a su vez, tienen un alto *pagerank*.<br><br>El *pagerank* tiene una escala de 0 a 10. Un sitio con una puntuación de 0 *pagerank* indica un sitio que no recibe ningún enlace o, al menos, que no recibe ningún enlace de un sitio web con *pagerank* alto (5 o más). Un sitio con una puntuación de 8 o superior indica un sitio que recibe numerosos enlaces, de los cuales, al menos una parte son enlaces de sitios web que a su vez tienen un *pagerank* alto (5 o más). Por ejemplo, la sede web de la CNN (www.cnn.com) obtenía un *pagerank* de 9/10 en diciembre de 2003<br><br>Google utiliza el *pagerank* de una página o de un sitio como uno de los principales elementos para el cálculo de relevancia de los resultados que entrega. |
| **Relevancia** | Capacidad de satisfacer una necesidad de información que presenta una página o un sitio web. Se dice que un recurso es muy relevante si es muy útil para solucionar a una necesidad de información.<br><br>La relevancia se mide siempre en relación a una necesidad de información dada que, a su vez, se expresa mediante una pregunta o ecuación de búsqueda. Si un usuario necesita información sobre automóviles de los años cincuenta, un recurso con información abundante y de calidad sobre turismos fabricados entre 1950 y 1960 será un recurso muy relevante. |

| | |
|---|---|
| | Los motores de búsqueda tratan de calcular la relevancia de manera automática, de modo que entregan los motores de búsqueda ordenados por el grado de probabilidad de resultar útiles para la necesidad de información expresada por el usuario. El cálculo de relevancia de cada motor de búsqueda combina diversas medidas, típicamente, la frecuencia y la densidad de las palabras clave de la pregunta. Google en particular utiliza también la popularidad (ver) como factor principal en su cálculo de relevancia. |
| ***Traffic Rank*** | Una medida del tráfico de un sitio web debido a *Alexa* (www.alexa.com), una empresa que realiza análisis y mediciones sobre tráfico en Internet y que proporciona también un directorio de sitios web. El *traffic rank* de un sitio indica tanto el número de usuarios que visita un sitio, como el número de páginas vistas por los usuarios.<br><br>El *traffic rank* es una cifra que varía de cero a infinito. Una baja cifra de traffic rank indica un gran número de visitas y de páginas vistas. Por tanto, cuanto más bajo es el número, mayor tráfico. Al contrario, cuanto más alto es el número, menor tráfico. Por ejemplo, el sitio de la *CNN* obtenía un 21, mientras que el sitio de *El País*, obtenía un 812 (ambos medidos en diciembre de 2003) |
| ***Spam*** | Prácticas destinadas a forzar una alta posición de una página web para una o más palabras clave, sin que tal posición tenga relación con la relevancia real de la página web. Los administradores de motores de búsqueda consideran fraudulentas las prácticas de *spam*. |

### 2.7.3. Tipología

Podemos ocuparnos ahora, para cerrar este grupo de apartados dedicados a la RI en la Web, de presentar una tipología de las cinco clases de herramientas para la *exploración* y la *explotación* de la información en la Web:

**Tabla 2.5: Herramientas de búsqueda en Internet**

| ***Navegación*** | ***Recuperación de Información*** |
|---|---|
| 1. Directorios o metadirectorios | 2. Motores de búsqueda<br>3. Multibuscadores<br>4. Difusión selectiva de la información<br>5. Bases de datos |

La tabla de la página siguiente ofrece una descripción sucinta de cada clase de herramienta junto con ejemplos de los servicios más característicos de cada clase:

**Tabla 2.6: Descripción de las herramientas básicas, uso típico y ejemplos principales**

| *Herramienta/ Descripción* | *Uso típico* | *Ejemplos* |
|---|---|---|
| **Directorios o metadirectorios**<br>Compilaciones de recursos hechas de manera intelectual ("a mano") y organizadas en forma de clasificaciones o estructuras jerárquicas de clases y subclases.<br><br>Existen por lo menos dos clases de directorios: directorios de recursos y metadirectorios, o directorios de directorios. | - Iniciar la exploración de un tema.<br>- Saber cuáles son los recursos considerados más importantes sobre un tema.<br>- Situar un tema de búsqueda en un contexto más amplio. | **Yahoo**<br>www.yahoo.com<br><br>**DMoz**<br>www.dmoz.org |
| **Motores de búsqueda**<br>Índices analíticos (índices invertidos, en términos técnicos) construidos por robots, es decir, de manera automática y consultables a través de lenguajes de consulta más o menos potentes. | - Explorar una gran parte de la Web de manera muy selectiva.<br>- Expresar necesidades de información muy concretas. | **Google**<br>www.google.com<br><br>**AllTheWeb**<br>www.alltheweb.com<br><br>**Scirus**<br>www.scirus.com |
| **Bases de datos**<br>Servicios de información que describen documentos o recursos digitales en un campo especializado. A diferencia de los motores de búsqueda, las descripciones que encontramos en una base de datos contienen informaciones con un alto valor añadido.<br>El primero de estos valores añadidos consiste en la selección de la información. Otros valores añadidos son la descripción estructurada por campos y la asignación de descriptores o de códigos de clasificación realizados de manera intelectual y a cargo de especialistas. | - Búsqueda de información retrospectiva sobre un campo concreto del conocimiento.<br>- Obtención de información a través de métodos muy selectivos.<br>- Búsqueda de documentos para uso en ciencia y tecnología. | **Sosig**<br>www.sosig.ac.uk<br>(Base de datos)<br><br>**IMDB**<br>www.imdb.com |

| | | |
|---|---|---|
| **Multibuscadores**<br>Sistemas capaces de enviar preguntas a varios buscadores a la vez. El servicio básico de los mejores metabuscadores incluye eliminación de duplicados y re-ordenación por relevancia o por agrupación en clases. En contrapartida, no permiten explotar todas las posibilidades de los motores de búsqueda individuales. | - Realizar una búsqueda muy exhaustiva enviando la misma pregunta a varios motores a la vez mediante una sola operación de búsqueda. | **Vivísimo**<br>www.vivisimo.com<br><br>**2.5 AskJeeves**<br>www.aj.com |
| **Difusión selectiva de la información**<br>Servicios que permiten registrar un perfil temático de búsqueda a nombre de un usuario. A petición del usuario, o periódicamente y de manera automática, se actualiza la búsqueda para localizar solamente los últimos documentos o informaciones publicadas sobre el tema que expresa el perfil. | -Mantenerse al día en un tema, típicamente después de haber realizado una búsqueda retrospectiva a través de un motor de búsqueda o de una base de datos o de ambos. | **My News**<br>www.mynewsonline.com<br><br>**Sosig**<br>www.sosig.ac.uk<br>(servicio *My Account*) |

## 2.8. Conclusiones

En relación a las técnicas de RI y de indización automática de documentos, Internet ha demostrado que, en los algoritmos, llamémosles "clásicos", como los que hemos examinado aquí, había una cierta cantidad de ideas preconcebidas.

Por ejemplo, la RI nunca se había enfrentado a un entorno tan heterogéneo como pueda ser el WWW. En este entorno, el rendimiento irregular de los motores de búsqueda convencionales demuestra el papel importantísimo de la selección y filtraje de calidad previa que tradicionalmente han llevado a cabo las bibliotecas y los centros de documentación. Por tal motivo, no es de extrañar que algunas soluciones a la búsqueda de información en la Web pasen por restituir ese papel previo de selección. El ejemplo más claro en este terreno lo constituye la iniciativa denominada *Resource Discovery Network* (www.rdn.ac.uk), donde los recursos de Internet son seleccionados, evaluados y descritos por equipos de expertos, aunque después las operaciones de recuperación se realizan con la asistencia, en modo automático, de un sistema de tesauros.

Otra aproximación diferente, pero en la misma línea de crear entornos más homogéneos la representa la iniciativa *Scirus* (www.scirus.com), un motor de búsqueda convencional en casi todo, excepto en que únicamente indiza servidores web de universidades e instituciones similares del mundo de la ciencia y la cultura.

En el futuro, los sistemas "inteligentes" de indización podrán incrementar su eficiencia, probablemente en base a: primero, considerar las propiedades lingüísticas de los textos,

y no tan sólo las estadísticas; y segundo, incorporar el uso de instrumentos de control terminológico como los tesauros, taxonomías u ontologías que permitan la categorización de los documentos y, por tanto, la reducción de la heterogeneidad y del espacio de búsqueda.

Esta última sería una relación muy adecuada de esfuerzo intelectual (o sea, hecho por personas) y de automatismo (o sea, de operaciones hechas por máquinas). Parece que es por aquí por donde podrán ir en el futuro los sistemas de RI de próximas generaciones. Con esfuerzo intelectual se construyen los tesauros u ontologías pero, una vez construidos, se podrían clonar tantas veces como hiciera falta, y su uso pasaría a ser automático en vez de manual, ya que los tesauros serían consultados y aplicados como resultado de reglas de producción de sistemas expertos.

En cualquier caso, y como ya hemos señalado en otra parte, la RI es un campo de trabajo y de estudios interdisciplinarios, la importancia del cual no dejará de aumentar mientras la Web vaya estando cada vez más presente en la vida de los ciudadanos, profesionales e investigadores.

## 2.9. Bibliografía


Abadal, E. *Sistemas y servicios de información digital*. Gijón: Trea, 2001, 147 pp.

Baeza-Yates, R.; Ribeiro-Neto, B. *Modern information retrieval*. New York: Addison-Wesley, 1999. 513 p.

Blair, D.C. *Language and representation in information retrieval.* Amsterdam: Elsevier, 1990. 335 pp.

Blair, D.C. “The challenge of commercial document retrieval, Part I: Major issues, and a framework based on search exhaustivity, determinacy of representation and document collection size”. *Information Processign and Management*, v. 38, 2002, pp. 273-291

Buckland, M. *Information and information systems*. Westport: Greenwood Pres, 1991, 225 pp.

Belkin, Nicholas J.; Croft, W. Bruce. “Retrieval techniques”. *Annual Review of Information Science and Technology*, n. 22, 1987

Chorafas, D. N. *Intelligent multimedia databases: from object orientation and fuzzy engineering to intentional database structures*. Englewood Cliffs, New Jersey: Prentice Hall, 1994, 360 pp.

Chowdhury, G.G. *Introduction to modern information retrieval*. London: Library Asociation, 1999, 451 pp.

Codina, L. "Sistemas automáticos de recuperación de información textual". En: Gomez Guinovart, J. *Aplicaciones lingüísticas de la informática.* Santiago de Compostela: Tórculo, 1994, pp. 63-86.

Codina, L. "Recuperación de información e hipertextos: sus bases lógicas y su aplicación a la documentación periodística". En: Fuentes, M. Eulália (ed.). *Manual de Documentación periodística*. Madrid: Síntesis, 1995, p. 213-230.

Codina, L. "Teoría de recuperación de información: modelos fundamentales y aplicación a la gestión documental". *Information world en español*, n. 38, octubre 1995, p. 18-22.

Ellis, D. *New horizons in information retrieval*. London: The Library Asociation, 1990, 138 pp.

Fox, Edward A. (1987). “Recuperación de información: investigación de nuevas posibilidades”. En: *CD-ROM: el nuevo papiro*. Madrid: Anaya, 1987.
Frakes, W. B.; Baeza-Yates, R. (eds). *Information retrieval: data structures & algorithms*. Englewod Cliffs: Prentice Hall, 1992, 504 p.
Gillman, Peter (ed.). *Text retrieval: the state of the art*. London: Taylor Graham, 1990, 208 pp.
Kowalski, G. *Information retrieval systems: theory and implementation.* Boston: Kluwer, 1997, 282 pp.
Lancaster, F. W. *Indexing and abstracting in theory and practice*. Champaing (IL): University of Illinois, 1998, 412 pp.
Losee Jr., R.M. *The science of information*. San Diego: Academic Pres, 1990, 293 pp.
Penrose, R. *La nueva mente del emperador*. Madrid: Mondadori, 1991. 597 p.
Rijsbergen, C. J. van (1981). *Information Retrieval*. Londres: Butterworths, 1981.
Salton, G.; Mcgill, M. J. *Introduction to modern information retrieval.* New York: McGraw-Hill, 1983, 448 pp.
Salton, G. *Automatic text procesing: the transformation, analysis, and retrieval of information by computer*. Reading (MA): Addison-Wesley, 1989, 530 pp.
Searle, John R. (1990). “¿Es la mente un programa informático?”. *Investigación y Ciencia*, nº 162, Marzo1990
Soergel, D. *Organizing information: principles of data base and retrieval systems.* Orlando: Academic Pres, 1985, 450 pp.
Sparck Jones, K; Willett, P. *Readings in information retrieval*. San Francisco: Morgan Kaufmann, 1997.
Van Slype, Georges (1991). *Los lenguajes de indización: concepción, construcción y utilización en los sistemas documentales*. Madrid: Fundación Germán Sánchez Ruipérez, 1991. 198 pp.
Vickery, B.; Vickery, A. *Information science in theory and practice*. London [etc.]: Bowker-Saur, 1987. 384 p.

**Sitios Web**

**Center for Intelligent Information Retrieval**
http://ciir.cs.umass.edu/

**Visualización de la Información**
http://www.infovis.net

**Agent-Based Systems**
http://www.agentbase.com/survey.html

**Search Engine Watch**
http://www.searchenginewatch.como

**Center for Networked Information Discovery and Retrieval**
http://www.cnidr.org

**The Information Retrieval Group**
http://ir.dcs.gla.ac.uk/

# 3 Sistemas de Gestión Documental: producción y administración de bases de datos

## 3.1 Introducción

En el capítulo 1 ya se ha hecho alusión a los conceptos de *base de datos*, y *de Sistema de Gestión de Bases de Datos*. Recordemos que, mientras una *base de datos* es un conjunto de *datos* (aunque no cualquier conjunto de datos, sino aquellos que están estructurados de una forma concreta), el concepto de *Sistema de Gestión de Bases de Datos* (SGBD a partir de ahora) es un programa informático. Obviamente, este programa informático sirve para crear, gestionar y explotar bases de datos. El eje central de este capítulo va a ser los Sistemas de Gestión Documental (SGD), aunque dedicaremos también un pequeño apartado a revisar las características de los Sistemas de Gestión de Bases de Datos de tipo *Relacional* (SGBDR) a fin de tener un mejor conocimiento, por comparación, de los sistemas documentales. En relación a estos últimos, distinguiremos a los Sistemas de Gestión de Bases de Datos Documentales (SGBDD), que se caracterizan por su capacidad para definir diccionarios de datos y que se ocupan básicamente de la gestión de referencias (véase 3.2) de los sistemas de indización o motores de búsqueda, cuya prioridad es la indización y que permiten la gestión de bases de datos de texto completo (véase 3.3).

### 3.1.1 Sistemas de Gestión de Bases de Datos Relacionales (SGBDR)

Los Sistemas de Gestión de Bases de Datos Relacionales (SGBDR) están concebidos para tratar información muy estructurada de tipo numérico o textual; una información que experimenta modificaciones constantes (p.e. los datos contables que cambian día a día). Estos programas acostumbran a disponer de facilidades para efectuar cálculos y tratamientos estadísticos básicos con los valores numéricos que gestionan. Son poco aptos para gestionar informaciones textuales relativamente extensas y no estructuradas (p.e. las que forman parte de un típico artículo de revista).

Lo cierto es que facilitan la gestión de datos de muy diversa naturaleza, desde listas de direcciones (directorios) hasta colecciones estadísticas y, como hemos destacado en otros apartados, están especialmente adaptados a las necesidades de la gestión empresarial y administrativa en general. En este sentido, actualmente la mayoría de las empresas y organizaciones disponen de uno o más SGBDR como núcleo de sus procesos de gestión, administración y toma de decisiones.

Este tipo de programas se denominan *relacionales* porque aplican el *modelo relacional*, una metodología de análisis de datos que se basa, al menos, en tres elementos fundamentales:

- *Un modelo de registro tabular*, es decir, que utiliza tablas para representar entidades. En las tablas, cada fila es una entidad (p.e., un cliente) y cada columna un atributo de la entidad (p.e. el apellido). Una base de datos relacional típica tendrá diversas tablas. Por ejemplo, una tabla para clientes, otra para vendedores, etc.

- *Un álgebra relacional,* que es una forma lógico-matemática de realizar operaciones con las filas y las columnas de las tablas que forman las bases de datos. Por ejemplo, existen operaciones que permite crear una tabla como respuesta a una pregunta. Las tablas de la respuesta se han creado con las filas de dos tablas distintas. Por ejemplo, una tabla que relaciona los datos de clientes y vendedores.

- *Capacidad multibase:* como consecuencia de los dos elementos anteriores, un sistema de gestión de bases de datos de tipo relacional debe ser capaz de abrir varias bases de datos a la vez.

En 1985, Codd detalló un conjunto de doce reglas que deberían cumplir los programas que quisieran considerarse relacionales.[2] Según como sea la aproximación a estas reglas se puede establecer una gradación de programas que iría desde sistemas simplemente tabulares, pero no relacionales, a sistemas totalmente relacionales (como *Oracle, Access* y otros), pasando por sistemas parcialmente relacionales, que se encontrarían entre ambos polos.

¿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 y si no necesitamos prestaciones de control terminológico.

Veamos los dos aspectos por separado. Los SGBDR no indizan todo el contenido de los campos de texto. Por defecto, los campos con mucha información 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 rápido, una búsqueda secuencial puede imitar el uso de un índice. Sin embargo, en cuanto crezca la base de datos, las prestaciones del sistema se degradarán. La experiencia indica que, a partir de unos pocos miles de documentos, un sistema relacional difícilmente podrá gestionar con eficacia su contenido. 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.).

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 en 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 bibliografías, etc.

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

[2] Algunas de estas reglas tenían un carácter fundacional, en el sentido que hacían referencia a la propia definición del modelo relacional, otras eran estructurales (la información se presenta por medio de tablas, etc.), otras hacen referencia a las reglas de integridad o de manipulación de los datos y, finalmente, existen otras que aluden a la independencia de los datos (física, lógica y de distribución).

### 3.1.2 Sistemas de Gestión Documental (SGD)

Los Sistemas de Gestión de Bases 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, es decir, para llevar a cabo las operaciones típicas de recuperación de información discutidas en el capítulo 2.

Como ya sabemos, 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*.

De acuerdo con nuestros propios análisis teóricos y empíricos, este modelo presenta, al menos, cinco importantes características:

*(i) 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 (*File Maker* o *Knosys*). El modelo irrestricto se refiere también a la posibilidad de que en una misma datos puedan convivir modelos de registros distintos (*CDS/ISIS* o *Inmagic*).

*(ii) Capacidad monobase o multibase indistintamente*
Es típico de algunos SGD que únicamente puedan abrir y operar una sola base de datos cada vez (*askSam, Knosys*). De aquí viene que se les denomine a veces sistemas planos, o bases de datos de datos planas. Sin embargo, cada vez son más los SGD con capacidad *multibase*, es decir, que pueden abrir y operar con más de una base de datos a la vez (*CDS/ISIS*, *Inmagic*, o *File Maker*).

*(iii) 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 temas 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.

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

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

***Tabla 3.1. 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 también el 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 3.2. 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 3.2 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 n. 0017 la palabra *Barcelona* aparece en el **Título** (campo 01) y en cambio en el registro n. 03401 aparece en el campo **Descriptores** (campo 04).

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 | U. Eco |
| 03 | Fuente | Vic: Editorial ZYX, 2002 |
| 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).

*(iv) Herramientas de control terminológico y/o lingüístico*
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.

*(v) 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, fundamentalmente, por dos razones: en primer lugar, porque el tipo de información que contienen es distinto 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.

### 3.1.3. Síntesis SGBDR *v* SGD

Las diferencias comentadas entre sistemas relacionales y documentales las sintetizamos y sistematizamos en la tabla 3.3. 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. Aun así, es útil comparar los modelos "puros" de cada categoría.

***Tabla 3.3. Principales diferencias entre SGBD***

| ***SGBDR*** | ***SGD*** |
|---|---|
| ***Estructura*** | |
| - Tabular (tablas para representar datos).<br>- Campos con longitud fija.<br><br>- No puede haber grupos de repetición.<br>- Tablas homogéneas | - Textual (modelo irrestricto).<br><br>- Campos y registro de longitud variable<br>- Campos repetibles<br>- 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 (*Idealist*), 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*). |
| - No disponen de controles terminológicos. | - Disponen de instrumentos de control terminológico para la indización, 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.). |

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:

- *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 documento completo. Algunos programas actuales de gestión documental siguen moviéndose en ese ámbito como una forma de

especialización. Los denominamos SGBDD y en el apartado 3.2 se tratan con mayor profundidad. Algunos ejemplos son *CDS/ISIS, Inmagic*, o *Knosys*.

.

- *Sistemas de indización*

Estos sistemas están especialmente orientados al tratamiento del texto completo de los documentos, además de la referencia. Son programas que no necesitan definir modelos de registro, aunque lo pueden hacer de modo opcional algunos de ellos. 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 indización o motores de búsqueda y en el apartado 3.3 se tratan con mayor profundidad. Algunos ejemplos son *Autonomy, BRS, Retrievalware,* o *Verity.*

Para diferenciar entre ambos modelos 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 indización tienen muy desarrollado el módulo de indización, que permite generar los índices invertidos de documentos extensos.

## 3.2 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. Agrupamos las funcionalidades en cinco módulos básicos:

- Definición de la base de datos
- Mantenimiento
- Indización y recuperación
- Salida e intercambio
- Administración y gestión de la base de datos

### 3.2.1 Definición 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*.

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.).

Como decimos, se trata de unas operaciones realizadas en el momento de crear la base de datos y que, normalmente, no necesiten ser alteradas una vez creada la base de datos.

Las principales prestaciones que permiten distinguir en este sentido a un buen SGBDD son las siguientes:

- *Asignación de tipos de dato a tipos de campo.*
Si un campo como, por ejemplo, el campo *autor* se asigna al tipo de dato *textual*, entonces será posible realizar una gama de operaciones distintas que si otro campo, por ejemplo, *alta*, se asigna al tipo de dato *fecha*. El tipo de dato *textual* nos permitirá, por ejemplo, realizar ordenaciones alfabéticas, mientras que el campo de tipo de dato *fecha* nos permitirá buscar por rangos de fechas, etc. Por último, un tipo de dato *numérico* como en *precio*, nos permitirá realizar operaciones aritméticas e incluso aplicar algún estadístico básico, dependiendo de la base de datos. Cuantos más tipos de datos proporcione un SGBDD mayores controles de calidad podrá tener la base de datos, así como mejores y más inteligentes formas de explotación. Los tipos de datos típicos son: *textual*, *numérico*, *lógico* y *fecha*; aunque la lista no es cerrada y algunos SGBDD utilizan tipos propios o exclusivos. Por ejemplo, *Idealist* dispone del tipo de datos *name*, lo que permite realizar ordenaciones alfabéticas de nombres interpretando correctamente la coma en los casos de inversión (p.e., en el caso de un nombre entrado en la forma "*Eco, Umberto*", *Idealist* sabría que debe ser mostrado como Umberto Eco, pero al mismo tiempo lo ordenaría por el apellido).

- *Posibilidad de disponer de diversos modelos de registro en una misma base de datos*.
Con frecuencia es importante poder disponer de distintos modelos de datos para tipos de documentos distintos (monografía, artículo, publicación periódica, etc.). De esta forma se consigue una mejor explotación de la información y se evitan errores en la entrada de los datos. Esta posibilidad puede quedar paliada por sistemas que admiten la posibilidad de abrir varias bases de datos a la vez. En este caso, cada base de datos se destina a un tipo de documento.

- *Definición de vistas e informes.*
Las vistas son versiones de cada modelo de registro adaptadas a categorías de usuarios. Una base de datos tiene al menos tres categorías de usuarios: el administrador, los operadores y los usuarios finales. Suele ser conveniente que cada clase de usuario tenga una versión del modelo de registro adaptado a sus necesidades de trabajo. Por ejemplo, el administrador querrá ver todos los campos, pero los usuarios finales, en cambio, no necesitan ver los campos de gestión (fecha de alta, por ejemplo). Por otro lado, no siempre deseamos explotar de igual modo la información. En ocasiones, necesitaremos que el resultado final se muestre en un único documento (informe) en forma de tabla; en otras ocasiones querremos que el resultado prescinda de ciertos campos, finalmente, podremos necesitar formatos de informes que unan campos de dos o más bases de datos distintas en un documento unificado, etc. El apartado de definición de informes es por tanto de gran importancia en un buen SGBDD.

- *Controles terminológicos*
Los controles terminológicos sirven para garantizar la coherencia y calidad de la información en la entrada de datos, así como para facilitar la recuperación de información en la consulta. Estos controles terminológicos pueden realizarse mediante

diversos medios, que van desde la definición de listas de palabras vacías hasta el uso de gestores de tesauros más o menos integrados en el programa pasando por el uso de listas de validación o descriptores admitidos para un campo determinado. Las listas de palabras vacías contienen los términos que **no** deben figurar en los índices; términos tales como artículos, preposiciones, etc. Se trata de términos sin significado propio y cuya aparición en los índices, además de degradar las prestaciones en cuanto a velocidad ofrecen una pobre imagen al usuario que desea conocer el contenido de la base mediante la consulta del índice.

- *Posibilidad de relacionar bases de datos.*
Algunos programas (p.e. *Inmagic*, *FileMaker* o *CDS/ISIS*) permiten establecer relaciones entre distintas bases de datos. Aunque no cumplen todas las especificaciones del modelo matemático relacional son muy útiles para la mayor parte de las necesidades de gestión documental. Así, p.e., Inmagic ha desarrollado un modelo relacional que permite relaciones bastante complejas entre bases de datos. De este modo, con Inmagic se han definido aplicaciones para bibliotecas que pueden relacionar diversas bases de datos: documentos, préstamo, usuarios, proveedores, etc. Cada modelo relacional particular de cada programa presenta prestaciones diversas, pero en general, se espera de la capacidad relacional al menos la posibilidad de cruzar datos de diversas bases e incluso de componer informes unificados con datos procedentes de bases de datos distintas.

- *Ordenación de los artículos iniciales y otras prestaciones de gestión bibliográfica.*
Dada su orientación típica a la gestión de documentos publicados (bibliografías), algunos programas proporcionan algunas prestaciones que facilitan la gestión de las mismas. Un ejemplo, sería la correcta ordenación de registros según un campo concreto (por ejemplo, *título*) ya que no tiene en cuenta los artículos iniciales como elementos de ordenación. Así pues, podemos escribir “El tiempo” teniendo la plena certeza que cuando se efectúe una ordenación por este campo el registro se ordenará dentro de la letra “t” de “*tiempo*”, palabra rica en significado, y no en la “e” de “*El”*, palabra muy pobre en significado.

### 3.2.2 Mantenimiento

El módulo de mantenimiento controla las operaciones de altas, bajas y modificaciones de registros. Las principales funciones que se pueden resaltar son las siguientes:

- *Facilidades en las operaciones de altas*
Un buen SGBDD puede proporcionar un amplio abanico de soluciones para la realización de las altas. Por ejemplo: definir plantillas con parte de los datos ya ingresados (fecha de alta, número del registro, nombre del operador, etc.), duplicación de registros en caso de grupos de documentos donde solamente cambian uno o dos campos (p.e., datos sobre los diversos volúmenes de una misma obra, datos sobre las distintas canciones de un mismo disco de un mismo autor, etc.). Otras facilidades están relacionadas con la posibilidad de seleccionar valores con uno o dos clics a partir de listas desplegables o de cuadros de selección, etc. Finalmente, un buen SGBDD debe contemplar la posibilidad de realizar altas de modo automático por importación de ficheros.

- *Realización de modificaciones globales*
Esta posibilidad es necesaria cuando se han de gestionar bases de datos voluminosas. Es decir, puede darse la necesidad de cambiar un descriptor, de modo que sea necesario que, donde aparece el término *Comunidad Europea* en los campos descriptores, aparezca ahora el término *Unión Europea*. Aunque existan cientos o miles de registros con el descriptor antiguo, un sistema de modificaciones globales convierte en trivial la operación si se hace tomando las precauciones adecuadas de seguridad.

- *Corrector ortográfico*
Dada la vocación textual de los SGBDD, cada vez es más habitual disponer de esta posibilidad en las operaciones de alta de documentos.

- *Asociación de ficheros externos al registro* (p.e. ficheros de imagen, sonido, etc.)
Cada vez es más frecuente la necesidad de disponer de documentos gráficos, sonoros o textuales asociados a los registros de la base de datos. La asociación debe ser inteligente, es decir, que al activar el enlace, el programa envíe la orden correspondiente al sistema operativo para que muestre el objeto enlazado (una imagen, p.e.) en el programa original (un editor o un visualizador de imágenes, en este ejemplo).

- *Detección de duplicados*
Se trata de una operación compleja ya que, en muchas ocasiones, no se trabaja a partir de campos clave que permiten diferenciar un registro de los otros.

### 3.2.3 Indización 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, los sistemas para buscar en ellos y las formas de mostrar y ofrecer los resultados a los usuarios. Las principales prestaciones diferenciales de los SGBDD son las siguientes:

- *Tipos de indización*
Los tipos básicos de indización son: palabra a palabra, grupos de palabras, campo entero o subcampo (en el caso que se puedan definir). La mayoría de programas sólo permiten la indización palabra a palabra, algunos de ellos permiten también la indización de grupos de palabras (es el caso de *Inmagic*) y tan sólo sabemos de un programa (*Winisis*) que permite los cuatro tipos. Otras variaciones en el control de la indización se refieren a indizar solamente las palabras de los campos presentes previamente en una lista (*Knosys*) o la de indizar las palabras que el operador humano marque o seleccione en un campo (*Knosys*). Igualmente, la indización puede estar influenciada por el uso o no de controles terminológicos. En el último caso, se habla de indización libre (o *free text*). En cambio, si la indización se realiza con la mediación de, por ejemplo, un tesauro, se habla de indización controlada.

- *Indización inmediata o diferida*
Es también interesante poder escoger entre la indización inmediata de los registros o la generación de los índices a posteriori, una vez se han introducido grupos importantes de registros.

- *Acceso a la información: exploración e interrogación*
Los tipos básicos de acceso a la información en SGBDD son la exploración y la interrogación. En cuanto a la primera, el sistema debe permitir la exploración *secuencial* (desplazamiento por todos o por grupos de registros secuencialmente, sin necesidad de formular ninguna consulta), la exploración por *índices* (de campo o globales) y la exploración por tesauro. En cuanto a la interrogación, debe permitir consultas asistidas mediante formularios (o medios similares) y consultas avanzadas, ya sea utilizando directamente el lenguaje de interrogación o por medio de formularios. Otro grupo de funciones relacionadas con este apartado se refieren a las prestaciones del lenguaje de interrogación. Las más habituales son: búsqueda por palabras, búsqueda por rangos de valores, búsqueda por frases, búsqueda booleana simple y búsqueda booleana compleja con uso de paréntesis. Prestaciones adicionales del sistema de acceso a la información pueden consistir en el almacenamiento y reutilización de consultas y la combinación de resultados de diversas consultas. Es importante poder acceder a todas estas opciones de forma rápida, lógica y mediante opciones bien agrupadas.

- *Tratamiento de caracteres*
Esto último se refiere, por ejemplo, a la sensibilidad a la búsqueda con acentos y a minúsculas y mayúsculas. Algunos programas, p.e. *Inmagic*, diferencian los términos acentuados de los que no lo son, mientras que otros programas convierten todos los términos a mayúsculas sin acentuar. Lo mismo sucede con el caso de las mayúsculas y minúsculas. Algunos programas facilitan mejor que otros la ordenación correcta de los caracteres latinos (ñ, acentos, etc.) en el fichero inverso.

### 3.2.4 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. Las principales funciones para la evaluación son las siguientes:

- *Formatos de importación*
En ocasiones la importación puede ser una tarea compleja, como poder diseñar y preparar (interpretar o descodificar) la información de un fichero de registros externos para poder incorporarlo a una base de datos. En otras ocasiones, se trata de la conveniencia de que el programa pueda interpretar directamente formatos bien establecidos (p.e., ascii, hojas de cálculo, dbase, etc.)

- *Formatos de exportación*
Por la misma razón, es muy importante que un SGBDD sea capaz de producir ficheros de salida en formatos estándar o en formatos de programas de amplia difusión (formato ascii, access, etc.).

- *Acceso a través de la web*
Cada vez, más SGBDD están adaptados para su consulta e incluso su manejo a través de la web, es decir, mediante un navegador de Internet estándar. *Inmagic*, por ejemplo, dispone de un programa denominado *Web Publisher*, que debe estar situado en el

ordenador servidor, y permite la consulta de la base de datos a través de Internet. *FileMaker*, por su parte, permite realizar todas las tareas, incluidas las de mantenimiento a través de Internet si se tiene acceso a un servidor de Internet.

### 3.2.5 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.

- *Sistema de seguridad*
Para asegurar la privacidad del acceso a la base de datos, el programa ha de permitir la creación de 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. Es necesario, además, distinguir diversos niveles de acceso: control del mantenimiento (sólo podrán realizar altas, bajas y supresiones de registros los usuarios explícitamente autorizados), acceso a la base de datos, control de impresiones, etc. En otro sentido, muy distinto, pero no menos importante, la seguridad se refiere a controles sobre la integridad de los datos y las facilidades para la realización de de copias de seguridad periódicas.

- *Programación y modificaciones en la interfaz*
Además de las vistas e informes, algunos programas van más allá y permiten configurar en buena parte el aspecto y las funciones que podrán ver los usuarios, así como permiten la creación de pequeños programas (denominados *scripts*) que automatizan parte de las funciones de la base de datos. Estos programas pueden ser invocados de manera automática cada vez que el usuario abre una base de datos o pueden activarse a través de botones y teclas de función. En este sentido, destaca con méritos propios *Inmagic*, que permite que el administrador (o cada usuario si dispone de privilegios) genere interfaces personalizadas para la consulta, el mantenimiento y otros apartados de la aplicación.

### 3.2.6 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 *WinIsis*) o de un nivel medio (en esta situación están *FileMaker* y *Knosys*).[3]

| | |
|---|---|
| ***Nombre*** | *CDS/ISIS - WinIsis* |
| ***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. |

[3] También podrían citarse *Idealist (*www.bekon.com) y *askSam* (www.asksam.com), como programas para uso personal.

| | |
|---|---|
| | - Utilización de subcampos.<br>- Indización con diversas técnicas (palabra a palabra, grupos de palabras, campo entero, subcampos).<br>- Alfabetización correcta de los campos de título (obviando los artículos iniciales).<br>- Definición de formatos de visualización siguiendo normativas de descripción bibliográfica.<br>- Compatibilidad con el formato ISO 2709.<br>- Posibilidad de relacionar bases de datos.<br>- Lenguaje de programación.<br>- Precio: gratuito.<br>Lista de distribución muy activa <http://www.bib.wau.nl/isis/isislist.html>.<br><br>Inconvenientes:<br>- Complejidad para el diseño de la base de datos.<br>- Sin soporte comercial. |

| | |
|---|---|
| ***Nombre*** | *FileMaker* |
| ***Productor*** | Claris <http://www.filemaker.fr/spain> |
| ***Distribuidor*** | Claris <http://www.filemaker.es> |
| ***Características*** | - De muy fácil utilización.<br>- Muy versátil, de hecho, es el programa con capacidad documental más integrado a la vez en el mundo ofimático.<br>- Capacidad relacional.<br>- Escasas herramientas de control terminológico.<br>- Lenguaje de *scripts.*<br>- Amplias posibilidades de personalización y modificación de la interfaz de la base de datos.<br>- Precio: bajo.<br>- Club de usuarios: http://www.fm-club.org/ |

| | |
|---|---|
| ***Nombre*** | *Inmagic DB/TextWork* |
| ***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.<br>- Capacidad relacional.<br>- Amplias posibilidades de adaptación y personalización de las interfaces de usuario.<br>- Amplias posibilidades de control terminológico.<br>- Gestión integrada de tesauros.<br>- Dos formas distintas de indización (palabra a palabra, por frases).<br>- Índices por cada campo.<br>- Posibilidad de programación mediante *scripts.* |

| | |
|---|---|
| | - Precio: medio.<br>- Dispone de club de usuarios (para más datos, contactar con el distribuidor). |

| | |
|---|---|
| ***Nombre*** | *Knosys* |
| ***Productor*** | Micronet <http://www.micronet.es> |
| ***Distribuidor*** | Micronet |
| ***Comentarios*** | - Fácil utilización.<br>- Tecnología española.<br>- Excelente interfaz de consulta.<br>- Posibilidades limitadas de control terminológico.<br>- No diferencia entre el fichero de definición de campos, entrada de datos y visualización.<br>- Precio: medio. |

### 3.3 Sistemas de indización o motores de búsqueda

Los motores de búsqueda, también denominados indizadores o sistemas de indización, se han hecho justamente famosos a raíz del importante papel que están jugando los buscadores de páginas web como *Google* o *AltaVista*. 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) que encuentra en Internet.

Lo cierto es que antes que existiera la web, ya existían estos sistemas de indización. 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 70 y a principios de los 80. Es conocida (y está muy bien documentada por su amplia implantación) la existencia de otros programas de este estilo que funcionaban con grandes sistemas al menos desde la década de los 80 como *STAIRS*, *Basis*, *DOCU-MASTER*, etc. También durante los años 80 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. Catherine Leloup, en el siguiente párrafo, los define a partir de las dos acciones que realizan: generar los índices y buscar en ellos.

“Un motor de indización y búsqueda es una herramienta que permite extraer de una información, principalmente textual, las palabras o términos que mejor la representan para almacenarlas en un índice. Esta misma herramienta es la que después recorre todo el índice, a fin de identificar los términos más relevantes en relación con la pregunta del

usuario, y escoge las informaciones que le suministrará como respuesta". (Leloup, 1998: 17)

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 utilizan 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 indización) es que en este último no existe 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. Dicho de otro modo, todo sistema de gestión de bases de datos documentales debe poseer, como uno de sus subsistemas, un motor de búsqueda. Sin embargo, lo contrario no es cierto, de modo que algunos SGD (como los indizadores) consisten únicamente en motores de búsqueda autónomos, al estilo de los programas que están detrás de *Google* o *AltaVista*.

Así pues, los motores de búsqueda son un tipo de SGD que genera índices analíticos (ficheros invertidos) a partir del análisis del texto completo de los documentos de una colección o fondo documental (o de toda la web). Las entradas de este índice contienen punteros a los documentos originales. El conjunto formado por el índice y los punteros se denomina *base de datos*. Esto induce a confusión, ya que una base de datos, en rigor, es un conjunto de registros, que a su vez están formados por campos, etc., como ya hemos discutido en las secciones precedentes.

En teoría, por tanto, las entradas de un índice no son lo mismo que los registros de una base de datos documental (al menos, no en el mismo sentido). Ahora bien, en informática, cosas muy diversas reciben a menudo los mismos nombres. Por ejemplo, las posiciones de datos en los chips de memoria RAM de un ordenador se denominan también registros. Por tanto, el conjunto de registros de la memoria RAM se puede considerar una base de datos. Por esta razón, a la larga, se ha establecido la convención de hablar de *base de datos* también para referirse al conjunto de entradas del índice más los punteros a los documentos que genera un sistema de indización.

La siguiente tabla intenta ilustrar estos conceptos.

**Tabla 3.4. Diversos significados del término *base de datos***

| | |
|---|---|
| ***Sentido estricto*** *(Contexto de la teoría de SGBD)* | Conjunto de registros (es decir, de representaciones de entidades) creados y/o administrados por un sistema de gestión de bases de datos. |
| ***Sentido amplio*** *(Contexto de los sistemas de información)* | Conjunto de datos, de cualquier tipo, gestionados por un sistema de información. Ejemplo: el conjunto de las entradas del índice de un motor de búsqueda. |
| ***Sentido metafórico*** *(Diversos contextos)* | Cualquier colección de datos o informaciones, incluso en contextos extra informáticos. Ejemplo: colecciones de fichas de papel. |

Para describir el funcionamiento de este tipo de programa partiremos de los módulos básicos que se han descrito anteriormente para los SGBDD (véase 3.2), es decir, administración de la base de datos, mantenimiento, indización, recuperación, y salida e intercambio de información.

### 3.3.1 Administración del fondo documental

El primer paso que hay que dar es la definición de la base de datos o de la colección de documentos (un término que también se utiliza en este contexto).[4] La colección ha de tener un nombre y una ubicación que indica en qué directorio del servidor se almacenan los datos, y puede estar formada por diversos tipos de documento y en diversos formatos.

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 indización 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 y/o cuáles son las direcciones de los servidores remotos en los que 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. Por ejemplo, si el documento encontrado es una página web, podremos verla en un navegador, pero si se trata de un documento de texto, podremos verlo en *Word* o en *WordPerfect*, etc.

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. Por ejemplo, los documentos *Word* creados con las últimas versiones suelen retener como parte de su contenido la fecha de creación, el nombre del autor y otros datos, incluyendo un resumen (generado automáticamente) y datos estadísticos sobre el documento (el lector puede comprobar la clase de metadatos que una aplicación como *Word* guarda de cada documento haciendo clic en Archivo > Propiedades). También es el caso de los documentos web que incorporan etiquetas de metadatos.

[4] La colección es el conjunto de los documentos que serán indizados. Estos documentos pueden estar en una unidad de disco y un único directorio, o en un conjunto de unidades y directorios.

### 3.3.2 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 (si es así, son pocos los datos que se introducen de esta forma) porque normalmente se dispone ya de ficheros informáticos con la información que se ha de procesar (documentos html, documentos de texto, de hojas de cálculo, gráficos, etc.). Por tanto, la introducción de los datos se realiza mediante operaciones de tratamiento de archivos (con o sin importaciones de los mismos). Estas operaciones pueden ser interactivas, mediante elecciones de menú, o totalmente automáticas, mediante la indicación al sistema de las unidades o directorios que el sistema debe explorar en busca de los archivos a tratar y/o importar.

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, rtf, html, xlc, pdf, eds, tiff, etc.). En cualquier caso, estos programas están preparados para indexar el texto completo de documentos creados al menos con los formatos más habituales.

Los documentos indizados suelen mantenerse en su formato original y lo único que necesita el sistema es saber dónde se encuentran y con qué aplicación están generados para así poder facilitar la visualización cuando sea necesario después de una operación de recuperación.

### 3.3.3 Indización

El motor de indización crea unos índices invertidos (véanse las tablas 3.1 y 3.2) que son la base de su sistema de recuperación, y a los que van a parar todos los términos de los documentos excepto los que figuran en el fichero de palabras vacías.

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. Como es habitual en todo SGD, cuando se actualiza la colección (se añaden o se retiran documentos) hay que reindizar de nuevo para actualizar los índices.

También hay que disponer de algún mecanismo que sirva para limitar de alguna manera las tareas de indización y así evitar que se indizen, por ejemplo, todas las carpetas de una unidad (si no desea que se haga así) o que ciertas extensiones de archivo no sean consideradas (p.e. archivos con extensión .bak, etc.) o, en el caso de páginas web, que solamente se indizen algunos niveles del sitio web, etc. En el caso de sitios web, lo que a veces se hace es indicar el directorio en el que se encuentran los ficheros que han de formar parte de la base de datos y no permitir que la indización se haga a ficheros que no estén dentro de la raíz indicada. Ejemplo: queremos crear una base de datos con los artículos de una revista digital que se llama *BiD* y que se encuentra en <http://www.ub.es/bid>. Dado que sus artículos contienen enlaces a muchas otras páginas que no están en la dirección antes apuntada, tal vez haya que limitar la indización a las páginas que están bajo la raíz antes apuntada, de lo contrario nos encontraremos con la sorpresa que estamos indexando todo el web.

### 3.3.4 Recuperación

En general, el proceso de consulta en sistemas de indización 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. En este apartado nos dedicaremos a describirlas de forma teórica y práctica, en tanto que constituyen un camino que puede servir para superar las limitaciones de los sistemas actuales RI.

Estos dos tipos de búsqueda han seguido dos caminos totalmente diferentes: la primera de ellas realizado un análisis que tiene en cuenta la morfología, la sintaxis y la semántica de los términos; la segunda, en cambio, prescinde totalmente de estas características y parte de la estructura binaria de los términos. Este binomio profundidad-superficie (Ellis, 1998) respecto a la forma de representar la información nos va a servir de línea argumental para este subapartado.

#### 3.3.4.1 Búsquedas semánticas

Al usuario no especializado, el que viaja por el web y busca información de cualquier tipo, le resulta mucho más fácil expresar una necesidad de información (p.e. “estoy buscando colaboraciones en prensa de Umberto Eco”) que formular una ecuación de búsqueda (p.e. “AU=Eco, Umberto AND TD=prensa”). Los usuarios vienen de contextos muy diferentes y utilizan su propio vocabulario, sus propias palabras que, con frecuencia, no coinciden con las empleadas por el sistema de información. Por ello ha de ser el sistema el que ha de actuar con "inteligencia" para poder entender la petición, traduciéndola a los términos que se utilizan en la base de datos. Las redes semánticas –utilizadas, entre otros programas, por *Verity* y *Retrievalware*– constituyen el instrumento esencial que ayuda a ampliar la búsqueda que realiza el usuario.

El objetivo de las búsquedas semánticas –o por conceptos– es poder ampliar la consulta de un término a todos aquellos que estén relacionados de alguna manera con él –derivaciones morfológicas, equivalencias lingüísticas, sinonimia, antonimia, términos generales, específicos, etc. El sistema permite, pues, consultar la base de datos utilizando tesauros, diccionarios multilingües o diccionarios especializados previamente elaborados que indiquen cuáles son las relaciones entre los términos. Las ventajas de esta prestación para el usuario que consulta un sistema RI son claras:

- *No hay que conocer previamente el lenguaje de indización* (las consultas se pueden realizar mediante el lenguaje natural).
El usuario puede formular su pregunta utilizando los términos del lenguaje natural que le sean más próximos. El sistema tiene la capacidad de procesarla para eliminar las palabras sin significación –palabras vacías– y para poder relacionar automáticamente los términos significativos que quedan con los que constan en su diccionario interno. Para ello, el sistema ha de disponer de un diccionario de sinónimos que le permite ampliar la consulta a todos los términos que se utilizan para hacer referencia a un mismo concepto.

- *Se pueden utilizar de forma controlada las relaciones jerárquicas del tesauro*.
En los casos en los que se haya utilizado un tesauro para la indización, es posible incorporar automáticamente no tan sólo las relaciones de equivalencia –que constituyen el caso más frecuente– sino también las jerárquicas y las asociativas. Por ejemplo, si estamos buscando documentos sobre la situación actual de los medios de comunicación en España, el sistema ha de ser capaz de seleccionarnos los documentos en los que consta el descriptor "medios de comunicación" y también aquellos que traten sobre la "prensa", la "radio" o la "televisión".

Este tipo de búsqueda incorpora una red semántica inteligente, que está formada por un número variable de términos, en diversos idiomas, y en la que se especifican diversas relaciones entre términos.

A continuación, pasamos a describir un par de ejemplos que pueden servir para ilustrar todo esto:

- *V|lex* (www.vlex.com)

Se trata de un portal de información jurídica que contiene diversas bases de datos de este ámbito. Dispone de una funcionalidad que realiza una asistencia inteligente en la búsqueda, sugiriendo al usuario un listado de términos que están relacionados con el que solicita en un momento determinado.

Así, por ejemplo, si un usuario busca información sobre “Impuestos” el sistema le muestra un conjunto de términos relacionados con el que se ha solicitado, tal y como se puede observar en la figura 3.1. A partir de aquí se puede ampliar o restringir la consulta a aquellos que se consideren más adecuados.


Figura 3.1. Consulta en V|lex sobre impuestos

Si realiza una petición de documentos relacionados con “copyright”, entonces se le sugiere la relación que podemos ver en la figura 3.2.

Figura 3.2. Consulta en V|lex sobre copyright

Si pensamos en un sistema que dispone de documentos en diversos idiomas, esta aplicación podría asistir al usuario traduciéndole sus términos de consulta a todos los idiomas existentes en la base de datos.

- *Enciclopedia Britannica* (http://britannica.com)

En el sitio web de esta famosa enciclopedia se pueden realizar búsquedas utilizando el lenguaje natural. En este caso, se trata de una aplicación que pretende ahorrar al usuario el conocimiento del lenguaje de indización que utiliza el sistema. Así pues, las preguntas no han de ser necesariamente formalizadas y el usuario puede preguntar cosas del estilo: "¿cuál es el tercer río más largo del mundo?". El motor de búsqueda de la enciclopedia ignora las palabras vacías (cuál, es, el, del, más) y busca los artículos de la enciclopedia que contengan el resto de los términos (tercer, río, largo, mundo), que serán presentados al usuario en función de unos mecanismos de ponderación internos.

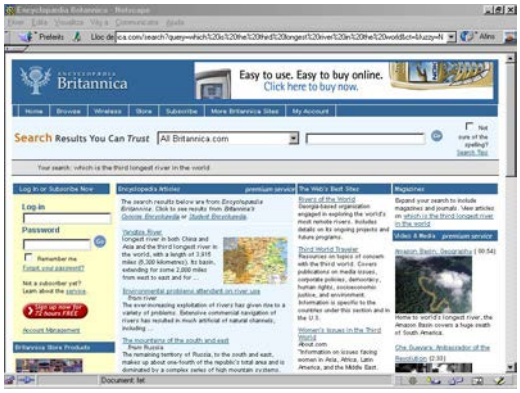


Figura 3.3. Consulta sobre cuál es el tercer río más largo del mundo en Britannica.com

En la figura anterior podemos ver, entre la lista de los resultados, el artículo del Yang-Tsé, el río más largo de Asia y tercero del mundo.

### 3.3.4.2 Búsqueda por patrones

Se trata de un sistema totalmente opuesto al anterior. En este caso nos estamos refiriendo 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 indización 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).

La búsqueda por patrones, o por reconocimiento de forma, permite buscar no tan sólo el término exacto que se introduce en la consulta sino todos aquellos términos que comparten en parte el mismo patrón, que tienen un parecido con el que ha indicado el usuario. La recuperación se realiza por aproximación a partir de unas reglas internas que rigen el proceso de comparación y que a veces incorporan elementos de la lógica difusa.

Las búsquedas patronales 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. Así pues, si buscamos, por ejemplo, “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.

El mismo modelo se puede aplicar a los documentos gráficos. Así pues, también se pueden buscar imágenes que respondan a un patrón como podría ser “una figura humana en el centro, un fondo azul, un coche rojo a la derecha”. A continuación, pasamos a describir una prestación de estas características que está implementada experimentalmente en el Museo del Hermitage y que utiliza el programa QBIC (IBM).

- *Museo Hermitage* (http://www.hermitagemuseum.org)

Dispone de una aplicación de búsqueda de imágenes a partir de formas y colores suministrada por la empresa QBIC (IBM) y que se encuentra en el apartado “Digital Collection”, opción “QBIC colour and layout searches”. Desde aquí podemos realizar consultas sobre obras artísticas del fondo del Museo a partir de peticiones no textuales. Así, por ejemplo, podemos dibujar un conjunto de formas en un recuadro (rectángulos, círculos, etc.) y añadirles color para que el sistema nos busque aquellos cuadros que tienen una semejanza con la disposición de los elementos y del color que hemos indicado. Los resultados, como puede comprobar cualquier persona que utilice esta opción, tan sólo son plenamente satisfactorios en algunas ocasiones y es por ello que está instalado de forma “experimental”. Aun así, no hay duda de que se trata de una vía que tiene abiertas muchísimas posibilidades.

En las dos figuras que se muestran a continuación reproducimos el sistema de consulta. En primer lugar, solicitamos una imagen con un rectángulo rojo en el cual se sobreinscribe un círculo amarillento. A continuación, nos aparecen los resultados, la

mayoría de ellos desestimables, aunque el primero tiene una notable coincidencia con las condiciones estipuladas.


Figura 3.4. Especificación de una consulta de imágenes

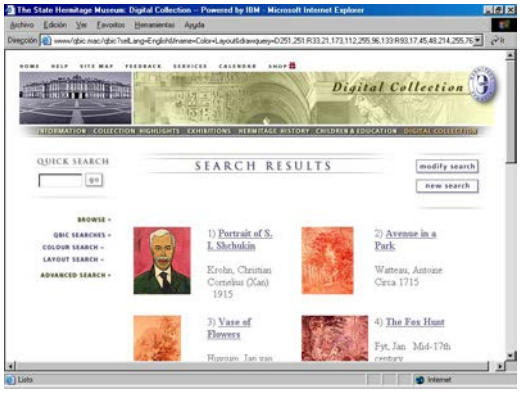



Figura 3.5. Resultados sobre la consulta

El funcionamiento de la aplicación parte de unos algoritmos internos elaborados por el programa y que, en muchas ocasiones, se basan en los métodos de las redes neuronales. Estas aplicaciones permiten indexar y recuperar información introduciendo patrones binarios en la información. Esto tiene dos importantes ventajas: por un lado, la independencia del idioma –cualquier información se almacena de la misma manera esté en el idioma en que esté– y, por el otro, la independencia del tipo de datos –tanto da que sea texto, imagen o sonido. Las palabras se descomponen en un mapa de bits y las búsquedas se realizan por medio de una comparación porcentual entre los mapas.

### 3.3.5 Ponderación de resultados

La utilización sin control de las búsquedas patronales y semánticas 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 que hay que disponer de instrumentos que ayuden a determinar cuáles son los más relevantes.

Existe una importante línea de desarrollo cuyo objetivo es la elaboración de algoritmos y sistemas que ayuden a determinar y ponderar la relevancia de los resultados. Muchos de ellos son sistemas de base estadística (que colocan en primer lugar los documentos con mayor frecuencia de aparición de los términos de consulta) aunque últimamente, gracias a la influencia de *Google*, se están imponiendo otros sistemas basados en la "popularidad" del documento. En el caso de *Google*, se utiliza el algoritmo *PageRank*, que se basa especialmente en el número y la calidad de los enlaces que se dirigen hacia una sede web, y que ya ha sido descrito en el capítulo 2.

### 3.3.6 Sistema de indización vs. SGBDD

Para finalizar, pues, si comparamos la estructura y funcionamiento de los motores de búsqueda con los módulos de un SGBDD clásico podremos constatar que los módulos en los que existen mayores diferencias son los de diseño de estructuras de registros (los motores de búsqueda no lo poseen) y en algunos apartados relativos a la indización y el mantenimiento. En lo que respecta al módulo de indización, está más desarrollada en los motores de búsqueda, ya que la cantidad de información que se ha de procesar es muy superior. En lo que se refiere al mantenimiento, hay que recordar que en los SGD clásicos la introducción de datos desde el teclado por parte de operadores humanos es una parte importante del proceso de creación, mientras que en los motores de búsqueda este apartado se resuelve, generalmente, indicando en qué directorios o servidores se encuentran los ficheros que se han de incorporar a la base de datos.

### 3.3.7 Aplicaciones

A continuación, vamos a indicar algunas de las principales aplicaciones de los sistemas de indización. Como se podrá comprobar después de la lectura de este apartado, no todas ellas son excluyentes, ya que podremos encontrar algún ejemplo práctico que podría formar parte de más de una categoría.

En estos ejemplos, el usuario puede comprobar las características que hemos descrito en la caracterización de los sistemas de indización: escoger entre una o más bases de datos; recuperación booleana, por patrones y, en menor medida, semántica; ordenación de los resultados por orden de relevancia; difusión selectiva de la información, etc.

- *Buscadores de recursos web*

Como ya hemos apuntado en la introducción, los buscadores de sedes y páginas web han sido los que han popularizado a los sistemas de indización, dándolos a conocer al gran público. Estos pueden ser generales (Google o Altavista), o particulares de una sola sede web, como puede ser el caso del buscador de la Universitat de Barcelona <www.ub.es/> o de cualquier otro organismo público o privado.

- *Indizadores en colecciones de texto completo (a veces llamados bases de datos de texto completo)*
Los motores de búsqueda también pueden utilizarse para crear bases de datos de texto completo de ámbitos temáticos o publicaciones distintos. En este caso existe 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.
P.e. *MyNews Online* <http://www.mynewsonline.com>, como servicio accesible a terceros o *ZyIndex* como sistema para crear esta clase de bases de datos en Intranets o servicios corporativos de información

- *Bases de datos (pseudo) etiquetadas*
En el caso anterior no se diferencian campos ni zonas dentro del texto completo. En muchas ocasiones, esto puede suponer una limitación para el usuario ya que no le permite afinar su consulta. Es por ello que cada vez es más frecuente la utilización de etiquetas, campos o metadatos para aumentar la precisión en el proceso de recuperación de la información. P.e. *Information Research* <informationr.net/ir>
En este caso se definen unas zonas del documento o se anexan metadatos.

- *Integración con sistemas gestores de intranets o trabajo en grupo*
Los programas de trabajo en grupo (p.e. Lotus Notes/Domino) tienen por objetivo permitir a las personas trabajar de forma conjunta y coordinada con las mismas versiones de ficheros (ya sean documentos de texto, hojas de cálculo, etc.). Estos programas se ocupan del control de las versiones de los ficheros, de su almacenaje, de su organización lógica y también, evidentemente, de la recuperación de su contenido. Esta última prestación, normalmente, corre a cargo de algún motor de búsqueda que se implementa dentro del sistema de trabajo en grupo. Así, por ejemplo, en el caso de Lotus Notes se ha integrado el motor de búsqueda de *Verity*.

### 3.3.8 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 presentes en el mercado español son:

| | |
|---|---|
| ***Nombre*** | AtomZ Search |
| ***Productor*** | AtomZ < http://www.atomz.com> |
| ***Comentarios*** | Dispone de una versión gratuita que permite la indización de páginas web. |

| | |
|---|---|
| ***Nombre*** | Autonomy |
| ***Productor*** | Autonomy <http://www.autonomy.com/Content/Autonomy/> |
| ***Comentarios*** | Dispone de un sistema de generación de taxonomías muy reputado. |

| | |
|---|---|
| ***Nombre*** | Oracle InterMedia |
| ***Productor*** | Oracle <http://www.oracle.com/intermedia/>. |
| ***Comentarios*** | Perfectamente acoplado con Oracle. |

| | |
|---|---|
| ***Nombre*** | RetrievalWare |
| ***Productor*** | Convera <http:// http://www.convera.com/Products/index.asp> |
| ***Distribuidor*** | OCS < http://www.ocstechnologies.com/pages/productos/convera.html > |
| ***Comentarios*** | Dispone de unas prestaciones de búsqueda semántica y patronal muy potentes. |

| | |
|---|---|
| ***Nombre*** | Search'97 |
| ***Productor*** | Verity <http://www.verity.com> |
| ***Distribuidor*** | Da Vinci Consulting Tecnológico (www.dvc.es) |
| ***Comentarios*** | Muy extendido en el mercado. Dispone de un módulo que permite aplicar taxonomías de forma automática. |

| | |
|---|---|
| ***Nombre*** | ZyIndex – ZyImage |
| ***Productor*** | ZyLab <http://www.zylab.nl> |
| ***Comentarios*** | Dimensiones más reducidas que los anteriores. |

**Bibliografía**


Baeza-Yates, R.; Ribeiro-Neto, B. *Modern information retrieval*. New York: Addison-Wesley, 1999. 513 p

Celma, Matilde; Casamayor, J.C.; Mota, L. *Bases de dades relacionals*. València: Universitat Politècnica de València, 1998. p. 1-20.

Codina, Lluís. "Cómo funcionan los servicios en Internet: un informe especial para navegantes y creadores de información (I)", *El profesional de la información*, vol 7, nº 5, mayo 1997, p. 22-27.

Codina, Lluís. *Sistemes d'informació documental: concepció, anàlisi i disseny de sistemes de gestió documental amb microordinadors*. Barcelona: Pòrtic, 1993. p. 25-34.

Codina, Lluís; Abadal, Ernest. "Gestió documental amb microordinadors: característiques, estructura i tecnologia dels sistemes de gestió documental". *Item*, núm. 11, 1992, p. 72-100.

Connolly, T.M.; et al. *Database systems: a practical approach to design, implementation and management*. Wokingham [etc.]: Addison-Wesley, 1995.

Eíto, Ricardo. "Sistemas GED e indizadores: ¿alternativas excluyentes o tecnologías complementarias?". *El profesional de la información*, vol 7, nº 9, septiembre 1998, p. 5-9.

Ellis, David; Ford, Nigel; Furner, Jonathan. "In search of the unknown user: indexing, hypertext and the world wide web", *Journal of documentation*, January 1998, 54 (1), 28-47.

Fidel, Raya. *Database design for information retrieval: a conceptual approach*. New York [etc.]: John Wiley & Sons, 1987. p. 1-18.

Figuerola, Carlos G.; Alonso, José L.; Zazo, Angel F. "Diseño de un motor de recuperación de la información para uso experimental y educativo", BiD, 4, juny 2000. <http://www.ub.es/bid/04figue.htm>

Frants, Valery I.; Shapiro, J.; Voiskunskii, Vladimir. *Automated information retrieval: theory and methods.* San Diego [etc.]: Academic Press, 1997. Cap 6. Automatic indexing of documents, p. 136-165.

García Figuerola, Carlos. "La recuperación de información en colecciones documentales multilingües". En: Pinto, María; Cordón, José A. *Técnicas documentales aplicadas a la traducción*. Madrid: Síntesis, 1999. p. 129-142

Gil Leiva, Isidoro. *La automatización de la indización de documentos*. Gijón: Trea, 1999. 221 p.

Gillman, Peter (ed.). *Text retrieval: the state of the art*. London: Taylor Graham, 1990, 208 p.

Information systems on-line course. INSY312 Introduction to database design: Lesson 1: Overview of database management systems. Mercer University. http://mumc.mercer.edu/etris/dbless1.htm [Consulta: 25/01/1999]

Kemp, A. *Computer-based knowledge retrieval*. London: Aslib, 1988. 399 p.

Leloup, Catherine. *Motores de búsqueda e indización*. Barcelona: Gestión 2000, 1998. 287 p.

Marcos Mora, Mari Carmen. "Motores de recuperación de la información: un análisis comparativo (parte II)", *El profesional de la información*, vol 7, nº 3, marzo 1998, p. 13-19.

Miguel, Adoración de; Piattini, Mario. *Fundamentos y modelos de bases de datos*. Madrid: RA-MA, 1997.

Moya, Félix de. *Los sistemas integrados de gestión bibliotecaria: estructuras de datos y recuperación de información*. Madrid: Anabad, 1995. p. 113-132.

Moya, Félix. "Técnicas avanzadas de recuperación documental". En: López Yepes, J. *Manual de ciencias de la documentación*. Madrid: Pirámide, 2002.

Muñoz, Jesús E. "Bancos de imágenes: evaluación y análisis de los mecanismos de recuperación de imágenes", *El profesional de la información*, vol. 10, nº 3 (marzo 2001), p. 4-18.

Nieuwenhuysen, P. "Criteria for the evaluation of text storage and retrieval software". *The electronic library*, vol 6, no.3 (june 1980), p. 160-166.

Olvera, Ma. Dolores. "Evaluación de sistemas de recuperación de información: aproximaciones y nuevas tendencias". *El profesional de la información*, noviembre 1999, 8 (11), 4-14.

Peña, Rosalía. *Gestión digital de la información: de bits a bibliotecas digitales y la web*. Madrid: Ra-ma, 2002. p. 105-115.

Rowley, Jennifer. "The controlled versus natural indexing languages debate revisited: a perspective on information retrieval practice and research", *Journal of information science*, 20, 2, 1994, p. 108-119.

Saffady, William. "Text retrieval products for libraries", *Library technology reports*, vol. 36, no. 2 (march-april 2000), p. 7-16.

*Search engine watch*. Danny Sullivan, editor. Internet.com, 1996-1999. <http://searchenginewatch.com>. [Consultat: gener 1999].

Sieverts, E.G. et al. "Software for information storage and retrieval tested, evaluated anc compared: Part 1 – General introduction". *The electronic library*, vol 9, no.3 (1991), p. 145-154.

Sieverts, E.G. et al. “Software for information storage and retrieval tested, evaluated anc compared: Part 2 – Classical retrieval systems”. *The electronic library*, vol 9, no.6 (1991), p. 301-316.

Sieverts, E.G. et al. “Software for information storage and retrieval tested, evaluated anc compared: Part 4 – Indexing and full-text retrieval programs”. *The electronic library*, vol 10, no.4 (1992), p. 195-206.

Soergel, Dagobert. *Organising information principles of data base and retrieval systems.* San Diego [etc.]: Academic Press, 1985. p. 41-51.

Tenopir, Carol. “Full text database retrieval performance”, *Online review*, 1985, vol. 9, No. 2, p. 149-163.

Tenopir, Carol; Lundeen, Gerald. *Managing your information: how to design and create a textual database on your microcomputer*. New York: Neal-Schuman, 1988. 226 p.

Willitts, John. *Database design and construction: an open learning course for students and information managers*. London: Library Association, 1992.

## 4 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 capítulo 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.

Con todo lo que hemos visto anteriormente, podríamos disponer perfectamente de una base de datos bien organizada y explotada con un SGBD determinado. A partir de este momento, se podría iniciar el proceso de distribución analizando cuáles han de ser las vías para que este producto pueda llegar al público deseado o a sus usuarios de la forma más rápida, sencilla y barata posible. 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 pequeños y medianos centros de información y documentación, e incluso de usuarios personales, la posibilidad de convertirse en productores y distribuidores de bases de datos. Estas eran unas funciones que, con la tecnología anterior, eran muy difíciles de desempeñar sin una infraestructura muy especializada y costosa. Por tanto, en el enfoque de nuestro texto, no vamos a perder de vista ese importante cambio y vamos a tratar los distintos apartados pensando especialmente en ellos.

En este capítulo nos centraremos en el análisis de los principales sistemas de distribución de bases de datos que existen en la actualidad profundizando en ellos según su importancia. Así pues, trataremos aspectos relacionados con la consulta local, la edición impresa de referencias y la edición de discos ópticos. Por último, abordaremos con especial detalle la estructura y funcionamiento de la distribución de bases de datos vía web, la fórmula actualmente más utilizada y que dispone de unas ilimitadas perspectivas de futuro. No se pormenorizará, en cambio, en la descripción de la vertiente más comercial de la distribución y, por tanto, no se estudiarán las estrategias de márqueting, ni los planes de comercialización, o los sistemas para el establecimiento de precios y de facturación. Pueden consultarse diversas referencias de Tomàs Baiget (1989 y 1995) para ampliar la información desde este punto de vista.

## 4.1 Sistemas de distribución

A continuación, vamos a describir y valorar someramente los distintos sistemas de distribución de bases de datos que actualmente se utilizan. En un primer bloque vamos a hacer referencia a la consulta local, la edición impresa de bibliografías y la edición óptica, tres formas que se pueden considerar tradicionales, dejando para el siguiente apartado la distribución por medio del web. No se toma en consideración el uso de disquetes ya que, debido a su limitada capacidad, tan sólo se puede utilizar para incluir pequeñas bases de datos que, además han de estar generadas con programas de amplia difusión (p.e. *MS Access*) ya que no pueden incluirse en el disquete.

### 4.1.1 Consulta local

La consulta local es el primer sistema que se utilizó para facilitar el acceso de los usuarios a la base de datos. No tiene ningún secreto. Se trata de permitir a los usuarios que puedan consultar la base de datos en el lugar en el que ésta se ha creado y, por ello, hay que poner a su alcance algún ordenador en el cual se haya cargado previamente la base de datos juntamente con la aplicación que permite gestionarla. En este contexto (muy habitual en pequeñas empresas o en pequeños departamentos) aunque la distribución es en efecto *conceptualmente* distinta de la producción, apenas hay diferencias *prácticas* entre ellas: una vez creada la base de datos, disponemos ya de forma intrínseca de un medio de distribución a la vez.

La ventaja principal de este método de distribución reside en el hecho de que no hay que preparar especialmente ni apenas realizar cambios en el programa de gestión de la base de datos, salvo pequeñas operaciones de administración de usuarios (creación de cuentas de usuarios, administración de passwords) o de configuración de vistas. Los usuarios, por tanto, consultan la base de datos utilizando la misma interfaz que proporciona la aplicación informática utilizada; a lo sumo con una interfaz personalizada con las propias herramientas que proporciona el SGBD.

Los inconvenientes, en cambio, son diversos. En primer lugar, parece claro que los usuarios se han de desplazar expresamente al centro de documentación o unidad donde está disponible la aplicación, con todas las molestias que, a veces, esto puede comportar. Si se opta por un acceso en red para evitar desplazamientos, se requiere de tal red, así como de diversos ordenadores y de un número igual de licencias en red (o para un número determinado de equipos) del programa de gestión documental. Finalmente, en todos los casos es necesario velar por la seguridad y tomar medidas para que los usuarios no puedan modificar la base de datos[5] ni tampoco puedan introducir ni ficheros ni programas en el ordenador.

Es obvio que la consulta local sigue siendo útil pero no se puede olvidar que con este sistema se va a llegar a un número siempre muy reducido de usuarios.

[5] P.e. *Knosys*, *Inmagic* y *CDS/ISIS* permiten, de formas diferentes, que el usuario tan sólo trabaje con las funciones de consulta y sin poder manipular, por tanto, la base de datos.

### 4.1.2 Edición impresa

En ocasiones, una forma de relativamente cómoda de distribuir la información de una base de datos ha consistido en editar su contenido en forma impresa, elaborándose bibliografías impresas, con todos o con una parte, de los registros de la base de datos. Aunque es cierto que se trata de un sistema que va a la baja, aún es habitual encontrar este modo de distribución en centros muy especializados, como el Servicio de Documentación de Historia Local de Cataluña (SDHLC) de la Universidad Autónoma de Barcelona, el Observatorio de la Comunicación Científica (OCC) de la Universidad Pompeu Fabra o el Centro de Recursos de l'Hospitalet de Llobregat. Esto sin contar con toda una seria de guías y repertorios que se distribuyen de forma impresa pero que utilizan una base de datos documental como medio de gestión y de producción (p.e. la mayor parte de los anuarios de los medios de comunicación que se distribuyen en forma impresa se elaboran con el soporte de bases de datos documentales).

Estas bibliografías se pueden elaborar automáticamente desde algunos SGD[6] y constan, básicamente, de dos partes: en primer lugar, un listado 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 listado general. A continuación, mostramos un ejemplo procedente del SDHLC.

[6] *Inmagic*, *CDS/ISIS*, *Pro-Cite* y, en menor medida, *Knosys*, disponen de sistemas para facilitar, más o menos, esta tarea.

[...]

24. Baeza-Yates, Ricardo; Ribeiro-Neto, B. Modern Information Retrieval. Addison-Wesley. Nueva York. 1999. 513 p.
25. Baeza-Yates, Ricardo; Saint-Jean, Felipe. "Análisis de consultas a un buscador y su aplicación a la jerarquización de páginas web'". BiD, núm. 10 (juny 2003) <http://www2.ub.es/bid/consulta_articulos.php?fichero=10baeza.htm> [Consulta: 3/03/2004].

[...]

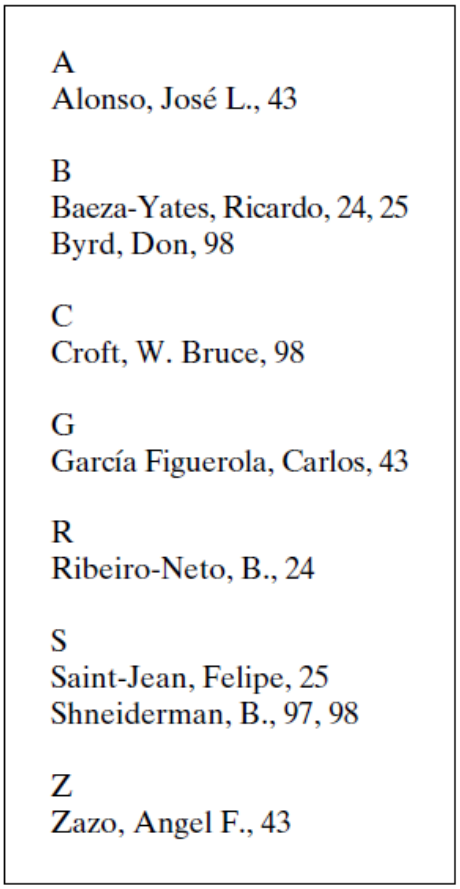
43. García Figuerola, Carlos ; Alonso, José L.; Zazo, Angel F. "Diseño de un motor de recuperación de la información para uso experimental y educativo", BiD, núm. 4, (juny 2000) <http://www.ub.es/bid/04figue.htm> [Consulta: 3/03/2004].

[...]

97. Shneiderman, B. "Dynamic queries for visual information seeking". En: Readings in information visualization: using vision to think. San Francisco: Morgan Kaufmann, 1999. pp. 236-243.
98. Shneiderman, Ben; Byrd, Don; Croft, W. Bruce. "Clarifying search: a user-interface framework for text searches", D-Lib Magazine, (January 1997) <www.dlib.org/dlib/january97/retrieval/01shneiderman.html> [Consulta: 25/09/01].

[...]

***Fig. 4.1. Listado global***



A
Alonso, José L., 43

B
Baeza-Yates, Ricardo, 24, 25
Byrd, Don, 98

C
Croft, W. Bruce, 98

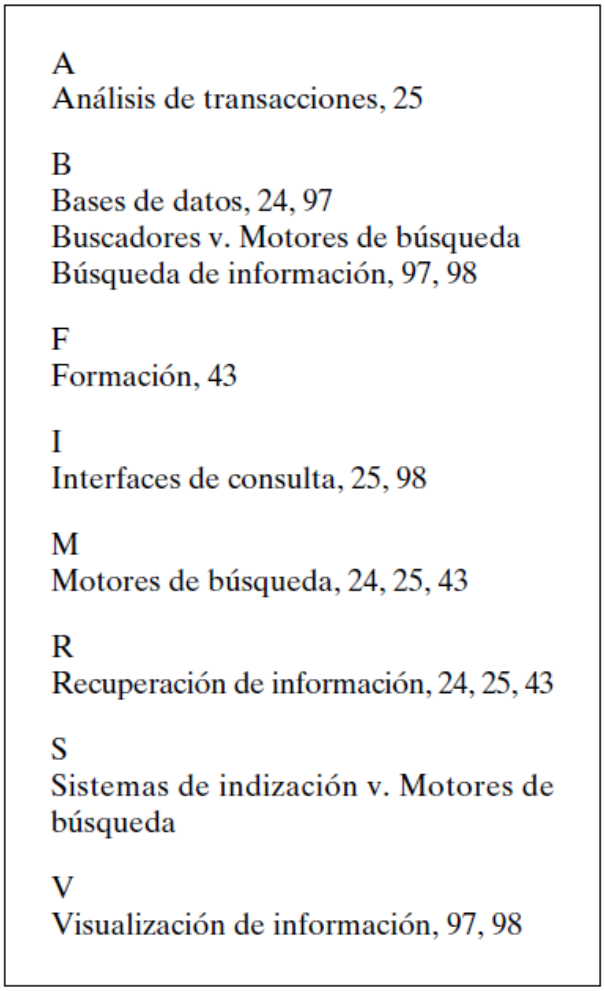
G
García Figuerola, Carlos, 43

R
Ribeiro-Neto, B., 24

S
Saint-Jean, Felipe, 25
Shneiderman, B., 97, 98

Z
Zazo, Angel F., 43

***Fig. 4.2. Índices de autoría***



A
Análisis de transacciones, 25

B
Bases de datos, 24, 97
Buscadores v. Motores de búsqueda
Búsqueda de información, 97, 98

F
Formación, 43

I
Interfaces de consulta, 25, 98

M
Motores de búsqueda, 24, 25, 43

R
Recuperación de información, 24, 25, 43

S
Sistemas de indización v. Motores de búsqueda

V
Visualización de información, 97, 98

***Fig. 4.3. Índices de materia***

Obviamente, la ventaja principal de este sistema reside en que el usuario no necesita equipo ni programa informáticos para poder consultar el contenido de la base de datos ya que toda la información se encuentra impresa y estructurada en la bibliografía. Además, a muchos usuarios les resulta más agradable consultar la información en forma impresa que en el monitor de un ordenador (como demuestra el hecho de que los anuarios de prensa mencionados siguen imprimiéndose en papel en paralelo a su disponibilidad en línea o en cd-rom).

Los inconvenientes más destacables son igual de obvios y coinciden con sendos problemas genéricos de la edición impresa respecto de la edición digital. En primer lugar, este tipo de obras tiene unos costes de impresión y de distribución muy altos no siempre recuperables con la venta ya que el mercado al que se dirigen es muy pequeño y con pocos recursos económicos. En segundo lugar, esta vía presenta notables dificultades para actualizar las obras ya que la edición impresa es poco ágil para facilitar la incorporación y difusión de los nuevos registros que se van incorporando a la base de datos. En general, en esta forma de distribución, la edición es tipo anual. Por último, por supuesto, se pierden las ventajas inherentes al mundo digital como las búsquedas

booleanas: en la edición impresa, con suerte, se dispone de uno o más índices analíticos además de los índices directos.

### 4.1.3 Edición óptica

La distribución de una base de datos en soporte óptico -ya sea cd o dvd- implica incorporar al disco compacto el programa de recuperación de la información (o, al menos, el módulo de consulta, a veces denominado *runtime*) que se ha de utilizar para poder acceder al contenido de la base de datos. La mayoría de los programas de recuperación de la información que se han señalado disponen de versiones sólo para consulta o *runtime* (p.e. *Knosys*, *CDS/ISIS*, etc.) que se pueden utilizar para la edición de la base de datos en soportes ópticos.

Las ventajas principales de este sistema residen, en primer lugar, en las facilidades para acceder a la información, ya que se puede consultar la totalidad de la base de datos utilizando las mismas funciones del programa con el que se ha creado y disponiendo, por tanto, de todas sus prestaciones. También hay que considerar que, actualmente, es posible utilizar fórmulas de autoedición en soportes ópticos, lo que siempre es más barato que la opción industrial.

En lo que respecta a los inconvenientes, se comparten algunos problemas que ya se han comentado anteriormente cuando se ha hecho referencia a la edición impresa: por un lado, las dificultades de actualización y, por otro lado, los costes derivados de la distribución (p.e. por correo postal), así como, a veces, la necesidad de adquirir licencias del programa de recuperación para su distribución en disco, la edición del disco, etc.

Como se puede ver a través de los comentarios anteriores, la valoración que se puede hacer de esta forma de distribución es, en parte, similar a la que se ha realizado de la edición de bibliografías impresas.

Por tanto, no es sorprendete que una de las tendencias que se están utilizando con éxito en la distribución consiste en la combinación de disco óptico con Internet, lo que algunos llaman discos híbridos (cd + línea). Se trata de sistemas que incluyen en el disco no tan sólo el contenido y el programa para permitir su consulta, sino también un pequeño programa que va descargando automáticamente los nuevos registros o nuevas referencias en línea (web o correo-e) y las integra a la colección global. De esta forma, el editor tan sólo tiene que preparar una edición anual de su base de datos, pero el usuario dispone de facilidad de actualización muy frecuentes. Un ejemplo de este sistema lo podemos encontrar en la base de datos *La Ley Actualidad* (Wolters Kluwer), desarrollada con *CD Web Publisher* (Verity). En el caso de obras dirigidas al gran público se pueden encontrar otros ejemplos (p.e. *Encarta* de Microsoft, *Enciclonet* de Micronet, etc.), que también facilitan la actualización mediante la descarga remota de los nuevos registros.

## 4.2 Consulta a través de Internet

Nadie duda que, actualmente, la Web es el sistema de distribución de bases de datos documentales más utilizado y el que cuenta con mejores perspectivas de futuro. 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.

Es decir, 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*, *Netscape*, *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.

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 pasarelas.

El principal y único inconveniente de este sistema de distribución no es otro sino el coste de la adquisición del programa que actúa como pasarela, ya que, por el momento, adquirir un SGDB junto con los programas CGI o similares para la distribución en Web implica pagar unos poco asequibles para las disponibilidades de centros de pequeñas y medianas dimensiones.[7] Esto último no deja de tener su lógica, puesto que adquirir una tecnología que permite consultar la base de datos a través de Internet es equivalente a adquirir en algunos casos una licencia de uso limitada únicamente por la capacidad del servidor web. Hay excepciones, afortunadamente, a esta lógica. En algunos casos, para un pequeño número de accesos concurrentes el SGBD no incrementa su precio (p.e. *FileMaker*).

### 4.2.1 Estructura

Vamos a explicar con un cierto detalle los elementos básicos que intervienen en este proceso y su funcionamiento. Son los siguientes:

- Navegador (P.e. *Explorer, Netscape*, *Mozilla*, etc.).
- Servidor httpd (p.e. *Internet information server*, *Apache*, etc.).
- Programa CGI (p.e. *Knosys Internet*, *WwwIsis*, *WebPublisher*).
- 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.

[7] P.e. *Wwwisis*, *Knosys Internet*, o *Inmagic WebPublisher* son las CGI correspondientes a *CDS/ISIS*, *Knosys Windows* e *Inmagic DB/Text*, respectivamente.

De todas las piezas enumeradas, quizá la menos conocida sea el programa CGI (Common Gateway Interface) que actúa de sistema de comunicación o pasarela[8] 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 (p.e. los términos de búsqueda) y otras correspondientes a parámetros generales (p.e. 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 de la 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, que se describirá con más detalle en el apartado 4.3, 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 (listado); y pantalla de visualización del documento completo.

### 4.2.2 Funcionamiento del proceso

A continuación, vamos a intentar resumir los pasos que se producen en la consulta a una base de datos por medio de un navegador web, desde que el usuario se conecta hasta que recibe una lista de resultados a su petición. De esta manera se puede comprobar cuál es la función de cada uno de los elementos descritos anteriormente.

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

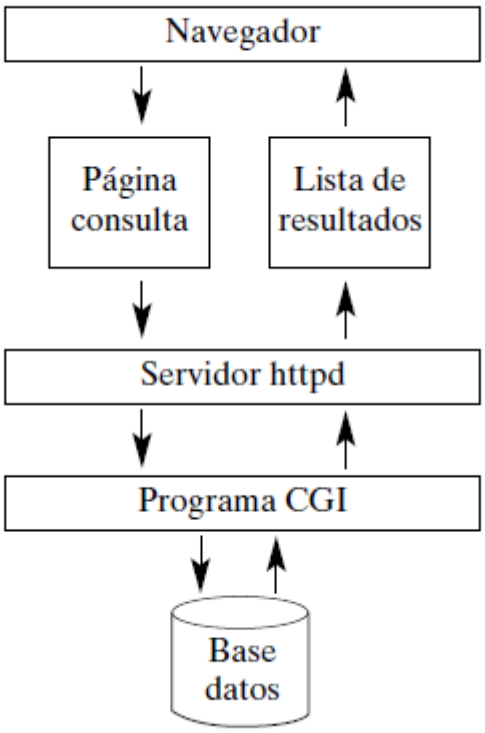



Figura 4.4. Esquema de funcionamiento de la consulta web

– Conexión
El navegador abre una conexión con el servidor httpd en el cual está instalado el programa CGI y la base de datos y se le muestra un formulario de consulta de la base de datos en html (interfaz de consulta).

— Petición del servidor
El usuario introduce los términos de búsqueda y las especificaciones para la visualización (formato, número de registros, orden, etc.) y ordena su ejecución (es decir, hace clic en el botón “Buscar”, o equivalente).

– Transferencia desde el servidor a la CGI
El servidor ejecuta el programa CGI y le transfiere los valores y las variables introducidos por el usuario en el formulario a través de la entrada estándar. (Estas variables se almacenan en un registro “virtual”, cada una de ellas con una etiqueta y un valor).

– Proceso de la petición
El programa CGI interpreta las variables y los valores que ha introducido el usuario y los ejecuta. Es decir, que efectúa una consulta a la base de datos con los términos y especificaciones indicados.

– Resultado
El programa CGI presenta los resultados de acuerdo con las instrucciones del usuario y genera un documento html (página de listado) que se envía al servidor httpd por medio de la salida estándar y éste, a su vez, transmite la información al navegador.

### 4.2.3 Mercado

La mayoría de sistemas de gestión documental disponen de aplicaciones (programas CGI) que permiten que las bases de datos se puedan consultar desde un formulario web. Anteriormente ya hemos avanzado algunos nombres comercializados. A continuación, se van a resumir las principales prestaciones que permiten distinguir entre los programas CGI y se va a presentar también una breve descripción de aquellos que tienen una mayor presencia en el mercado español. Hay que recordar, no obstante, que las CGI no pueden solventar las limitaciones que pueda tener un determinado sistema de gestión documental. Así pues, si *FileMaker* o *Knosys* son dos programas de la gama media y *CDS/ISIS* o *Inmagic* de la gama alta, ésto no se va a poder mejorar por el hecho que las CGI de unos sean mejores que las de otros.

- Asistente
La mayoría de las aplicaciones cuentan con un asistente que permite crear de forma simple una interfaz de consulta básica (es decir, con página de consulta, listado y documento). Este asistente, no obstante, no permite generar aplicaciones complejas para las cuales normalmente será necesario conocer el lenguaje de programación de la CGI y desarrollarlas directamente.

- Mantenimiento de la base de datos
No es frecuente encontrar, en los programas que describimos en este apartado, con la posibilidad de realizar operaciones de mantenimiento de la base de datos (introducir, modificar o borrar registros). *WwwIsis* sería la excepción ya que es posible configurarlo no tan sólo para la consulta de los registros sino también para incorporar nuevos registros de la base de datos, y así facilitar su mantenimiento desde un formulario web.

- Consulta
Entre las prestaciones de consulta que pueden diferenciar un programa de otro se pueden destacar las siguientes: poder consultar más de una base de datos a la vez; poder relacionar bases de datos; poder consultar los índices de todos los campos, etc.

- Visualización
En lo que respecta a este apartado se puede hacer referencia a diversas prestaciones entre las que resaltamos las siguientes: poder establecer uno o diversos criterios de ordenación de los resultados (p.e. fecha, autor, título, etc.); poder escoger el formato de visualización, o el número de registros a visualizar.

- Estadísticas
Algunos de los programas a los que nos vamos a referir llevan incorporada una funcionalidad que facilita la generación de estadísticas o de informes sobre las consultas realizadas a la base de datos. Estos datos pueden ser muy útiles, posteriormente, para contribuir a la evaluación del uso de las bases de datos soportadas.

| | |
|---|---|
| ***Nombre*** | File Maker |
| ***Productor*** | Claris <www.filemaker.com> <www2.filemaker.fr/spain> |
| ***Distribuidor*** | Claris |
| ***Comentarios*** | - Dispone de un asistente que permite elaborar rápidamente la interfaz de consulta. |
| ***Ejemplos*** | BIGPI (Geología de la Península Ibérica): http://www.bib.ub.es/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.<br>- 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.<br>- Se pueden mostrar los índices de campo.<br>- Se pueden relacionar bases de datos. |
| ***Ejemplos*** | Coordinadora Documentació Biomèdica - http://www.doc6.es/cdb |

| | |
|---|---|
| ***Nombre*** | WwwIsis + GenIsis |
| ***Productor*** | Bireme <http://www.bireme.br/wwwisis.htm> <http://perso.wanadoo.fr/pierre.chabert/> |
| ***Distribuidor*** | Bireme |
| ***Comentarios*** | - Dispone de asistente (GenIsis).<br>- Se pueden realizar operaciones de mantenimiento desde el web (entrada, modificación, supresión de registros).<br>- Se pueden relacionar bases de datos.<br>- Se pueden mostrar los índices de campo.<br>- Permite ordenar los registros de acuerdo a distintos criterios. |
| ***Ejemplos*** | Ejemplos de instalaciones en: <http://www.bireme.br/wwwisis/I/listsites.htm> |

### 4.3 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, esta interfaz está formada por un conjunto de formularios (páginas html en el caso de bases de datos distribuidas a través de la Web) de los cuales

podríamos destacar las siguientes: formato de consulta, resultados, visualización del documento completo, información general y ayudas.

El objetivo de este apartado consiste en determinar cuáles son los elementos básicos que han de estar presentes en cada uno de los formularios o de las páginas antes citadas para contribuir a facilitar el proceso de recuperación de la información por parte de los usuarios.

Existe gran cantidad de bibliografía sobre las interfaces de consulta de bases de datos. A modo de presentación, vamos a hacer referencia a dos obras muy citadas y reputadas que servirán para mostrar orientaciones distintas respecto a este ámbito y para situar el enfoque de este apartado.

En primer lugar, vamos a referirnos al famoso manual de usabilidad de Jakob Nielsen (2000), que incluye un apartado dedicado a la interfaz de consulta de bases de datos ("Opciones de búsqueda"). El marco de referencia son los sistemas de recuperación que se pueden encontrar en el web –normalmente, motores de búsqueda a texto completo y con pocas prestaciones para consultar sobre zonas determinadas o campos–, y que son consultados por el gran público. Es por ello que la principal recomendación consiste en solicitar la inclusión de un recuadro de búsqueda y en ocultar y utilizar lo mínimo la búsqueda avanzada, la que permite utilizar el lenguaje de interrogación y combinar términos con la lógica booleana. Como se puede comprobar después de su lectura, el tratamiento y las recomendaciones que se presentan transpiran sencillez y simplicidad.

En segundo lugar, situamos las *Guidelines for OPAC displays*, establecidas para normalizar la visualización de registros bibliográficos procedentes, en especial, de catálogos de biblioteca. Estas directrices se componen de dos partes. En la primera, *"Principles"*, se enumeran treinta principios referentes a la visualización de la información de los registros bibliográficos contenidos en el catálogo. La segunda parte, *"Recommendations"*, se entretiene en precisar recomendaciones específicas respecto de la visualización del contenido de unos determinados campos (autores, obras, materias, clasificación, etc.). En este caso, el nivel de especificación y detalle, así como el tipo de usuario al que se dirigen, se encuentra en las antípodas de lo que se caracteriza en el texto de Nielsen.

Nuestra aproximación va a situarse en medio de ambos polos, y tiene por objetivo precisar cuál podría ser el canon, el modelo teórico, sobre el que se puede fundamentar el diseño de interfaces web de consulta a bases de datos. No llegaremos al nivel de detalle de las *Guidelines* ya que no se hace referencia a las características de la visualización del contenido de los campos, sino que nos centraremos en determinar cuáles han de ser los elementos formales de una interfaz de consulta.

El ámbito de aplicación de nuestra propuesta no se refiere tanto a los catálogos de biblioteca, cuyas funciones están muy normalizadas y sobre los cuales existe una extensa bibliografía al respecto, sino que nos centraremos en las bases de datos documentales en Internet — ya sean bibliográficas o de texto completo— una de cuyas características fundamentales es la integración en un mismo entorno de distintos tipos de objeto, ya sea texto ascii, formatos gráficos (pdf, tiff, jpg, etc.), sonoros o de vídeo.

Dado que el objetivo de este apartado es mostrar un posible canon, en el sentido del uso estándar más aceptado para la consulta de bases de datos a través de la Web, no trataremos experiencias punteras o de laboratorio, sino que, al contrario, pondremos el énfasis en presentar y describir cuáles son los elementos básicos de mayor aceptación en relación a la interfaz de consulta a bases de datos.

### 4.3.1 Qué es una interfaz de consulta

Una interfaz (llamada a veces interfase) es algo que une dos sistemas distintos. 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:[9]

> "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 ya ha sido descrito anteriormente, nuestro objetivo consiste en determinar cuáles son los elementos que han de formar parte de la interfaz de consulta. Parece claro que, para conseguirlo, necesitamos disponer de una definición operativa que nos permita encarar y diseccionar el problema.

Para precisar cuál es la estructura básica de la interfaz de consulta y cuáles son los elementos que han de estar presentes en cada una de sus partes, podríamos basarnos en un esquema que tome como modelo las fases del proceso que sigue el usuario para realizar una consulta[10] o, por el contrario, tener en cuenta la estructura de formularios y páginas que se han de construir desde la aplicación CGI.

---

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

[10] Según este modelo, siguiendo a Marchionini (1995) o a Shneiderman (1997), podríamos descomponer las partes básicas que conforman la interfaz de consulta basándonos en las fases que se establecen en el proceso de consulta a una base de datos, y de cada una de ellas determinar los elementos fundamentales

En nuestro caso, no obstante, vamos a tomar en consideración la segunda de las opciones, y partiremos de los modelos de página que deben elaborarse para que la aplicación funcione de manera adecuada como interfaz de consulta. Se trata de las siguientes:

- Consulta
- Lista de resultados
- Visualización del documento completo
- Páginas de información general
- Páginas de ayuda

La elección de esta opción tiene una relación directa con la perspectiva que tomaremos en consideración. En nuestro caso nos situamos en el lado de quien debe diseñar la interfaz y, por tanto, se describirán los elementos con los que se puede operar para construir una adecuada interfaz a una base de datos.

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.

Para la exposición vamos a seguir la metáfora de la composición de un texto impreso, que está formado por diversas piezas y elementos (una cubierta, portada, sumario, encabezado, paginado, etc.) y que se han de disponer de una determinada manera. De la misma forma que existe un consenso amplio sobre cuáles han de ser los elementos que han de estar presentes en la composición de un texto impreso, deberíamos poder decir lo mismo de la interfaz de consulta. Así pues, se trata de determinar cuáles son los elementos formales que han de formar parte de la interfaz de consulta de una base de datos para que su función se pueda llevar a cabo con las máximas garantías de éxito.[11]

### 4.3.2 Página de 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 que contendrá diversos recuadros de texto para que se puedan introducir los términos de búsqueda, así como también incluirá los operadores de búsqueda presentes en el sistema.

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

---

que han de estar presentes. Así pues, p.e., Marchionini describe con detalle el proceso de búsqueda de información (information seeking process) estableciendo tres grandes subprocesos divididos, a su vez, en etapas: comprensión (definición del problema), planificación (selección de un sistema de búsqueda, formulación de una pregunta, ejecución de la búsqueda), evaluación y uso (análisis de los resultados, extracción de información, repetición, iteración o finalización). Shneiderman, por su parte, contempla las siguientes fases, a cada una de las cuales asocia unas determinadas funciones: formulación (colecciones o bases de datos, campos, términos, variantes), ejecución, revisión de resultados, y gestión de resultados (refinar consulta, enviar por correo).

[11] En principio, no se intentará hacer un ejercicio de jerarquización de estos elementos como, p.e., hace Luisa Sabin-Kildiss (2001) en su estudio.

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.

- *Consulta simple o básica*

Está concebida para la mayoría de usuarios que, normalmente, buscan información sobre un concepto, sin combinarlo con ningún otro, y que no saben precisar en qué campo concreto se encuentra (si se trata de un autor, una materia, o un título). Debe caracterizarse por su simplicidad y, en general, dispondrá de un único recuadro de texto en el que se podrá introducir el término (o términos) de consulta que el sistema buscará en cualquiera de los campos, si los hubiera, de la base de datos.[12]

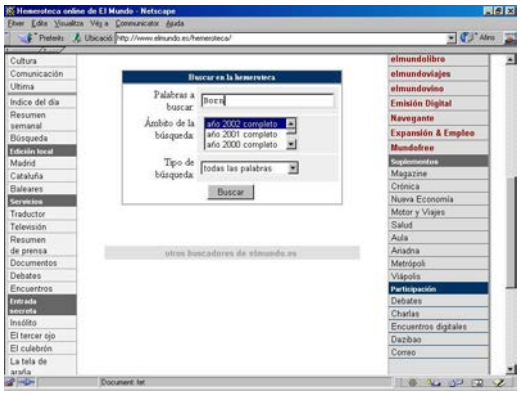



***Fig. 4.5. Ejemplo de página de consulta simple***

- *Consulta avanzada o combinada*

Está pensada para los usuarios que necesitan realizar consultas que combinen más de un término, que pueden estar en campos distintos y que utilizan diversos operadores booleanos. Está página debe incluir diversas líneas con recuadros de texto para que se puedan combinar los términos y, además, se pueda escoger el operador booleano apropiado.

Este tipo de consulta es siempre aconsejable, pero no siempre es necesario que aparezca en primer lugar, al menos no se aconseja que aparezca en primer lugar en sedes web de carácter general, ya que pueden confundir al usuario. Según los estudios de Nielsen (2000), a los usuarios les es difícil diferenciar y aplicar correctamente los operadores booleanos. Es por ello que recomienda que la página de búsqueda que permite utilizarlos quede totalmente separada de la búsqueda simple, y que no se invite especialmente a su utilización:

[12] En algunos casos, dependiendo de las características de los usuarios que van a consultar la base de datos, puede contemplarse la posibilidad de acotar la búsqueda a un campo determinado.

"Es importante usar un nombre intimidatorio como 'búsqueda avanzada' para desanimar a los usuarios principiantes a que lleguen a la página y puedan hacerse daño a sí mismos." (Nielsen, 2000: 227)


***Fig. 4.6. Ejemplo de página de consulta avanzada***

*- Consulta de índices*

Además de la estructura básica a la que antes se ha aludido, una buena interfaz ha de permitir también la consulta a los índices de los campos, especialmente los de autores, títulos y materias (temas).

Los índices se desplegarán a petición del usuario, es decir, cuando haga clic en el botón correspondiente. Su objetivo consiste en ayudar al usuario a afinar su consulta.

La estructura de una página de índices será, más o menos, la siguiente:

| |
|---|
| Número o recuadro de selección - Término - Núm. documentos asociados |

También debe contemplarse la opción de volver a la página de consulta o aún mejor la posibilidad de ejecutar desde aquí la consulta con las condiciones que se hayan especificado.

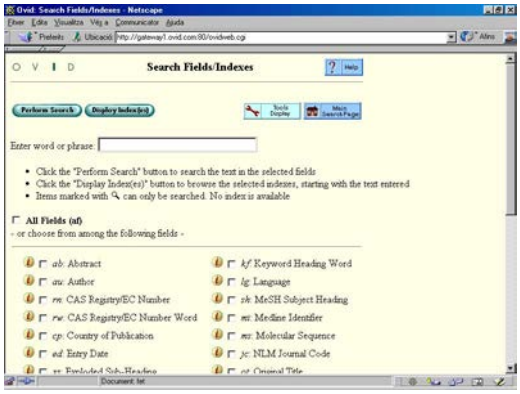



***Fig. 4.7. Estructura de un índice***

Los índices de campo contienen información muy valiosa para ayudar a que el usuario clarifique su consulta. Existe un problema técnico para facilitar su visualización ya que son muy voluminosos (p.e. título, materia, etc.) y permitir su consulta, sin restricción, implica enviar miles de términos a la memoria temporal del ordenador cliente con las dificultades que ello conlleva. Para solventar este problema, algunas aplicaciones lo envían fragmentado (piden al usuario que indique a partir de qué término quiere consultar).

Esta forma de acceso normalmente se combina con la consulta simple y/o la avanzada.

- *Acceso por categoría temática*

Poder consultar los registros a partir de una categoría temática constituye un instrumento muy útil de recuperación de la información. Son diversos los estudios que señalan las ventajas de este sistema (Lim, 1999; Vizine-Goetz, 1996): simplifica la consulta a los usuarios poco experimentados; permite ampliar o limitar las consultas; sitúa las materias dentro de un contexto, etc.

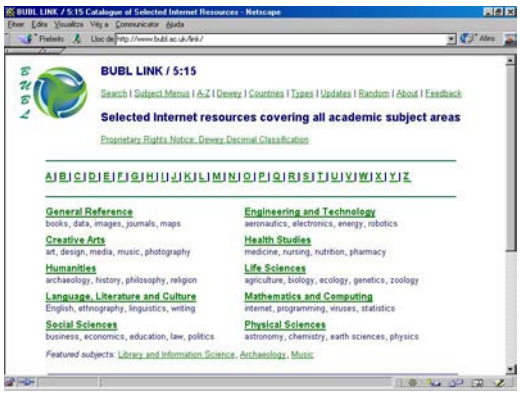



***Fig. 4.8. Ejemplo de categoría temática***

#### 4.3.2.1 Elementos básicos

Por otra parte, los elementos y prestaciones fundamentales que han de estar presentes en una página de consulta son los siguientes:

- *Identificación de la página o base de datos*
Hay que incluir un título que indique cuál es la página ("consulta simple" o "consulta avanzada") o base de datos en la que se encuentra el usuario. Su presencia facilita especialmente la orientación.

- *Especificación de la base de datos (o del fondo o colección o subsitio).*
El usuario ha de saber, con mucha claridad, cuál es la base de datos o fondo sobre el que está buscando. En algunos casos, cuando se consulta un gran distribuidor de bases de datos (p.e. *Dialog, Lexis-Nexis,* etc.) ésto constituye un aspecto fundamental. Si es posible, se ha de poder señalar más de una colección para poder realizar búsquedas simultáneas. Pensando siempre en el gran público, Nielsen (2000: 225) insiste mucho en dejar bien claro el ámbito sobre el cual se está realizando la búsqueda, y desaconseja la división de un sitio web en áreas especializadas (equivalentes a fondos, colecciones o bases de datos homogéneas entre sí y distintas de las demás) porque a los usuarios les cuesta entender la estructura y no saben si están buscando en todas ellas o tan sólo en alguna en concreto.

- *Sistema de recogida de información del usuario.*
Para facilitar que el usuario pueda indicar los términos de consulta se puede recurrir a cualquiera de los recursos básicos para recopilar información de un usuario en un formulario web, es decir, los cuadros de entrada de texto, botones de radio, casillas de verificación, listas desplegables, etc.

- *Acotación de la búsqueda a un campo o conjunto de campos.*
Cuando el usuario introduce el término de búsqueda tiene que saber si se está realizando una búsqueda global (a todos los campos) o, por el contrario, está limitada a un campo determinado. Cuando se trata de este último caso, hay que precisar cuáles son los campos que se han de destacar (serán los que tienen mayor interés para el usuario).

- *Utilización de los operadores booleanos (y de otros operadores)*
Puede utilizarse una ventana del formulario para permitir que los usuarios avanzados que conocen la sintaxis booleana puedan expresar búsquedas complejas, pero también es importante disponer una serie de ventanas con iconos asociados con cada uno de los operadores de manera que, a usuarios que estén inseguros con la sintaxis puedan limitarse a desplegar y seleccionar operadores booleanos como opciones.

- *Visualización de los índices. (v. 4.3.2)*
Recordemos en este sentido que, al describir la consulta a partir de índices ya se ha indicado que acostumbran a presentarse como un elemento más de la página de consulta.

- *Informaciones breves para ayudar en la consulta.*
Independientemente de la existencia de las páginas de ayuda, puede ser muy útil disponer de breves mensajes que sirvan para orientar al usuario sobre cómo tiene que

introducir los términos (p.e. los autores invertidos, las fechas en formato aaaa/mm/dd, etc.), u otros aspectos de interés y clarificadores.

- *Elección de la forma de presentación de los resultados:*
  - Formato de visualización del listado
  Especificar qué tipo de formato se desea para el listado (breve, extenso, con tabla, sin tabla, etc.).
  - Número de registros a visualizar
  Determinar la cantidad de registros que se desea que se incluyan en la página de listado (ya sean grupos de 5, 10, 20 o n referencias).

- *Elección del sistema de ordenación de los resultados.*
El criterio puede ser alfabético (ordenación por autor, título o cualquier otro campo) por fecha, por ponderación (es decir, ordenación según precisión, colocando en primer lugar los registros que más se ajustan a la petición del usuario), etc. El usuario puede tener la opción de escoger más de un criterio (p.e. fecha y autor) o, si no, el sistema automáticamente aplica un segundo y tercer criterios, si es necesario. En bases de datos dirigidas al gran público se recomienda configurar por defecto la opción de ordenación por ponderación.

- *Botones para la ejecución de acciones.*
La acción principal a realizar es la de buscar, y el botón que permite ponerla en práctica ha de estar destacado. Por otro lado, también puede encontrarse un botón para borrar (cancelar).

- *Registro de las búsquedas realizadas (historial).*
En muchos casos, poder releer el historial de búsqueda sirve de ayuda para la elaboración de la estrategia de búsqueda.

- *Acceso multilingüe.*
Posibilidad de escoger el idioma de la interfaz de consulta.

- *Navegación entre páginas de la interfaz.*
Desde la página de consulta ha de poderse desplazar con facilidad a las otras páginas de consulta, a las ayudas, a la descripción de las bases de datos, etc.

- *Datos identificativos (productor, fecha, lugar, etc.).*
Es recomendable incluir las menciones de responsabilidad y de identificación de la interfaz de consulta.

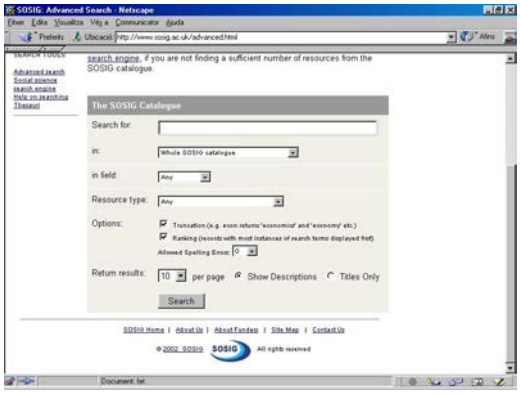



***Fig. 4.9. Ejemplo de página de consulta***

Como hemos detallado, 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 del listado de resultados: cuántos registros se presentarán, en qué formato, en qué orden aparecerán, etc.

### 4.3.3 Página 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 del 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 que utilizan metáforas visuales más o menos desarrolladas; estas visualizaciones pueden ser de carácter bidimensional o tridimensional (Moya, 1999; Shneiderman, 1999). Este último sería el caso de KartOO (www.kartoo.com), el metabuscador francés que presenta los resultados en una especie de mapa a base de esferas que jerarquiza e interrelaciona los resultados.

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 precisión (ponderación), la relevancia u otros criterios libremente configurables por el usuario (fecha, autor, etc.).


*Fig. 4.10. Ejemplo de página de resultados (Ovid)*


*Fig. 4.11. Ejemplo de página de resultados (El Mundo)*


*Fig. 4.12. Ejemplo de página de resultados (Bireme)*


***Fig. 4.13. Ejemplo de página de resultados (Kartoo)***

La unidad de información del listado de resultados acostumbra a ser el documento unitario o el registro, cuando buscamos en bases de datos bien estructuradas, aunque a veces también puede ser la página o la sede web, especialmente cuando se trata de buscadores de carácter general.

#### 4.3.3.1 Elementos básicos

La página de resultados ha de presentar, como mínimo, los siguientes elementos y funciones:

- *Identificación de la página o base de datos*
Hay que recordar el nombre de la base de datos, en el caso que exista la posibilidad de consultar más de una.

- *Información sobre el término de búsqueda y los resultados obtenidos.*
Es importante incluir un breve mensaje que recuerde cuál ha sido la consulta, cuántos son los documentos que se ajustan a ella y el número de ellos que se están visualizando.

- *Lista con la descripción básica de los documentos.*
Se trata de proporcionar la información de los campos básicos para poder facilitar la selección de los documentos que pueden ser de mayor interés.
Aspectos que pueden variar de un listado a otro:

- *Estructura*
Tal y como decíamos anteriormente, el listado puede tener estructuras diversas. En registros textuales, la presencia, o no, de tablas es lo que puede condicionar más su formato. En registros gráficos acostumbra a incluirse la posibilidad de consulta panorámica (p.e. galería de fotos).
- *Inclusión del nombre del campo*
En algunos casos, no será necesario indicar el nombre del campo, ya que se puede identificar con facilidad su contenido sin necesidad de incluir su denominación.
- *Casilla de selección*
Se ha de incluir una casilla para poder seleccionar aquellos documentos del listado que sean de interés, y así poder elaborar un subconjunto propio.

- *Indicación de tipo de documento (objeto).*
Hay que prever los casos en los que en la base de datos se encuentran documentos de distintos tipos (fotos, vídeos, textos, etc.). El listado de resultados ha de disponer de algún recurso visual para diferenciarlos.

- *Agrupación de resultados por categorías.*
Imprescindible en los casos en que es posible el acceso por categorías temáticas.

- *Información sobre errores o ausencia de resultados.*
Es fundamental proporcionar mensajes informativos para comunicar al usuario las posibles incidencias que se puedan producir en el proceso de consulta a la base de datos.

- *Opciones de gestión de los registros o documentos.*
Algunos sistemas facilitan la realización de operaciones diversas: grabar los resultados, imprimir los registros, enviarlos por correo electrónico, etc. Estas opciones han de ser posibles no tan sólo para la lista global sino también para el subconjunto de registros marcados por el usuario.

- *Elección de la forma de presentación de los resultados* (v. 4.3.2.1)
    - Formato de visualización
    - Número de registros a visualizar

- *Elección del sistema de ordenación de los resultados.* (v. 4.3.2.1)

- *Reformulación de la búsqueda.*
Se ha de permitir realizar búsquedas sucesivas —ampliar o restringir las condiciones de búsqueda— sobre el conjunto de registros que se ha encontrado.

- *Encontrar documentos similares.*
Esta opción aparece especialmente en bases de datos de texto completo y permite recuperar documentos que tienen similitudes de contenido con cualquiera de los que aparecen en la lista.

- *Navegación entre registros de la base de datos.*
Se trata de que estén activos como enlaces el contenido de algunos campos básicos (autor, materia, etc.) para facilitar la navegación directa a un autor, materia, etc. desde cualquier página de listado.

- *Avance y retroceso en las páginas del listado.*

- *Navegación entre páginas de la interfaz* (v.4.3.2.1).

### 4.3.4 Visualización de los documentos

Desde la página de listado se ha de 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 y, además, se puede pedir la visualización de la referencia o de los metadatos.

Es recomendable que la página de visualización de documentos contenga los siguientes elementos y opciones:[13]

- Identificación de la página o base de datos.

- Indicación del número de registro que se está visualizando.
Dentro del conjunto recuperado.

- Opción de cambio de formato de visualización.

- Resaltar los términos de búsqueda.
Esta opción no siempre es posible en la lista de resultados ya que, en general, se visualiza un número reducido de campos, los cuales no siempre contendrán los términos de búsqueda.

- Distintas resoluciones.
Para el caso de documentos gráficos (fotografías, especialmente) es importante poder disponer de la opción de descargar el documento con distintos grados de resolución.

- Navegación entre registros de la base de datos (v. 4.3.3.1).

- Avance y retroceso entre los registros seleccionados.

- Navegación entre páginas de la interfaz (v.4.3.3.1.1).

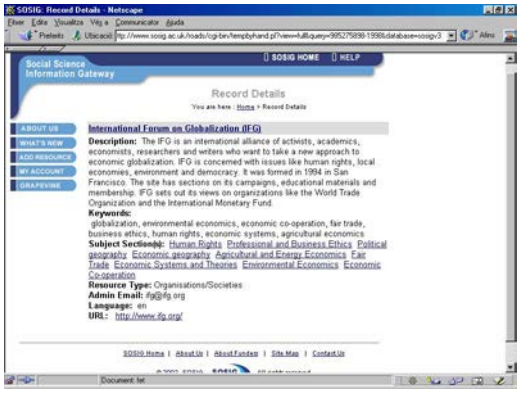



***Fig. 4.14. Ejemplo de página de documento (Sosig)***

[13] Una buena parte de ellos ya ha sido descrita en el apartado anterior y, por tanto, no volverán a ser tratados.


***Fig. 4.15. Ejemplo de página de documento (El Mundo)***

### 4.3.5 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:

—*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 se proporcionan han de permitir contextualizar el contenido de la base de datos, así como mostrar su alcance.

—*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.[14]

—*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.

### 4.3.6 Conclusiones

"*Beyond boolean*" es un artículo escrito hace en la década de los ochenta (Hildreth, 1987) que describía las limitaciones de los catálogos en línea de la época e indicaba un amplio conjunto de recomendaciones sobre cómo deberían que ser los OPACS del futuro —"la tercera generación", en palabras suyas. El texto recopila un amplio abanico

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

de ideas para ayudar a incrementar las prestaciones de usabilidad de los catálogos en línea. Se formulaba en este artículo un principio básico que aún hoy es vigente:

> "Un catálogo en línea de acceso público debe trabajar de modo inteligente con el usuario, implicándose en un diálogo con sentido para extraer expresiones de las necesidades de información del usuario (que pueden variar durante el transcurso de la búsqueda) y para mejorar los resultados de la actividad de búsqueda del usuario. Algunos corolarios de este principio […]: Nunca asuma que el usuario podrá navegar de forma efectiva a través de una base de datos cada vez más compleja, presentada con opciones de recuperación cada vez más sofisticadas, sin una generosa ayuda y asistencia del sistema en línea." (Hildreth, 1987: 665)

Respetando este principio teórico con el cual es fácil estar de acuerdo, nuestro estudio nos lleva a destacar un conjunto de elementos que constituyen el canon actual de las interfaces de consulta a bases de datos documentales, y que se presentan de forma resumida en la tabla siguiente.

**Tabla 4.1. Elementos de una interfaz de consulta a bases de datos**

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

| | |
|---|---|
| | - Elección de la forma de presentación de los resultados<br>  - Formato de visualización<br>  - Número de registros a visualizar<br>- Información sobre errores o ausencia de resultados.<br>- Opciones de gestión de los registros o documentos.<br>- Elección del sistema de ordenación de los resultados.<br>- Reformulación de la búsqueda.<br>- Encontrar documentos similares.<br>- Navegación entre registros de la base de datos.<br>- Avance y retroceso en las páginas del listado.<br>- Navegación entre páginas de la interfaz. |
| Documento | - Identificación de la página o base de datos<br>- Indicación del número de registro que se está visualizando.<br>- Opción de cambio de formato de visualización<br>- Resaltar los términos de búsqueda.<br>- Distintas resoluciones.<br>- Navegación entre registros de la base de datos.<br>- Avance y retroceso entre los registros seleccionados.<br>- Navegación entre páginas de la interfaz. |

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. Así pues, la evaluación de una interfaz puede basarse en la presencia (inclusión) de los elementos antes reseñados.

En cualquier caso, hay que tener presente dos cuestiones adicionales:

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

- *Universalidad*
Por otro lado, 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 han de presentar ya configurados, sin dejarles libertad de elección para personalizarlos.

Para finalizar, vamos a presentar una reflexión sobre cómo se puede complementar el canon que se ha descrito a fin de simplificar aún más el diálogo, la relación, entre el usuario y la base de datos.

En un artículo de José Vicente Rodríguez y Tomás Saorín (Rodríguez, 1998) se caracterizan dos modelos de interfaces de consulta que los autores denominan respectivamente: “OPAC-formulario” –el cual tiene por objetivo fundamental

"recuperar"– y "OPAC-guía" –que persigue "mostrar" el contenido de las bases de datos.

En el primer caso, se trata de consultar la información introduciendo los datos en un formulario genérico que normalmente es opaco al contenido de la base de datos, que viene a ser una caja negra que tan sólo deja ver sus contenidos cuando existe un requerimiento preciso y formalizado. El objetivo fundamental de este modelo, el más extendido actualmente y que se ha descrito en el texto precedente, es "recuperar". En el segundo caso, se trata de interfaces que tienen por objetivo "mostrar" el contenido de las bases de datos, preparando los contenidos para facilitar su difusión. Se trata de elaborar consultas ya preparadas que incidan en temas de especial interés para el usuario a fin que, sin necesidad de preguntar nada en concreto, pueda obtener la información que le interesa.

Pongamos un ejemplo. Supongamos que tenemos que preparar la interfaz de consulta del fondo histórico de un periódico. Podríamos construir un conjunto de páginas a base de formularios o, por el contrario, y de forma complementaria, disponer de una guía temática ya elaborada que se centrara en unos determinados temas de interés (p.e. atentados durante la II República, el 23-F, la coronación del Rey, etc.) que irían variando en función de los intereses de la actualidad y que consistieran en unas consultas estándar bajo formulario pero que permanecen ocultas para el usuario.

Este modelo no constituye ninguna novedad. Hace muchos años que las bibliotecas elaboran guías de lectura para "mostrar", dar a conocer, los fondos de los que disponen basándose en un tema o autor de actualidad. De esta forma se sugieren temas de interés a los usuarios, que no hace falta que pasen por el catálogo para conocer el fondo de la biblioteca. Este tipo de prestaciones dibujan un modelo de interfaz de consulta que es un tanto laborioso de preparar pero que puede complementar a la perfección el canon clásico, en especial cuando va dirigido a un público amplio y no especializado.

## Bibliografía


Abadal, Ernest. "Elementos para la evaluación de interfaces de consulta de bases de datos". *El profesional de la información*, vol. 11, núm. 5 (septiembre-octubre 2002), p. 349-360.

Abadal, E.; Criach, D.; Cuadrado, M; Gascón, J.; Omella, E. "Centres de Documentació i Biblioteques de Sabadell en xarxa: una iniciativa per a incrementar els serveis de la biblioteca pública al municipi". Ernest Abadal; [et al.] En: *7es. Jornades Catalanes de Documentació*. Barcelona: Col·legi Oficial de Bibliotecaris-Documentalistes de Catalunya, 1999. p. 127-135.

Abadal, E.; Cuadrado, M; Gascón, J.; Omella, E. "Disseny i creació de bases de dades bibliogràfiques amb CDS/ISIS: l'experiència de SABA-DOC". *BiD: textos universitaris de biblioteconomia i documentació*. Núm 3 (desembre 1999). <http://www.ub.es/bid/03abadal.htm>

Abadal, Ernest; Martínez, Raúl. "Distribució de bases de dades en el web amb *Knosys Internet*". *BiD: textos universitaris de biblioteconomia i documentació*, núm 4 (juny 2000). <http://www.ub.es/bid/04abadal.htm>

Ahlberg, C.; Shneiderman, B. "Visual information seeking: tight coupling of dynamic query filters with starfield displays". En: *Readings in information visualization:*

*using vision to think*. Ed. by S.K. Card, J.D. Mackinlay, B. Shneiderman. San Francisco: Morgan Kaufmann, 1999. p. 244-250.

Altuna Esteibar, Belén. "Comportamientos de uso y estrategias de búsqueda de los usuarios de catálogos automatizados: breve revisión de la investigación". En: *Miscelánea homenaje a Luis García Ejarque*. Madrid: Fesabid, 1992. p. 103-111

Arant, Wendi; Payne, Leila. "The common user interface in academic libraries: myth or reality", *Library Hi Tech*, Vol. 19, No. 1 (2001), p. 63-76.

Archuby, Gustavo. *Wwwisis: manual de procedimientos para bibliotecarios*. [fitxer electrònic]. Versión 0. La Plata, septiembre 1998. 39 p.

Asensi, Viviana; Pastor, Juan Antonio. "Propuesta de un modelo de interfaz genérica para sistemas de recuperación de información". *Scire*, vol. 4, nº 1 (ene-jun. 1998), p. 71-88.

Ashenfelter, J.P. *Choosing a database for your web site*. New York [etc.]: John Wiley & Sons, 1999. 443 p.

BAIGET, T. "La distribució de bases de dades a Espanya". En: *3es Jornades Catalanes de Documentació*. Barcelona: SOCADI; COBDC, 1989. p. 101-141.

BAIGET, Tomàs. "25 años de teledocumentación en España", *Revista Española de Documentación Científica*, vol. 21, núm. 4 (1998), p.373-387.

Bechini, Mònica; Burguillos, Ferran; Díaz, Albert. "Confección de categorías y recuperación de la información en Internet". En: Congreso ISKO-España (5º: 2001: Alcalá de Henares). *La representación y organización del conocimiento* [CD-ROM] *: metodologías, modelos y aplicaciones : actas del V Congreso ISKO-España: 25-27 de abril de 2001, Alcalá de Henares (Madrid).* Alcalá de Henares: Sociedad Internacional para la Organización del Conocimiento, Capítulo Español: Facultad de Documentación, Universidad de Alcalá, 2001. [p.404-414].

Beumala, Àngel et al. "Base de datos de recursos Internet científico-técnicos: Ep! (enlaces politécnicos). En: Jornadas Españolas de Documentación (6as: València, 29-21 octubre 1998). *6as Jornadas Españolas de Documentación: los sistemas de información al servicio de la sociedad: actas de las jornadas*. València: Fesabid; Avei, 1998. p. 149-156.

Brisaboa, Nieves R., et al. "Sistema de consulta vía web para el Instituto Andaluz de Patrimonio Histórico". En: Jornadas de Bibliotecas Digitales (2ª: Almagro, 2001). *JBIDI'2001: Jornadas de Bibliotecas Digitales*. [Ciudad Real]: Universidad de Castilla La Mancha, 2001. p. 99-116.

Brusilovsky, P. "Methods and techniques of adaptive hypermedia", *User Modeling and User Adapted Interaction. Special issue on adaptive hypertext and hypermedia*, Pittsburgh, 1996.

Card, S.K.; Mackinlay, J.D.; Shneiderman, B. *Readings in information visualization: using vision to think*. San Francisco: Morgan Kaufmann, 1999.

Choo, Chun Wei; Detlor, Brian; Turnbull, Don. "Information seeking on the web: an integrated model of browsing and searching", *First Monday*, Vol. 5, no. 2 (February 2000). <firstmonday.org/issues/issue5_2/choo/index.html>. [Consulta: 25/09/01]

Choo, Chun Wei; Detlor, Brian; Turnbull, Don. *Web work: information seeking and knowledege work on the world wide web*. Dordrecht [etc.]: Kluwer Academic, 2000. "Chapter 5 Models of information seeking on the world wide web", p. 133-158.

Chowdhury, G.G.; Chowdury Sudatta. *Introduction to digital libraries*. London : Facet, 2003. Chap. 8: Information access and user interfaces, p. 152-177.

Codina, Lluís. "Evaluación de recursos digitales en línea: conceptos, indicadores y métodos", *Revista española de documentación científica*, vol. 23, nº 1, (enero-marzo 2000), p. 9-44.

Codina, Lluís. "Parámetros e indicadores de calidad para la evaluación de recursos digitales". En: *VII Jornadas Españolas de Documentación. La gestión del conocimiento: retos y soluciones de los profesionales de la información*. Bilbao: Universidad del País Vasco, 2000. p. 135-144.

Cooper, Michael D. "Design considerations in instrumenting and monitoring web-based information retrieval systems", *Journal of the ASIS*, vol. 49, no. 10 (1998), p. 903-919.

Crestani, Fabio, Funte, P. de ; Vegas, J**.** "Diseño de una interfaz de consulta para la recuperación de documentos estructurados". En: Jornadas de Bibliotecas Digitales (2ª: Almagro, 2001). *JBIDI'2001: Jornadas de Bibliotecas Digitales*. [Ciudad Real]: Universidad de Castilla La Mancha, 2001. p. 85-97.

De Groote, Sandy. "PubMed, Internet Grateful Med, and Ovid: a comparison of three Medline Internet interfaces", *Medical reference services quarterly*, vol. 19, no. 4, winter 2000, p. 1-13.

*DESIRE information gateways handbook* [en línea]. DESIRE, c1999-2000, last updated 26 April 00. <http://www.desire.org/handbook/>. [Consulta: 27/04/2001].

Espelt, Constança. "Improving subject retrieval: user-friendly interfaces and effectiveness", *BiD: textos universitaris de biblioteconomia i documentació*, núm 1 (juny 1998). <http://www.ub.es/bid/01espel1.htm> [Consulta: 27/09/01]

Fernández, Mª Jesús; Angós, José Mª; Salvador, José A. "Interfaces de usuario: diseño de la visualización de la información como medio para mejorar la gestión del conocimiento y los resultados obtenidos por el usuario". En: Congreso ISKO-España (5º: 2001: Alcalá de Henares). *La representación y organización del conocimiento* [CD-ROM]*: metodologías, modelos y aplicaciones: actas del V Congreso ISKO-España: 25-27 de abril de 2001, Alcalá de Henares (Madrid).* Alcalá de Henares: Sociedad Internacional para la Organización del Conocimiento, Capítulo Español: Facultad de Documentación, Universidad de Alcalá, 2001. [p.506-517].

Fox, Edward A. et al. "Users, user interface, and objects: Envision, a digital library". *JASIS*, vol. 44 no. 8 (1993), p. 480-491.

Frías Montoya, José Antonio; Martín Rodríguez, Fernando**.** "El análisis transaccional como técnica de recogida de datos para el estudio del comportamiento de los usuarios del catálogo en línea", Congreso ISKO-España EOCONSID'99 (4º. 1999. Granada). *Actas de las VI Congreso ISKO-España EOCONSID'99 : Representación y Organización del Conocimiento en sus distintas perspectivas : su influencia en la recuperación de información*, p. 427-434.

García Marco, Francisco J. "De la consulta de catálogos a la gestión de información: tensiones hacia el cambio en el diseño de OPACS", *Boletín de la ANABAD* (1991), p. 325-334.

García Marco, Francisco J. "Interfaces amigables para la recuperación de la información bibliográfica", *Scire*, vol. 1, nº 1 (ener.-jun. 1995), p. 127-148.

*Guidelines for OPAC displays*. Prepared for the IFLA Task Force on Guidelines for

Gutiérrez de Mesa, José A.; Hilera, José R. "Generación de documentación hipermedia en Internet a partir de información multimedia en bases de datos", *Cuadernos de documentación multimedia*, núms.6-7, 1997-1998, p. 135-140.

Head, Alison J. *Design wise: a guide for evaluating the interface design of information resources.* Medford: CiberAge Books, 1999. 196 p.

Hearst, Marti A. "Sistemas para consultar la red", *Investigación y ciencia* (mayo 1997), p. 44-49.

Hearst, Marti A. “User interfaces and visualization”. En: Baeza-Yates, Ricardo; Ribeiro-Neto, B. *Modern information retrieval*. New York: ACM; Harlow: Addison-Wesley, 1999. p. 257-323.

Herrero, Víctor. "La conexión con bases de datos Microisis a través del World Wide Web", *Bol. Anabad*, 2, abril-junio 1998, p. 309-316.

Hildreth, Charles R. “Beyond boolean: designing the next generation of online catalogs”, *Library trends*, (Spring 1987), p. 647- 667.

Joint, Nicholas. “Designing interfaces for distributed electronic collections: the lesons of traditional librarianship”. *Libri*, vol 51 (2001), p. 148-156.

Lim, Edward. “Pasarelas temáticas del Sudeste Asiático: análisis de sus métodos de clasificación [en línia]”. En: IFLA Council and General Conference (65a: Bangkok: 1999). *Conference proceedings.* <http://ifla.inist.fr/IV/ifla65/papers/011-117s.htm>. [Consulta: 16/07/00].

Marchionini, Gary. *Information seeking in electronic environments*. Cambridge: Cambridge University, 1995. 224 p.

Marchionini, Gary; Komlodi, A. “Design of interfaces for information seeking”, *Annual review of information science and technology*, Vol. 33, (1998), p. 89-130.

Marchionini, Gary; Plaisant, C.; Komlodi, A. “Interfaces and tools for the Library of Congress National Digital Library Program”, *Information processing & management*, Vol. 34, No. 5 (1998), p. 535-555.

Marcos, Mari-Carmen. “HCI (human computer interaction): concepto y desarrollo”, *El profesional de la información*, vol. 10, nº 6 (junio 2001), p. 4-16.

Marcos, Mari-Carmen. “Interacción persona-ordenador en las interfaces de rcuperación de información”. En: Jornadas Españolas de Documentación (8as: Barcelona, 6-8 febrero 2003). *Fesabid 2003: los sistemas de información en las organizaciones: eficacia y transparencia*. Barcelona: Fesabid, 2003. p. 463-476.

Martínez, Victoria. "Un modelo para el uso de internet en los centros de información juvenil", *El profesional de la información*, vol 8, nº 7-8, julio-agosto 1999, p. 22-33.

Moya, Félix; Herrero, Víctor. “Investigaciones en curso sobre interfaces gráficos en dos y tres dimensiones para el acceso a la información electrónica”, *Cuadernos de documentación multimedia*, nº 8, (1999) <www.ucm.es/info/multidoc/multidoc/revista/num8/moya.html>. [Consulta: 25/09/01]

Nackerud, Shane A. “The potential of CGI: using pre-built CGI scripts to make interactive web pages”, *Information technology and libraries*, December 1998, p. 222-229.

Nielsen, Jakob**.** *Usabilidad: diseño de sitios web*. Madrid [etc.]: Prentice Hall, 2000. “Opciones de búsqueda”, p. 224-245.

OPAC Displays by Martha Yee. Draft version. November 24, 1998.

Pastor, Juan A. et al. “Proyecto SABIO: Sistema de Acceso a Bases de Información Organizada”. En: Jornadas Españolas de Documentación (6as: València, 29-21 octubre 1998). *6as Jornadas Españolas de Documentación: los sistemas de información al servicio de la sociedad: actas de las jornadas*. València: Fesabid; Avei, 1998. p. 695-702.

*Readings in information visualization: using vision to think*. Ed. by S.K. Card, J.D. Mackinlay, B. Shneiderman. San Francisco: Morgan Kaufmann, 1999.

Rodríguez Muñoz, José V.; Saorín, Tomàs. “Modelado documental de servicios de información en web”, *El profesional de la información*, vol. 7, nº 9 (septiembre 1998), p. 10-18.

Sabin-Kildiss, Luisa; Cool, C.; Xie, H. “Assessing the functionality of web-based versions of traditional search engines”, *Online* (march-april 2001), p. 18-26.
Shneiderman, B. “Dynamic queries for visual information seeking”. En: *Readings in information visualization: using vision to think*. Ed. by S.K. Card, J.D. Mackinlay, B.Shneiderman. San Francisco: Morgan Kaufmann, 1999. p. 236-243.
Shneiderman, Ben; Don Byrd and W. Bruce Croft. "Clarifying search: a user-interface framework for text searches", *D-Lib Magazine*, January 1997. <www.dlib.org/dlib/january97/retrieval/01shneiderman.html> [Consulta: 25/09/01]
Sinclair, J.; McCullough. *Creación de bases de datos en Internet*. Madrid: Anaya Multimedia, 1997. 504 p.
Stein, Lincoln D. *How to set up and maintain a web site*. Reading [etc.]: Addison-Wesley, 1997. 793 p.
Tittel, Ed et al. *La biblia de la programación CGI*. Madrid: Anaya Multimedia, 1997. 734 p.
Vizine-Goetz, Diane. “Using library classification schemes for Internet resources [en línia]”. En: OCLC Internet Cataloging Project Colloquium (1996: San Antonio, Texas). *Proceedings of the OCLC Internet Cataloging Colloquium*. [Dublin, Ohio]: OCLC, 1996. <http://www.oclc.org/oclc/man/colloq/v-g.htm>. [Consulta: 23/07/00].
Warren, Scott. “Visual displays of information: a conceptual taxonomy”. *Libri*, vol. 51 (2001), p. 135-147.

# 5 Metodología de análisis y desarrollo de bases de datos documentales

## 5.1 Qué podemos esperar de una metodología

En el contexto de los sistemas de información, el término *metodologías* suele generar equívocos. Se suele esperen de ellas cosas que, en realidad, no pueden dar. En concreto, se les pide lo mismo que proporcionan, por ejemplo, los algoritmos en matemáticas, es decir, una solución segura a un problema bien planteado.

Sin embargo, en el desarrollo de sistemas de información no existe nada parecido a los algoritmos (ni a las recetas de cocina). ¿Para qué sirve entonces una metodología? La experiencia nos dice que una metodología sirve, exactamente, para que el resultado final se deba en la mayor medida posible *a la planificación* y, en la menor posible, *al azar o al ensayo y error*. Nada más. Pero nada menos...

Mediante la planificación, un profesional tiene derecho a esperar un grado de éxito mucho mayor que si toma las decisiones al azar o por el método del ensayo y error. Por contra, por muy correcta que sea una metodología, un lego no hará nada bueno con ella. O sea, las metodologías no producen milagros ni eximen de tener una buena base de conocimientos profesionales.

Por tanto, la diferencia entre utilizar o no utilizar una metodología para desarrollar una base de datos radica en la proporción final que puede atribuirse: *a)* al azar; *b)* al ensayo y error; *c)* a la planificación. La cuestión clave radica en que la parte correspondiente a la *planificación* debe ser la que tenga *mayor influencia* en el resultado final de un sistema de información, por razones obvias. Hemos dicho que son razones obvias, pero pese a todo las expecificaremos: es sabido que, con un tiempo ilimitado y con un presupuesto infinito, simplemente por medio del ensato y error y sin necesidad de metología alguna, podemos llegar a la solución óptima. El *pequeño* problema, es que en la vida real nadie tiene presupuestos infinitos ni un tiempo ilimitado.

Por otro lado, es también habitual que las metodologías suenen como un mero puñado de consejos de sentido común, lo cual induce a veces a un cierto y peligroso menosprecio hacia ellas. El problema radica en que, si bien muchos aspectos de las metodologías parecen de sentido común, su contrario también lo parece. Así pues, con una metodología, por lo menos sabemos cuáles de las muchas cosas que *parecen* razonables *son* razonables.

Pongamos un ejemplo: supongamos que alguien afirma que el mejor procedimiento para diseñar una base de datos es escoger un buen equipo informático, después elegir un programa que sea compatible con el mismo y, a continuación, diseñar la base de datos.

Por desgracia, se sabe de gente a la cual el consejo le ha parecido tan adecuado que lo han llevado a la práctica con resultados, por supuesto, bastante lamentables. No les hubiera sucedido así si hubieran conocido uno de los aspectos más básicos del diseño de sistemas de información que aconseja comenzar siempre un proyecto estudiando en primer lugar los aspectos lógicos y no los físicos, o comenzar por la fase de análisis y no por la de implantación, etc. Sin embargo, cuando se explican esa clase de principios

a una audiencia, invariablemente, todo una parte de ella cree que está recibiendo un mensaje de sentido común.

## 5.2 Qué es una metodología

Por otro lado, unas meras reflexiones o unos consejos no son, a pesar de todo, una auténtica metodología. ¿Qué cosas forman parte, por tanto, de una auténtica metodología? Entendemos que, en sistemas de información, toda metodología debe contemplar, como mínimo, tres elementos o tres grupos de elementos, que aquí llamaremos *aparatos*:

*a)* Aparato conceptual
*b)* Aparato instrumental
*c)* Aparato procedimental

El primer aparato, o grupo de elementos conceptuales, tiene la misión de proporcionar a los responsables de desarrollo de sistemas de información unas bases conceptuales mínimas que faciliten su entendimiento de todo el proyecto y la comunicación entre los diferentes actores involucrados en el proceso. En el aparato conceptual se definen las entidades básicas que intervienen en el proyecto y se proporcionan puntos de vista estratégicos.

El aparato instrumental es el responsable de proveer los instrumentos de análisis y de diseño, es decir, es aquella parte de la metodología que, precisamente, a veces se ha confundido, incorrectamente, con un algoritmo.

Finalmente, el aparato procedimental establece las fases y los procedimientos básicos, señalando sus objetivos, así como identifica y describe los productos que deben obtenerse de cada fase de análisis, incluido el producto final.

Así pues, y de acuerdo con lo expuesto, se describirá aquí una metodología de desarrollo de bases de datos documentales que no es un algoritmo, es decir, que no libera, mágicamente, de la obligación de tener una buena formación para poder aplicarla con éxito, pero que ayudará a reducir al mínimo posible los riesgos debidos a la improvisación.

Para elaborar esta metodología, los autores han seguido tres principios o tradiciones, cada una de ella bien cimentada en su propio campo: (2) la tradición del análisis de sistemas, de fuerte influencia informática, que podemos representar en autores como Yourdon; (2) la *Soft System Metodology*, desarrollada por Chekland, muy utilizada para el análisis general de problemas y a su vez muy influenciada por la teoría general de sistemas; por último, hemos tenido siempre como background general (3) las metodologías propias de las Ciencias de la documentación relacionadas con la descripción y el tratamiento de información cognitiva.

A través de diversos proyectos en los que han participado los autores, la metodología que se expone aquí ha sido a su vez puesta a prueba en el terreno de la realidad y refinada con los resultados de la experiencia.

## 5.3 Aparato conceptual

### 5.3.1. Modelos

El punto de partida consiste en considerar la futura base de datos como un *sistema de información* que mantiene registros sobre alguna parte del *mundo real*. A esta parte del mundo real la podemos denominar *sistema objeto*. Por tanto, una base de datos, vista de esta forma, es un sistema *que mantiene registros* para describir o representar cosas del mundo real (documentos, ideas, personas, objetos, etc.).[15]

De este modo, el proceso de análisis y diseño puede concebirse como el intento de obtener un modelo lo más fiel posible de aquella parte de la realidad (sistema objeto), que resulta de interés para el sistema de información. Tenemos entonces el par conceptual: <sistema de información (*S1*) *vs.* sistema objeto (*S2*)> y la relación que les une es que el primero, *S1*, es un modelo del segundo, *S2*, exactamente en el mismo sentido en que un mapa es un modelo de un territorio.

Esta relación nos dice algo que, a primera vista parece obvio, pero que, en la práctica, se olvida con mucha frecuencia: una base de datos es un modelo y su misión como tal es parecerse lo más posible a aquello que intenta representar. Si el sistema objeto consiste en fotografías (imaginemos la base de datos de imagénes de un diario o de una revista), entonces la base de datos documental debe contemplar las características de esa parte del mundo real que son las fotografías, y deberá tener en cuenta el hecho de que una fotografía tiene características formales (ByN, color), características icónicas (aparecen determinadas cosas o personas en las fotografías), que tienen una fecha de toma, que tienen un autor que posee unos derechos de reproducción sobre las mismas, etc.

No tiene sentido entonces esa expresión, por desgracia tan frecuente, que dice así más o menos: “lo siento, no podemos buscar fotografías por el nombre del autor porque no está previsto en el sistema”. Ese tipo de declaraciones son en realidad declaraciones de fracasos y si se ajustaran a la realidad, deberían decir lo siguiente: “lo siento, pero el diseño de este sistema de información es tan deficiente que no ha sido capaz de reproducir aspectos esenciales de la realidad que se supone debería representar”.

### 5.3.2. Subsistemas

El segundo punto de partida consiste en considerar que, desde el punto de vista de las bases de datos documentales, todo sistema objeto *(S2)* se compone al menos de dos componentes o subsistemas:

> *a)* La empresa o el sistema social que necesita de la base de datos o, dicho de forma más abstracta, el *sistema de actividades humanas (SAH)* en el que se inscribe la futura base de datos.
> *b)* El conjunto de cosas, entidades o documentos que deberán ser descritos y representados en la base de datos, o dicho de forma más abstracta, el *sistema de entidades registrables (SER)* que estará incluidas en la base de datos.

[15] En el apartado 1.2.1 ya se ha hecho referencia esta concepción de la base de datos.

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, nos puede convenir considerar que, a su vez, dentro del SAH podemos distinguir entre el poseedor o propietario del sistema y los usuarios o beneficiarios del sistema (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 a 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 beneficiarios 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 a las entidades en sí mismas (piensen 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*: como documentos en papel (bases de datos bibliográficas), films (bases de datos de cinematografía), obras de arte (bases de datos de museos) o monumentos (bases de datos de patrimonio arquitectónico).
>
> *c) Personas:* datos biográficos de personajes históricos o de personalidades contemporáneas, cargos de la Administración, etc.
>
> *d)Conceptos:* como 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).

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 entidad-relación (que se explica más adelante) 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.

## 5.4 Aparato instrumental

El aparato instrumental de una metodología proporciona los instrumentos de análisis que puede utilizar el analista. En concreto, tres son los instrumentos principales que se pueden emplear: el modelo entidad-relación, desarrollado originalmente por Chen (1976), el diccionario de datos y la norma ISBD.

### 5.4.1 Modelo Entidad-Relación

El modelo entidad-relación (o modelo E-R) ayuda a detectar sin ambigüedades las entidades que formarán parte de la base de datos, es decir, las clases de cosas que estarán descritas y representadas en la base de datos. La fuente de ambigüedad suele provenir del hecho de que no siempre es fácil distinguir si alguna cosa es una entidad o es un simple atributo de una entidad. Más adelante pondremos ejemplos sobre esto. El modelo E-R utiliza los siguientes conceptos:

- Entidad
- Atributo
- Relación

Según este modelo, si las bases de datos representan a cosas u objetos del mundo real, tales cosas deben ser identificables y deben tener algunas propiedades. A los objetos sobre los cuales una base de datos almacena información se les denomina, como ya sabemos, *entidades*, y pueden ser como ya hemos señalado cosas, personas o conceptos.

La única restricción aplicable aquí es que las entidades que han de estar representadas en una base de datos deben ser identificables y, por tanto, debe ser posible señalar a una cualquiera de ellas sin ambigüedad. Por tanto, los visitantes de un museo, por ejemplo, no pueden ser entidades, ya que no están identificados (un museo puede saber cuántos visitantes tiene cada día, pero no sabe quiénes son). En cambio, los usuarios de una biblioteca sí pueden ser entidades ya que todos deben estar identificados antes de tener

derecho al uso de la biblioteca (en la base de datos de una biblioteca, además de documentos, están representados los usuarios).

Los *atributos*, por su parte, son las propiedades relevantes que caracterizan a una entidad. En este sentido, el término relevantes significa lo siguiente: relevantes para el problema de información que se está considerando solucionar mediante la base de datos. Teniendo en cuenta que, en principio, los atributos de una entidad son virtualmente ilimitados, será labor del documentalista seleccionar en cada caso cuáles son los que se consideran más relevantes.

El modelo distingue entre *tipo* de entidad y *ocurrencia* de entidad. Un tipo de entidad define un conjunto de entidades constituidas por datos del mismo tipo, mientras que una ocurrencia de entidad es una entidad determinada y concreta. Cuando se diseña una base de datos el objetivo del documentalista debe consistir en definir un *tipo* de entidad, que obtiene estudiando ocurrencias concretas de entidades.

#### 5.4.1.1. Registros y campos *vs.* entidades y atributos

Ahora ya podemos relacionar registros y campos con entidades y atributos. En primer lugar, podemos decir que un registro es una representación de una entidad en la base de datos y, por lo tanto, cada registro describe a una entidad. Por ejemplo, en una base de datos bibliográfica, cada documento se describe en un registro.

En segundo lugar, si los registros describen entidades del mundo real, los campos del registro corresponden a los atributos de la entidad. De este modo, si un tipo de entidad posee los atributos A, B, C, (por ejemplo, la entidad tiene un autor, un título y una fecha) el modelo de registro debe poseer los campos *A*, *B*, *C* (*autor*, *título*, *fecha*). Ahora bien, cuando se estudia el concepto de campo en una base de datos, es necesario diferenciar entre los siguientes conceptos:

*a) Etiqueta* del campo
*b) Valor* del campo
*c) Dominio* del campo

La etiqueta es el nombre del campo, es decir, una constante que identifica una zona del registro. Esta constante suele ser una corta cadena de caracteres como *Autor*, *Título*, etc.. El valor es una variable, y se refiere al contenido concreto de un campo concreto y puede ser distinto para cada campo de cada registro. El dominio, por su parte, es un concepto lógico, y se refiere al conjunto teórico del cual puede tomar sus valores un campo. Por ejemplo, el dominio del campo *Año*, puede ser el conjunto formado por los años de publicación de los documentos.

**Figura 5.1: Un ejemplo de registro que describe (representa) a un libro**

| | |
|---|---|
| *Título* | ADN: El secreto de la vida |
| *Autor* | James D. Watson |
| *Fuente* | Madrid: Taurus, 2003 |
| *Año* | 2003 |
| *Páginas* | 475 |
| *ISBN* | 84-306-0514-2 |
| *Descriptores* | ADN, Biología, Evolución, Genoma humano |

Veámoslo con este ejemplo. De acuerdo con el registro de la figura 5.1, el segundo campo (*Autor*) o zona de información se puede analizar así:

- *Etiqueta del campo* (no cambia nunca): *Autor*
- *Valor del campo* (puede ser diferente para cada registro): James D. Watson
- *Dominio del campo*: nombres de los responsables intelectuales de los documentos.

### 5.4.1.2. Generalizaciones y abstracciones

Al igual que distinguimos ente tipo y ocurrencia de entidad, debemos diferenciar también entre modelo de registro y ocurrencia de registro. Un tipo de entidad se forma por abstracción y/o generalización. Abstracción o generalización significa que se ignoran ciertos aspectos distintos de diversas ocurrencias de entidad y se forma con todas ellas un tipo unitario, o que se generalizan a todas las entidades ciertos rasgos que presentan regularmente ciertas entidades.

Por ejemplo, supongamos que aplicando el modelo E-R a un problema de información (por ejemplo, una base de datos para automatizar el archivo de un medio de comunicación), nos muestra como primer resultado los siguientes tipos de entidades:

1. Artículos de revistas
2. Artículos de prensa diaria
3. Capítulos de libros
4. Libros
5. Informes
6. Fotografías de personajes
7. Fotografías de sucesos
8. Fotografías de estudio
9. Infografías

Una simple generalización reduce los nueve tipos de entidades a dos, puesto que las entidades 1 a 5, pueden reducirse, por abstracción, a una sola: *Documentos escritos*, y los tipos de entidades 5 a 9 al tipo de entidad: *Documentos gráficos.* La entidad *Documentos escritos* deberá tener un atributo denominado *Tipo de documento*, que permitirá describir qué clase de documento es: artículo, libro, etc. Por su parte, la entidad *Documentos gráficos*, deberá tener también un campo denominado *Tipo de documento*, que permitirá indicar si es una fotografía de personas, fotografía de paisajes, o si es una infografía, etc.

### 5.4.1.3. Relaciones

El tercer componente del modelo E-R son las relaciones. Vamos a examinarlas aquí. Las entidades del mundo real pueden tener relaciones entre ellas y, mientras las entidades suelen nombrarse mediante sustantivos, las relaciones se nombran mediante verbos. Por ejemplo, consideremos el caso de una base de datos sobre teatro español. Un análisis intuitivo nos revelaría la existencia de dos entidades relevantes para el sistema: *[obras]* y *[autores]*, y veríamos que entre ambas entidades existe la relación

<escriben>, que significa más explícitamente que *[autores teatrales]* <escriben> *[obras de teatro]*.

Un aspecto importante de la relación es su *grado*, el cual indica el número de elementos que pueden participar en cada uno de los extremos de la relación, en este caso *[autores]* y *[obras]*. Este grado puede ser de uno a uno (1:1), de uno a muchos (1:n) y de muchos a muchos (n:m). Una manera típica de representar estas relaciones y su grado es utilizando diagramas y expresiones textuales. En estos diagramas, las entidades se representan como rectángulos y las relaciones como rombos.

Así, por ejemplo, la relación que existe entre el número de ISBN y un libro es una relación de 1:1 (se lee "relación de uno a uno") porque un número de ISBN se asigna a un solo libro, y cada libro tiene un solo número de ISBN.

En cambio, la relación entre profesores y universidades es de 1:n, (se lee “de uno a muchos”) porque cada profesor pertenece a una sola universidad, y una universidad tiene muchos profesores.

Finalmente, una relación de tipo n:m (“de muchos a muchos”) sería la que existe entre autores de teatro y obras de teatro, porque un autor puede escribir diversas obras de teatro, y una obra de teatro puede estar escrita por varios autores y justamente ese es el significado de las letras *n* y *m*.

Además, la participación de la entidad puede o no ser obligatoria, lo cual significa que una entidad obligatoria interviene siempre en la relación. Por ejemplo, en la relación entre ISBN y libros, la participación de la entidad *[libros]* es obligatoria, porque siempre que hay un número de ISBN hay un libro, en cambio lo contrario no es cierto, porque hay libros que no tienen número de ISBN.

Esta última parte del análisis entidad-relación (grado y participación) es muy importante en el diseño de bases de datos relacionales, porque ayuda a modelar los datos de la empresa y a representarlos en tablas normalizadas.

En cambio, en sistemas documentales no es tan importante porque éstos no suelen utilizar el modelo relacional, ni necesitan modelar relaciones complejas entre entidades, como las que se dan en los sistemas de gestión administrativos.

En algunos sistemas documentales, las entidades, de hecho, no mantienen relaciones entre ellas que deban ser reflejadas en el diagrama E-R. Por ejemplo, en una típica bases de datos documental bibliográfica no suele existir ninguna relación entre las entidades representadas (típicamente artículos de revista y monografías) que deba ser tenida en cuenta en el modelo E-R.

En tales situaciones, el modelo E-R aporta una importante claridad conceptual y proporciona una terminología común a todos los miembros que participan en el diseño. Sin embargo, el propósito de las herramientas de diseño no es tanto proporcionar soluciones para situaciones que son bien conocidas, sino para las situaciones no conocidas o menos típicas y, en este sentido, el modelo E-R puede resultar de ayuda.

Por ejemplo, y volviendo al caso anterior, donde se nos pide diseñar una base de datos sobre teatro español. Supongamos que tenemos dudas sobre el siguiente aspecto: no sabemos si considerar que el autor (y todos sus datos biográficos) son atributos de la obra de teatro, o bien si considerar que autor y obras de teatro son entidades distintas, como hemos supuesto en la presentación del caso.

Si adoptáramos el primer punto de vista, tendríamos que diseñar un único modelo de registro, donde los atributos del autor serían otros tantos campos, junto con los atributos de la obra de teatro. En cambio, si adoptamos el segundo punto de vista, necesitaremos diseñar dos modelos de registro, uno para obras de teatro y otro para autores. Puede ser que la simple intuición no indique cuál es el camino correcto en este o en otros casos parecidos, pero si queremos estar seguros de no equivocarnos en nuestra decisión, siempre podemos aplicar el siguiente procedimiento:

*1º.* En caso de duda, tratar los objetos como entidades distintas.
*2º.* Determinar la relación entre entidades.
*3º.* Determinar su grado.
*4º.* Si la relación es de grado *1:1,* entonces se trata de una sola entidad y un solo modelo de registro es suficiente para representarla. Por ejemplo, el número de ISBN es, de hecho, un atributo de la entidad libro, y para representarla es suficiente un solo registro, con un atributo que incluya el número de ISBN.
5º. Si la relación es de grado *n:1,* se trata de dos entidades y, por lo tanto, necesitamos dos modelos de registro, uno para cada entidad, y cada uno de ellos debe contar con un campo con un dominio común.
*6º.* Si la relación es de grado *n:m,* se trata también de dos entidades, pero en este caso, además, podemos necesitar singularizar la relación, con lo cual podemos necesitar añadir a los dos modelos de registro antes considerados un tercer modelo de registro que represente a la relación y que incluya un campo común de cada entidad. Esta última clase de modelos de registro son necesarios cuando la relación es dinámica. Por ejemplo, en una biblioteca con servicio de prestamos, la relación entre libros y usuarios, que es el préstamo, necesita estar representada en un registro, de modo que cada préstamos crea un registro que contienen la relación entre el libro prestado y el usuario que lo tiene en préstamo.

En nuestro ejemplo, la aplicación de esa regla nos indicaría que la decisión acertada consiste en utilizar dos modelos de registro: uno para representar obras de teatro y otro para representar autores teatrales. Como la relación autor-obra no es dinámica (los autores de una misma obra no van cambiando con el tiempo), no sería necesaria el tercer modelo de registro para representar la relación .

¿Qué sucedería si no procediéramos como indica esta norma? En tal caso, la carga de datos sería poco eficiente, porque para autores muy prolíficos tendríamos que entrar los mismos datos tantas veces como obras de teatro hubiera escrito.

En general, si un autor ha escito *n* obras de teatro, tendríamos que repetir sus datos *n* veces. Además, la redundancia, como es sabido, genera inmediatamente inconsistencias, y tendríamos enseguida, por ejemplo, diversas fechas de nacimiento para un mismo autor. Es evidente que si no detectamos ese error de diseño a tiempo, no tardará en hacerse evidente en algún momento de la fase de carga de datos, pero no debería ser menos evidente que si podemos evitar el error en la fase de diseño estaremos trabajando

con mucha mejor calidad (ahora que está tan de moda este tema) que si necesitamos llegar a la implantación para detectar los errores, tal vez después de meses de trabajo que, de golpe, se revelarán inútiles.

Una advertencia final sobre el modelo E-R. Primero, cuando se utiliza para diseñar bases de datos relacionales, las reglas para tomar decisiones son más complejas, porque la descomposición de datos a la que obliga el modelo relacional implica la necesidad de representar no sólo las entidades, sino también las relaciones entre entidades mediante tablas adicionales. Los interesados en esos aspectos de diseño pueden consultar Jackson (1990).

### 5.4.2 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, por tanto, 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

Por 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. P.e. "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 ",". P.e. edición óptica, publicación digital, documentación.

Etiqueta: FPublicación
Dominio:
La fecha de publicación de la noticia, indicada con el siguiente formato:
DD/MM/AAAA.
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. P.e. 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:

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

2º. Que el *Tipo* se refiere, en cambio, al tipo de dato que admite el campo. Los tipos de datos suelen ser: numérico, alfanumérico, fechas 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 "alfanumérico" 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 fechas 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.

3º. 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.

4º. 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 na 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 5.1: Grupos de campos en una base de datos documental**

| ***Campos*** | ***Explicación*** |
|---|---|
| **De control** | Son aquellos que tienen por objetivo controlar la gestion interna del registro. Por ejemplo, el número del registro (ID), la fecha de entrada, la fecha de modificación, etc. Són aquells que tenen per objectiu controlar la gestió interna del registre. P.e. número de registre, data d'entrada, data de revisió, etc. |
| **Descriptivos** | Se utilizan para describir las características de las entidades o documentos de la base de datos (autor, título, fecha, etc.) |
| **Temáticos** | Para representar el contenido o tema del documento o entidad representada en la base de datos (resumen, descrptores, etc.) |
| **Derechos** | Indican, en su caso, que restricciones o derechos limitan la utilización del documento y/o qienes estám en posesión de los mismos |
| **Ubicación** | Indican, en su caso, la ubicación o localización del documento original. Estos datos 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. |

### 5.4.3 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 testeado.

Por ejemplo, los atributos estructurales de cualquier clase de documento pueden ser adecuadamente modelados siguiendo las normas internacionales ISBD. Recordemos que 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.

Entendemos que tal postura es un error: primero, porque 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, el programa documental debería permitir, como es el caso de diversos de ellos (p.e. 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.

## 5.5 Aparato procedimental

El principio general de diseño 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, sin embargo, suele suceder, ya que, en la práctica, existen muchas formas de violar ese principio general a causa de malos hábitos de trabajo.

Otra manera de enfocar incorrectamente este proceso consiste en querer abordar directamente el diseño del sistema de información e, incluso en querer visualizarlo por completo en nuestra mente, sin saber antes nada del sistema objeto (la empresa u organismo y los tipos de entidades). El resultado, claro está, será una visión caótica. Todas las interrogantes se agolparán en nuestra mente y seremos incapaces de despejar una sola de ellas.

Lo correcto en ambos casos es comenzar a diseñar los aspectos lógicos (nivel conceptual) e ignorar, de momento, los aspectos físicos; y, por otro lado, comenzar por analizar el sistema objeto y, sólo después de conocerlo bien, iniciar el diseño del sistema de información.

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

1. Análisis
2. Diseño
3. Implantación

Otra forma de enfocar el ciclo de vida de 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 5.1 :Dirección del diseño en el ciclo de vida 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 ciclo de vida, 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). Las fases de implantación pueden subdividirse en cuatro subfases (a, b, c, d, 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 ciclo de vida. En concreto, en esta metodología se propone la división de fases del cuadro sinóptico 5.2:

**Cuadro 5.2: Ciclo de vida de una base de datos documental**

*1. Análisis*
1a. Análisis de la empresa u organización (sistema de actividades humanas), incluyendo su entorno
1b. Análisis de las cosas u 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, etc.)

*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 ciclo de vida.
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 parece estrictamente lineal, en realidad, el proceso de diseño también tiene mucho de intercativo, 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 del ensayo y error, como ya se advirtió, sino que lo integra de un modo controlado de 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, tiene que 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.

En todo caso, 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, implantación) tiene unos objetivos, debe producir unos resultados concretos y utilizar unas herramientas determinadas.

### 5.5.1 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, de una base de datos en este caso.

Como ya vimos anteriormente, 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 registrable (SER):* las cosas, personas o conceptos que estarán representadas 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, deberá conocerse lo mejor posible antes de iniciar cualquier actividad de diseño.

El resultado de esta fase de análisis es una descripción textual que puede incluir gráficos de ser necesario, sobre el SAH, que suele denominarse *Informe de funciones* o *Informe de oportunidad,* [16] y que debe incluir, como mínimo, los siguientes aspectos:

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

[16] En otras versiones de esta metodología, a este informe se le denominaba "Modelo esencial". Como es fácil suponer, el nombre es lo de menos. Ahora se opta por el nombre de "Informe de funciones" o "Informe de oportunidad" para utilizar expresiones más estandarizadas.

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 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 dee estudio o análisis. Por 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 resultado de esta fase debe consistir en la identificación clara y sin ambigüedades no solamente de las cosas, personas o 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.

El *Informe de funciones* debería ser aprovado 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.

### 5.5.2 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 los 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 temáticoe 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 incluir 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, y/o a qué aplicaciones podrá dar soporte.

5. *Una propuesta de tratamiento documental*.

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*.

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-relación y (3) el diccionario de datos.

### 5.5.3 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 y solicitar documentación sobre sus programas, prestaciones, 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):

*1.1*. Grado de compatibilidad con la plataforma informática de la empresa o corporación.

*1.2.* Grado de satisfacción de los requerimientos establecidos en el diseño conceptual.

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

*1.4*. Base instalada de clientes: ¿pueden proporcionar referencias de otros clientes? ¿Existe un club de usuarios de la aplicación?

*1.5*. 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.)

*1.6*. Presupuesto

El orden indicado no es significativo: en algunos casos, el punto 1.6 puede ser primordial mientras que el punto 1.1. 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 1.1 al 1.6 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***.

*2.1*. 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.

*2.2*. 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, listas de valores predefinidos, etc.).

***4. Realización de pruebas con una colección-test*** de documentos o de entidades a ser representadas para comprobar la consistencia de los modelos y esquemas de registros detallados en las fases previas y contenidos en el diccionario de datos.

***5. Realización de los 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*** para la carga de datos: facilidades para dar altas, realizar exportaciones, 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 o tipos de consultas que requieren algún reajuste en los campos, etc.

***8. Test de usabilidad***. Se puede aplica 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 como las resuelves, además también les pediremos que proporcionen 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, con estos cambios 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:

*i). 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.

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

*iii). Los operadores.* Son quienes realizan la carga de datos. Operadores y analistas pueden ser las mismas personas o pueden ser personas distintas. En este último caso, a veces, los analistas realizan su trabajo sobre plantillas que después, vuelcan los operadores en la base de datos.

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

***11. Inicio del proceso de carga de datos*** 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:

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

*ii). 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 copia de seguridad, 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 otros programas de base de datos.

*iii). 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 ambas opciones, las bases de datos documentales suelen proveer opciones. 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 capítulo 6 se dedica a analizar, de forma pormenorizada, todo lo referente a evaluación y control de calidad en bases de datos).

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

iv). *Política de mantenimiento y explotación*. Se editará la versión 1 del *Libro de estilo de la base de datos*, que incluye:

a). La versión definitiva del modelo conceptual.

b). La normativa de tratamiento documental.

c). Política de formación del personal técnico y organización de sesiones de formación de los usuarios finales.

***12. Acciones de promoción***, en su caso.

## 5.6 Conclusiones

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 y/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 nuestra profesión 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 nuestro campo puedan demostrar los beneficios de una adecuada formación académica, del trabajo bien realizado y de la planificación, porque en nuestro 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”.

**Bibliografía**


Abadal, E. “Diseño y creación de una base de datos en un medio de comunicación”. En: Fuentes, M.E. *Manual de documentación periodística*. Madrid: Síntesis, 1995. pp. 196-211.

AENOR. *Norma UNE 50-106-90. Documentación. Directrices para el establecimiento y desarrollo de tesauros monolingües.* Madrid: AENOR, 1990, 47 pp.

Baiget, T. *Análisis de sistemas de información*. Barcelona: Institut Català de Tecnologia, 1986, 64 pp. (documento reprografiado).

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

Checkland, P. B.; Scholes, J. *Soft systems methodology in action*. Chichester: Wiley , 1990.

Chen, P.P-S. "The entity-relationship model: towards a unified view of data". *ACM transactions on databases systems*, v. 1, n. 1, 1976, pp. 9-36.

Codina, L. *Sistemes d'informació documental: concepció, anàlisi i disseny de sistemes de gestió documental amb microordinadors*. Barcelona: Pòrtic, 1994, 224 pp.

Codina, L. “Metodología de creación de bases de datos documentales”. Parte I. *Information world en español*, n. 33, abril 1995; Parte II. *Information world en español*, n. 34, mayo 1995;

Codina, L. "Metodología de análisis de sistemas de información y diseño de bases de datos documentales: aspectos lógicos y funcionales". En: Baró, J.; Cid, P. (eds.). *Anuario SOCADI de Documentación e Información 1998*. Barcelona: SOCADI, 1998, pp. 195-210.

Curras, E. *La información en sus nuevos aspectos.* Madrid: Paraninfo, 1988, 307 p.

Jackson, G. A. *Introducción al diseño de bases de datos relacionales*. Madrid: Anaya, 1990, 203 pp.

Lewis, P. *Information systems development*. London: Pitman,1994, 260 pp.

Osborne, L.; Nakamura, M. *Systems analysis for librarians and information professionals*. 2nd ed. Englewood, CO: Libraries Unlimited, 2000. 261 pp.

Puig Torné, J. *Proyectos informáticos: planificación, desarrollo y control*. Madrid: Paraninfo, 1994.

Underwood, P. G. *Soft systems analysis and the management of libraries, information services and resource centres*. London: Library Association, 1996, 198 pp.

Van Slype, G. *Los lenguajes de indización: concepción, construcción y utilización en los sistemas documentales.* Madrid: Fundación Germán Sánchez Ruipérez, 1991, 198 pp.

Walker, D.W. *Sistemas de información basados en ordenador.* Barcelona: Marcombo, 1991

Yourdon, E. *Análisis estructurado moderno*. México: Prentice-Hall Hispanoamericana, 1993, 735 pp.

# 6 Evaluación de bases de datos

## 6.1 Introducción

En los capítulos 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 (www.askeric.org); *Archives Le Monde*, la base de datos de prensa del periódico *Le Monde* (archives.lemonde.fr); la base de datos de fotografías *AGE Fotostock* (www.agefotostock.com) o la base de datos de recursos de Internet *Sosig* (www.sosig.ac.uk).

Ahora bien, una vez estas bases de datos se encuentran en el mercado, la pregunta legítima es la siguiente: ¿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, van a constituir el eje de este capítulo.

Diversos autores[17] han hecho referencia a los dos principales enfoques que han presidido los estudios de evaluación de sistemas de recuperación de la información. En primer lugar, existe una línea de investigación que se centra en el análisis de las prestaciones de los sistemas de recuperación de la información. En este caso, se trata de analizar la eficacia de recuperación de los sistemas de RI y se toma como objeto de análisis los distintos componentes de los sistemas de RI, es decir, los aspectos operativos. El ejemplo clásico son los experimentos de Cranfield.[18] El principal criterio de evaluación se basa en la relevancia de los resultados, a fin de calcular los índices de precisión y exhaustividad (ver cap. 2) que otorgan un valor cuantitativo al SRI. Esta línea de trabajo ha tenido una línea de continuación en las TREC (Text REtrieval Conference). Más recientemente, los estudios que analizan las prestaciones de los directorios y motores de búsqueda en Internet acostumbran a adoptar esta orientación ya que se dedican a analizar las prestaciones de búsqueda, la cantidad de información recuperada, los resultados, etc.[19]

La segunda línea de investigación, más reciente, es la que toma al usuario como objeto prioritario. En esta perspectiva, más centrada en el usuario, adquieren mayor relieve aspectos como la *utilidad* de la información recuperada, la *usabilidad*, el *contexto* de la búsqueda y la *satisfacción* del usuario. Un ejemplo de tales estudios es *SERVQUAL*, un modelo de evaluación de productos o servicios basado en las expectativas del usuario que Tom Wilson (1998) utiliza para realizar una encuesta europea sobre calidad en las

[17] Podemos destacar a Dervin y Nilan (1986), Vickery (1987) o Ingwersen (1992) y, en España, a Dolores Olvera (1999) o Francisca Abad (2002).

[18] Fueron llevados a cabo por Ciril Cleverdon en el Cranfield Institute of Aeronautics (Gran Bretaña) a finales de los cincuenta y principios de los sesenta y tenían como objetivo evaluar la eficacia de distintos lenguajes de indización utilizando los índices de precisión y exhaustividad para medir el rendimiento del SRI. El contexto experimental estaba formado por una colección de documentos cerrada (unos 1400), un conjunto de unas 300 preguntas y unos juicios de valor binarios (sí/no) sobre la precisión de los documentos en relación a las preguntas.

[19] El estudio evaluativo de motores de búsqueda de Chu y Rosenthal (1996) es un ejemplo claro. Se analizan aspectos de los SRI (precisión, tiempo de respuesta, prestaciones de búsqueda, etc.) de Altavista, Excite y Lycos.

bases de datos y que Xie (1998) también retoma en un estudio basado en esta metodología. En estos casos, se trata de tener en cuenta la capacidad del SRI para adaptarse a las características del usuario y servirlo de forma adecuada. En este contexto, es muy importante disponer de distintas técnicas de recogida de datos sobre cuál ha sido el comportamiento del usuario o la utilización que éste ha hecho de la base de datos. Como vemos, esta preocupación por el usuario y su contexto comporta que se valoren mucho más los aspectos de carácter cualitativo y que, inevitablemente, los resultados estén teñidos con cierto grado de subjetividad.

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 RI 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 preferente 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 curso de los años se han ido presentando análisis de estos distintos objetos. El enfoque que aquí presentamos quiere tener un carácter integrador y, por tanto, es aplicable tanto a catálogos de bibliotecas, bases de datos especializados o motores de búsqueda.

### 6.1.1 Evaluación y calidad

En los primeros párrafos ya se han mencionado los dos conceptos que van a marcar la pauta interna de este capítulo: se trata de “evaluación” y de “calidad”. En este capítulo no se van a tratar excesivos detalles sobre cuestiones generales de ambos conceptos, sino que se van a aplicar directamente al objeto por antonomasia de nuestro texto, las bases de datos documentales.

La evaluación es un proceso por medio del cual se pretende obtener un juicio de valor o una apreciación sobre un objeto, una actividad, un proceso o sus resultados, a partir de unos criterios o normas que se toman como modelo. En suma, evaluar implica determinar el valor de alguna cosa. Por eso, evaluar implica recoger una serie de datos sobre un producto o servicio determinado y compararlos con una serie de criterios o estándares previamente establecidos. Así pues, en toda evaluación es imprescindible obtener una medida (“lo que tenemos”) y, además realizar una operación de comparación con un estándar (“lo que deberíamos tener”). Fijémonos, no obstante, que para realizar esta comparación hay que llevar a cabo una tarea previa consistente en

establecer unos indicadores o criterios de evaluación. A esta labor vamos a dedicar el grueso de este capítulo.

La calidad, el otro concepto clave, es el conjunto de características que debe estar presente en un producto o servicio para satisfacer las necesidades de sus usuarios. Se trata, por tanto, de un atributo fundamental de cualquier servicio o producto que se coloca en el mercado. Como es bien sabido, el interés por la calidad procede del sector industrial y se sitúa en Japón en la década de los 50. A partir de aquí se extendió por todos los países industrializados y se amplió a todos los productos y servicios y a todos los sectores; en el ámbito específico de la documentación, por ejemplo, se dispone ya de una amplia bibliografía sobre actuaciones realizadas en servicios y productos documentales.

Aunque sería difícil proporcionar una definición académica de "calidad de una base de datos", no pasaría lo mismo sobre cuáles han de ser los elementos clave de esta calidad, aspecto en el cual sería más fácil encontrar el consenso. Sin entrar en detalles, se podría afirmar que la calidad de una base de datos es aquello que hace decidirse a un usuario a realizar consultas en la base de datos *A* y no en otra. Se trata de un conjunto de elementos bien ensamblados que hace referencia, en primer lugar, (1) a unos materiales de interés, bien organizados y estructurados (contenido), pero también al (2) sistema de RI (el continente) que tiene por ideal ofrecer al usuario el mayor número de documentos relevantes que se ajusten a su pregunta con el mínimo esfuerzo por su parte y, finalmente, (3) a un buen sistema de distribución y de promoción de este producto (accessible, a buen precio, etc.).

## 6.2 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 indicadores 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 (p.e. evaluación de revistas, evaluación de sistemas de recuperación de la información, etc.)

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 y que son los siguientes: consistencia, alcance y cobertura, alcance temporal, grado de errores y exactitud, accesibilidad y facilidad de uso, integración (armonización con otras bases de datos similares, del mismo distribuidor), salida de la información, documentación que acompaña la base de datos, apoyo al usuario y formación y relación calidad-precio. Aunque no es la primera propuesta seria, se trata del estudio más citado, con mucha diferencia, por los investigadores de este ámbito.[20]

[20] Con anterioridad, otros autores como Cleverdon (1974), Van Rijsbergen (1975) o Auster (1979) también habían realizado propuestas de criterios para evaluar la calidad de los SRI. Sin embargo, el trabajo de SCOUG, quizá más amplio, ha sido el más utilizado como referente de partida.

Estos criterios han constituido el punto de partida para muchos sectores profesionales de la Documentación implicados en evaluación y también para estudios publicados posteriormente como los de Wilson (1998), Xie (1998), Rodríguez Yunta (1998) o Johnson et al (2001), 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 los noventa 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.

En lo que se refiere a grupos de trabajo y de investigación se puede citar a la *Finnish Society for Information Services* que, en 1989, creó un grupo de trabajo para evaluar la calidad de las bases de datos finlandesas y que elaboró una amplia lista de criterios agrupados en cinco categorías: telecomunicaciones, programa de recuperación, contenidos, ayudas a la búsqueda y costes (Juntunen, 1991). Posteriormente, en 1993, se inició el proyecto *EQUIP* (*European Quality in Information Programme*), iniciado en 1993, con el objetivo, en este caso, de examinar y promover la aplicación de la gestión de la calidad al sector de la información.

Uno de los trabajos realizados fue la elaboración y tabulación de una encuesta para valorar la presencia de diez indicadores de calidad para bases de datos. Los criterios tomados en consideración parten de *SCOUG*. Las encuestas solicitaban a los productores que priorizasen estos indicadores (y entonces establecieron rankings por países, tipo de base de datos, tipo de organización, etc.). También se puede hacer mención de la tarea del *Centre for Information Quality Management* (CIQM)**,** creado también en 1993 con el apoyo de la *Library Association* y del *UK Online User Group* (UKOLUG) y actualmente un servicio de la compañía *Information Automation Limited* (www.i-a-l.co.uk/ciqm_index.html), que se ha convertido en un centro de documentación especializado en problemas relacionados con la calidad de las bases de datos.

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:
a) *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 análisis e indización, los documentos escogidos, etc.
b) el *sistema de recuperación de la información (el continente):* 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 podría estar disponible en diferentes sistemas de recuperación de la información)
c) *la gestión y administración de la base de datos:* 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. 6.1. Criterios de evaluación**

| | |
|---|---|
| **Contenido de la base de datos** | Grado de exactitud y precisión<br>Errores gramaticales y mecanográficos<br>Errores de omisión<br>Fiabilidad de los datos<br>Registros duplicados<br><br>Alcance y cobertura<br>Grado de cobertura o alcance temático<br>Cobertura geográfica y lingüística<br>Grado de inclusión<br>Estructura<br>Tamaño<br>Nivel de crecimiento<br><br>Actualización<br>Grado de actualización<br>Periodo de actualización<br><br>Consistencia<br>Consistencia de la catalogación<br>Consistencia en el análisis de contenido |
| **Sistema de recuperación de la información** | Prestaciones del lenguaje de interrogación<br>Precisión<br>Exhaustividad<br>Tiempo de respuesta<br>Utilidad<br>Formatos de visualización<br>Interfaz (Amigabilidad) |
| **Gestión de la base de datos** | Documentación sobre la base de datos<br>Atención al usuario<br>Precio y sistema de facturación<br>Sistema de distribución |

### 6.2.1 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 sistema de RI 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 competencia directa del productor de la base de datos.

#### 6.2.1.1 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 con 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", que podría ser otra forma de decir lo mismo.

- *Errores gramaticales y mecanográficos*

O'Neill (1988) publicó en ARIST un artículo que presentaba una amplia panorámica sobre lo que se ha escrito sobre detección y corrección de errores en bases de datos bibliográficas. Posteriormente, Spinak (1995) y Ortego (1996) han publicado sendos artículos dedicados principalmente a este tipo de indicadores que incluyen errores gramaticales (ortografía, sintaxis, semántica) y también errores mecanográficos, todos ellos muy presentes y habituales en el entorno web.

En lo que se refiere a la ortografía, uno de los errores más frecuentes es la acentuación. La presencia de términos sin acentuar puede constituir un problema ya que algunos SRI no son capaces de buscar el mismo término acentuado o sin tilde. Para solventar los errores ortográficos existe una solución simple y barata: usar un buen corrector. Por otro lado, no hay que olvidar que afectan poco a la recuperación de la información.

Los errores sintácticos remiten a problemas de concordancia de género, número, caso, etc. de las palabras, aspecto que tiene que ser corregido manualmente, y los semánticos, a las variaciones de significado que se pueden producir porque se digita una palabra (eficacia) por otra (eficiencia), ambas correctas ortográficamente.

Finalmente, en lo que respecta a los errores mecanográficos, se han definido diversos tipos: permutación (*biblitoeca* per *biblioteca*), omisión (*odenador* por *ordenador*), sustitución (*sintasi* por *sintaxis*), repetición de letras, inserción (de letras o de espacios en blanco).[21] Hacer un repaso a los índices de campo, generando un listado, puede ser útil para detectar errores de mecanografía, que se pueden corregir mediante sistemas automáticos de detección y corrección.

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 adecuada. Son especialmente preocupantes en áreas como las finanzas, o la información medica o jurídica en las cuales se pueden tomar decisiones muy importantes sobre la base del contenido de la información recuperada.

- *Errores de omisión*

Los registros incompletos (p.e. falta fecha de publicación, idioma, tipo de documento, etc.) constituyen errores de omisión y pueden ser fácilmente detectables y prevenibles. Para detectarlos, se puede pensar en rutinas automáticas que indiquen en qué registros se encuentran campos vacíos, y para prevenir este error, lo mejor utilizar campos

[21] Ortego (1996: 507) los describe con más detalle y se remite a estudios precedentes.

obligatorios o de campos con entrada automática (p.e. fecha de modificación o de entrada, analista, etc.).

- *Fiabilidad de los datos*
En este caso, se refiere a la exacta correspondencia del contenido de los registros con los documentos a los que representan. Un ejemplo de poca fiabilidad lo encontramos en en la base de datos del ISBN la cual, debido al bajo control existente en el suministro de información a la base de datos (se realiza antes de la publicación del libro), incluye registros de libros pendientes de publicación, con datos inexactos del autor, título, páginas, o clasificación. En el caso de base de datos de recursos web, hay que estar atentos a la presencia de enlaces inexistentes o sin actualizar.

- *Registros duplicados*
El origen de este problema son los errores de exactitud que se han descrito en este mismo apartado y también las inconsistencias (v. 6.2.1.4). 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.

Para poder controlar este problema, es necesario realizar una comprobación antes de entrar un nuevo registro en la base de datos. También se pueden elaborar sistemas automáticos que comparen el contenido de determinados campos-clave para detectar los que son idénticos.

#### 6.2.1.2 Alcance y cobertura

- *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 usuario puede considerar que una base de datos no cubre adecuadamente el ámbito temático que declara (porque faltan revistas importantes o no se incluyen completamente, por ejemplo). El sistema que se puede utilizar para medir este indicador consiste en determinar la proporción de revistas consideradas de máximo interés para un área temática que forman parte del conjunto de revistas vaciadas por la base de datos.

- *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. Google ha sido calificado como el buscador que tiene un carácter más internacional, ya que incluye una alta proporción de páginas que no son estrictamente del ámbito anglosajón; este aspecto ha sido muy bien valorado siempre por sus usuarios, que no están tan concentrados en Estados Unidos como pasa con otros motores de búsqueda.

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

- *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.). Se puede valorar el número de campos y la existencia de listas de validación y control. En este apartado, las bases de datos científico-técnicas tienen un importante valor añadido respecto de aquellos buscadores web que no utilizan metadatos.

- *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 indicador 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?

- *Nivel de crecimiento*
En este caso, 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.

#### 6.2.1.3 Actualización

- *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. Es decir, se mide la proporción de información más moderna respecto del conjunto de la base de datos. Un sistema para determinarlo consiste en buscar el número de documentos correspondientes al año en curso que ya están introducidos en la base de datos. Por otro lado, también se tiene que tener en cuenta que no haya vacíos significativos en el curso de los años.

- *Periodo de actualización*
Se refiere a la periodicidad con la que se actualizan los registros de la base de datos. En el caso de las bases de datos de recursos web existe una particularidad que no se presenta en las bases de datos que parten de fuentes o documentos impresos, y reside en el hecho de que prácticamente ningún registro se puede considerar “permanente” o “definitivo”. Así pues, en este caso, idealmente, cada registro tendría que actualizarse cuando se producen cambios en el original, lo que pasa con mucha frecuencia. En la práctica, no obstante, los motores de búsqueda tienen establecidos unos periodos fijos para revisar los documentos ya indizados.

#### 6.2.1.4 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. Hay que hacer notar que, para algunos autores, se trata de un criterio transversal, que puede ser aplicado a otros

indicadores, aunque es una consideración que no tendremos en cuenta. Las inconsistencias facilitan la duplicación de registros (más frecuentes en el caso de catálogos colectivos), ya analizado antes (v. 6.2.1.1) y se pueden diferenciar dos ámbitos de aplicación: en el análisis formal y en el de contenido.

- *Consistencia 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 más problemas en la recuperación. Las inconsistencias en los puntos de acceso son las que provocan que un mismo título de revista esté entrado de dos o tres formas (el de la cubierta, el del lomo o uno de abreviado) o el caso ya conocido de las diversas entradas de los autores latinos en las base de datos anglosajonas.[22] Los términos que constituyen puntos de acceso fundamentales a los registros (autores, 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.

- *Consistencia en el análisis de contenido*
Nos referimos, en este caso, 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. El grado más elevado de inconsistencia se produce cuando se utiliza más de un lenguaje de indización o clasificación en la misma base de datos, lo que sucede en algunos catálogos colectivos. Normalmente, no obstante, las inconsistencias se deben a los cambios de criterio en la aplicación de los lenguajes documentales en el transcurso del tiempo, o a la divergencia de aplicación entre diversos analistas.

### 6.2.2 Sistema de recuperación de la información

La mayoría de los elementos que se pueden traer a colación en este apartado ya han sido analizados en capítulos anteriores.[23] Es por ello que vamos a realizar un somero 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.

[22] Si consultamos la base de datos LISA encontramos a una entrada para Abadal, Ernest y otra Falgueras, Ernest Abadal que corresponden a distintos artículos, el primero de ellos firmado tan sólo con un apellido y el segundo, con los dos apellidos. Este tipo de error distorsiona y dificulta los estudios bibliométricos que se realizan sobre productividad de los autores.

[23] El capítulo 5 para aspectos relacionados con el diseño de la base de datos y los capítulos 3 y 4 para lo que se refiere a las prestaciones del programa de recuperación de la información. En el apartado 3.3. y 3.4 ya se han suministrado numerosos elementos que sirven como indicadores para evaluar los programas de recuperación de la información, y que están relacionados con los distintos módulos que los componen (administración de la base de datos, entrada de datos, indización y recuperación, salida e intercambio).

De hecho, estos criterios son totalmente independientes de los anteriores. De hecho, se da el caso de bases de datos (que son consultables por distintos SRI[24] con lo cual, si se evalúan, hay que precisar cuál es el programa de recuperación que se tiene en consideración.

- *Prestaciones del 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, de mostrar y consultar el tesauro, la búsqueda en lenguaje natural, búsquedas semánticas, etc. (ver cap. 2)

- *Precisión*
Mide la capacidad del SRI 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 (veáse cap. 2).

- *Exhaustividad*
Se trata de la proporción de documentos relevantes que se suministran en respuesta a una determinada petición respecto del total de documentos precisos que existen en la base de datos (veáse cap. 2).

- *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.

- *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 subjectivo (satisfacción del usuario con los resultados).

- *Formatos de visualización*
Posibilidad de seleccionar diferentes formatos ajustados a las necesidades de los usuarios. 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.

- *Interfaz*
El objetivo perseguido es la amigabilidad, es decir, una presentación clara, sencilla, intuitiva, etc. del SRI. Los elementos que forman parte de la interfaz de interacción del usuario han sido sumarizados y descritos en el apartado 4.3: 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.

[24] 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.

### 6.2.3 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 márqueting y promoción. No están directamente relacionados con el contenido ni tampoco con el SRI, sinó que forman parte del nivel administrativo de la base de datos.

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.

- *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. Las *bluesheets*, las famosas "hojas azules", de Dialog son uno de los mejores ejemplos de documentación explicativa sobre la estructura de la base de datos (incluyen, entre otras informaciones, un modelo de registro con indicación de los campos de búsqueda). También es importante que el sistema de ayuda sea contextualizado.

En muchos casos, esta información está estructurada en forma de pregunta-respuesta (FAQs). Tradicionalmente, en el sector de las bases de datos científico-técnicas se precisaban complejos manuales que tenían que ser exactos, precisos y, sobretodo, actualizados. También es frecuente la posibilidad de consultar el tesaurus (impreso o en línea).

- *Atención al usuario*
La existencia de cursos de formación dirigidos a diversos tipos de usuario, o de un servicio más o menos permanente 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.

- *Precios y sistema de facturación*
En este apartado hay que valorar 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 (que si los registros visualizados, los descargafdos, el tiempo, etc.). Se trata de un criterio que, de momento, no se tiene en cuenta para los buscadores web.

- *Sistemas de distribución*
Se tiene que analizar 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.

## 6.3 La perspectiva del usuario

Los indicadores antes reseñados pueden mesurarse 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 mandamos ejecutar una petición de información hasta que nos aparece en pantalla un listado con los resultados. Ahora bien, existe también un tiempo subjetivo que indica cuál es la satisfacción del usuario con ese tiempo de respuesta. Lo mismo podríamos decir del valor o utilidad de los resultados, de los formatos de salida, de la interfaz de consulta, o de la documentación sobre la base de datos, indicadores que podríamos medir en función de esa satisfacción del usuario que antes apuntábamos.

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 existen ejemplos de utilización 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 permiten seguir las acciones que el usuario ha llevado a cabo, pero desconociendo su contexto (la pregunta que se formula) ni sus impresiones (satisfacción, utilidad, etc.).

El objetivo principal de este tipo de estudios es conocer el uso que se hace de una aplicación (sistema de recuperación de la información) para poderla mejorar, enmendar o reorientar. Los anglosajones los denominan estudios de las búsquedas de los usuarios (user searching studies) y se trata de una area de investigación en expansió que se aplicó tradicionalmente a los catálogos en línea (OPACS), las bases de datos científico-técnicas y, más recientemente, al web (*web searching studies*). De esta forma se complementan los aspectos más cuantitavos y operativos que hemos venido tratando hasta ahora.

### 6.3.1 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, opiniones o actitudes de los usuarios respecto a las bases de datos que han consultado (p.e. 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 encuestado, 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 radica 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 inquiere directamente por sus opiniones. Por el contrario, se trata de sistemas lentos y caros (tienen que pasarse personalmente) y no se pueden administrar a un conjunto muy grande de usuarios (en especial, la entrevista).

Vamos a comentar un par de ejemplos de utilización de esta técnica para conocer el grado de satisfacción de usuarios de bases de datos. Tom Wilson (1998) describe los resultados de una encuesta realizada en 1993 por EQUIP sobre la percepción de la calidad en bases de datos en línea. Fue atendida por unos 600 usuarios de 12 países europeos, que rellenaron 989 cuestionarios. Wang et al (1999) también utilizan los cuestionarios para realizar un estudio basado en la metodología SERVQUAL que tiene por objetivo evaluar la calidad de los motores de búsqueda en Internet teniendo especialmente en cuenta la perspectiva del usuario.

También se ha utilizado para conocer la opinión de los productores y distribuidores. Sería el caso del cuestionario sobre calidad y bases de datos que Annick Duflos (1995: 48-78) envió a productores y distribuidores franceses con el objetivo primordial de conocer las políticas de calidad que tenían establecidas.

### 6.3.2 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 que se puede aplicar en situaciones en las que los usuarios no son capaces de responder adecuadamente un cuestionario (p.e. 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.

Un ejemplo claro es el estudio de Raya Fidel (1985), que utilizó la observación para analizar el comportamiento de un grupo de especialistas de distintas áreas temáticas durante sus búsquedas de información. Esta técnica se complementaba con una entrevista al final de la sesión para clarificar aspectos que no habían sido accesibles a la observación de sus acciones. En ocasiones, esta técnica no se utiliza de forma exclusiva sino combinada con cuestionarios o análisis de transacciones. Lo podemos comprobar

en el estudio del uso de Yahoo por parte de niños de Dania Bilal (2001), que tiene por objetivo mejorar el diseño de la interfaz de Yahooligans! mediante el conocimiento de las necesidades de sus usuarios y de su comportamiento. Se utiliza la observación (registran en vídeo el comportamiento de los niños), cuestionarios (para conocer el nivel de experiencia en la consulta al web) y entrevistas (al profesorado). Finalmente, una investigación de Louise T. Su (1992), que tiene por objetivo identificar el mejor criterio de evaluación de una base de datos y que está centrada en el usuario. El estudio se realiza a partir de las preguntas que formulan 40 usuarios reales a distintas bases de datos, tomando como referente una lista de 20 indicadores para la evaluación. La investigación utiliza la observación (graba el proceso de búsqueda) pero también el resto de técnicas descritas en este apartado: cuestionarios autoadministrados, entrevistas, y también análisis de transacciones.

### 6.3.3 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 queden se registran en un fichero.

El análisis de transacciones se ha utilizado de forma extensiva desde hace unos 30 años para evaluar a sistemas de gestión de bibliotecas (la parte pública, los OPACS) y es muy utilizada. Jansen y Pooch (2001) realizan una revisión bibliográfica sobre los diversos trabajos que, sobre estudios de recuperación de información en el web, se han publicado mostrando como la gran mayoría utilizan el análisis transaccional como base para el estudio.

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); duración de la conexión. En la figura 6.1 podemos ver un ejemplo real de la consulta de unas bases de datos de prueba gestionadas con KnosysInternet.

**Figura 6.1. Ejemplo de fichero de logs**

```
QRY 26Abr2000 16:46:09 <- [User: 161.116.140.19 Base: PERALADA]: QUERY 12 .EN Tipologia (Universitat)
QRY 26Abr2000 16:46:13 <- [User: 161.116.140.19 Base: PERALADA]: QUERY 12 .EN Tipologia (Manual)
QRY 26Abr2000 16:46:18 <- [User: 161.116.140.19 Base: PERALADA]: QUERY 12 .EN Tipologia (Llibre de
notes)
QRY 26Abr2000 16:46:22 <- [User: 161.116.140.19 Base: PERALADA]: QUERY 12 .EN Tipologia (Encants)
QRY 26Abr2000 16:46:24 <- [User: 161.116.140.19 Base: PERALADA]: QUERY 12 .EN Tipologia (Capbreu)
QRY 26Abr2000 19:06:44 <- [User: 161.116.140.14 Base: AARTICLE]: QUERY 31 .EN Autor (Deleuze  Gilles)
QRY 26Abr2000 19:06:47 <- [User: 161.116.140.14 Base: AARTICLE]: QUERY 31 .EN Autor (Fortuny Bonet
Francesc)
QRY 26Abr2000 19:06:50 <- [User: 161.116.140.14 Base: AARTICLE]: QUERY 31 .EN Autor (García Calvo
Agustín)
QRY 26Abr2000 19:10:50 <- [User: 161.116.140.11 Base: FOTOCELH]: QUERY 20 paper .Y .EN Col-lecció (celh)
QRY 26Abr2000 19:11:25 <- [User: 161.116.140.11 Base: FOTOCELH]: QUERY 20 .EN Col-lecció (Foto Arxiu
Josep Bonastre)
QRY 26Abr2000 18:37:29 <- [User: 161.116.140.26 Base: SACU]: QUERY 22 .EN Periodicitat (Mensual)
```

```
QRY 26Abr2000 18:37:36 <- [User: 161.116.140.26 Base: SACU]: QUERY 22 .EN Idioma (Anglès)
QRY 26Abr2000 18:37:40 <- [User: 161.116.140.26 Base: SACU]: QUERY 22 .EN Idioma (Català)
QRY 26Abr2000 18:42:09 <- [User: 161.116.140.23 Base: MRR]: QUERY 29 .EN Autor/es (piquero)
QRY 26Abr2000 18:19:37 <- [User: 161.116.140.17 Base: TDUC1]: QUERY 2 .EN Universitat (Universitat de
Barcelona)

ADM 26Abr2000 11:19:57 <- [IP = 161.116.140.56]: ADMINUSER 161.116.140.56
ADM 26Abr2000 11:19:57 <- [User: 161.116.140.56]: REINICIABASES
ADM 26Abr2000 18:06:02 <- [IP = 161.116.140.56]: ADMINUSER 161.116.140.56
ADM 26Abr2000 18:06:02 <- [User: 161.116.140.56]: ABRECERRADAS

SYS 26Abr2000 14:40:19 Cerrando base de datos "ejemplo5"
SYS 26Abr2000 14:40:19 Cerrando base de datos "CONMAR"
SYS 26Abr2000 14:40:19 Cerrando base de datos "TDUC1"
SYS 26Abr2000 14:40:19 Cerrando base de datos "ANTIQ"
SYS 26Abr2000 14:40:19 Cerrando base de datos "CDMCA"
SYS 26Abr2000 18:06:02 Abriendo base de datos "d:\alumnes\39698821\AARTICLE.DOK"
SYS 26Abr2000 18:06:02 Tipo Servidor: Oro
SYS 26Abr2000 19:03:10 Abriendo base de datos "d:\alumnes\53121024\BESCOLA.DOK"
SYS 26Abr2000 19:03:14 Abriendo base de datos "d:\alumnes\52788980\RONYONS.DOK"
SYS 26Abr2000 19:03:14 Abriendo base de datos "d:\alumnes\43541716\CONFE.DOK"
SYS 26Abr2000 19:03:15 Abriendo base de datos "d:\alumnes\46232528\MRR.DOK"
SYS 26Abr2000 19:03:16 Abriendo base de datos "d:\alumnes\46320409\AGRIARXI.DOK"

SYS 26Abr2000 11:20:01 Abriendo base de datos "d:\alumnes\43092696\FOTO.DOK"
ERR 26Abr2000 11:20:02 ERROR: $202 Base corrupta
ERR 26Abr2000 11:20:02 ERROR: No se ha podido abrir "d:\alumnes\43092696\FOTO.DOK"
SYS 26Abr2000 11:20:03 Abriendo base de datos "d:\alumnes\72438163\CLIMURB.DOK"
ERR 26Abr2000 11:20:04 ERROR: Parametros inicializacion incorrectos SELECT_F
SYS 26Abr2000 11:20:08 Abriendo base de datos "d:\alumnes\35028791\DONES.DOK"
ERR 26Abr2000 11:20:08 ERROR: Parametros inicializacion incorrectos SELECT_F
ERR 26Abr2000 18:46:26 ERROR: Parametros inicializacion incorrectos SELECT_FIELDS 33 Y Y TÍTOL
AUTOR EDITORIAL ANY EDICIÓ CIUTAT COL·LECCIÓ DESCR-FÍSICA ISBN MATÈRIA DESCRIPTORS
RESUM
```

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[25] que el usuario formula al sistema.

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.

Se trata de la técnica más simple para recoger datos sobre la interacción usuario-SRI, a distancia, sin necesidad de presencia humana y, por tanto, sin molestar ni condicionar las acciones del usuario. Además, se puede aplicar a un gran número de usuarios, 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.

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

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 SRI. De hecho, tal y como señala Frías (1999) las encuestas y el análisis de transacciones permiten conocer cosas distintas, ya sean las opiniones de los usuarios y sus percepciones sobre el SRI (cuestionario o entrevista) o las acciones llevadas a cabo durante el proceso de consulta.

Dos ejemplos recientes de aplicación los encontramos en Jansen et al (2000), que presentan los resultados del análisis transaccional de casi 52.000 preguntas formuladas por 18.000 usuarios a Excite, y en Baeza et al (2003), un estudio donde se analizan 777.351 consultas realizadas en agosto y septiembre del 2001 en el buscador chileno TodoCL con el fin de complementar los algoritmos de jerarquización tradicionales de resultados.

## 6.4 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 SRI 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 va a permitir 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 también se interesarán por recoger, con las técnicas directas o indirectas que han sido descritas, la utilidad y el grado de satisfacción de los usuarios.

En una reseña de un seminario sobre calidad en bases de datos en CD-ROM que tuvo lugar en 1993 (Casale, 1993: 311), Diane Richards, de *Inspec*, describe varias actuaciones llevadas a cabo por este productor que resumen a la perfección las líneas maestras de lo que se tiene que hacer para garantizar la calidad del contenido de la base de datos. Se trata de una adecuada combinación de sistemas automáticos (ficheros de validación, correctores, etc.), con recursos humanos (personal bien formado para analizar correctamente la información, creando grupos de trabajo específicos para el control de calidad, etc.) y la existencia de un manual de procedimiento que es seguido correctamente por todo el mundo. Sobre este último elemento se detecta una gran coincidencia por parte de muchos autores. Así, p.e. Cerezo et al.(2002:142) señalan también la falta de manuales de procedimiento en los centros productores de bases de datos como uno de los principales motivos que explican la baja calidad en bases de datos, ya que determinadas operaciones se llevan a cabo de formas distintas lo que ocasiona, como ya hemos visto, faltas de consistencia notables en el contenido.

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 recuperación 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 como 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, ¿por qué no hacerlos extensibles a las bases de datos de su departamento de documentación?

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

## 6.5 Bibliografía


Abad, Francisca. "Evaluación de las operaciones de análisis y difusión de la información". En: *Manual de ciencias de la documentación*. Madrid: Síntesis, 2002. p. 671-690.

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

Azorín, Virtudes; Fernández, Fco; Morillo, Matilde. "Evaluación de la calidad en la gestión de bases de datos iconográficas: las fotografías de historia del arte del Centro de Estudios Históricos del CSIC". En: *Actas VI Jornadas Españolas de Documentación*. Valencia: FESABID, 1998. p .127-140.

Auster, E. et al. "A system evaluation of the Educational System for Ontario". *Journal of the ASIS*, vol. 30 (1979), p. 33-40.

Baeza-Yates, Ricardo; Saint-Jean, Felipe. "Análisis de consultas a un buscador y su aplicación a la jerarquización de páginas web". *BiD: textos universitaris de Biblioteconomia i Documentació*, núm. 10 (juny 2003) <http://www2.ub.es/bid/consulta_articulos.php?fichero=10baeza.htm> [Consulta: 27/8/2003].

Banks, Julie. "Are transaction logs useful?: a ten year study". *Journal of Southern Academic and special librarianship*. Vol. 1, no. 3 (2000). <http://southernlibrarianship.icaap.org/content/v01n03/banks_jo1.html> [Consulta: 20/05/02]

Bash, Reva. "Measuring the quality of data: report on the Fourth Annual SCOUG Retreat". *Database searcher*, vol. 6, no. 8 (october 1990), p. 18-23.

Bash, Reva. "Decision points for databases". *Database* (August 1992), p. 46-50.

Bilal, Dania. "Children's use of the Yahooligans! web search engine (II): cognitive and physical behaviors on research tasks", *Journal of the American Society for Information Science and Technology*, vol 52, No. 2 (2001), p. 118-136.

Breeding, Marshall. "Strategies for measuring and implementing e-use". *Library technology reports* (May-June 2002).

Casale, M. "CDROM database quality". *Online&CDROM review*, vol. 17, no. 5 (1993), p. 310-312.

Cerezo, Eva; Alonso, B.; Gómez, Ana. "Evaluación de la calidad en la automatización de bibliotecas", *El profesional de la información*, vol. 11, nº 2 (marzo-abril 2002), p. 141-146.

Chu, H.; Rosenthal, M. “Search engines for the World Wide Web: a comparative study and evaluation methodology”. *ASIS 1996: Annual Conference Proceedings*. ASIS, 1996. <http://www.asis.org/annual-96/electronicproceedings/chu.html> [Consulta: 25/08/2003]

Cleverdon, Cyril W. “The Cranfield test of index language devices”. *Aslib proceedings*, vol 19 (1967), p. 173-192.

Cleverdon, Cyril W. “User evaluation of information retrieval systems”. *Journal of Documentation*, 30 (June 1974), p. 170-180.

Cooper, Michael D. “User patterns of web-based library catalog”. *JASIS*, vol. 52, no. 2 (2001), p. 137-148.

Dervin, B.; Nilan, M. “Information needs and uses”. *Annual review of information scince and technology*, vol. 21 (1986), p. 5-33.

Duflos, Annick. *Las criteres de évaluation des banques de données: le démarche qualité chez las professionnels de l’information électronique*. Paris: ADBS, 1995. 146 p.

Ellis, D. “The physical and cognitive paradigm in information retrieval research”. *Journal of documentation*, vol. 48 no. 1 (1992), p. 45-64.

Extremeño, Ana. “Análisis cualitativo de la base de datos Ecosoc”, *El profesional de la información*, vol. 7, nº 10, (octubre 1998), p. 4-11.

Fidel, Raya. “Towards expert systems for the selection of search keys”. *Journal of the American Society for Information Science*, vol. 37, no. 1 (1986), p. 37-44.

Frías, José Antonio; Martín, Fernando. “El análisis transaccional como técnica de recogida de datos para el estudio del comportamiento de los usuarios del catálogo en línea”. En: Congreso ISKO-España EOCONSID'99 (4º. 1999. Granada). *Actas de las VI Congreso ISKO-España EOCONSID'99 : Representación y Organización del Conocimiento en sus distintas perspectivas : su influencia en la recuperación de información*. Granada: ISKO, 1999. p. 427-434.

Harter, Stephen; Hert, Carol. “Evaluation of information retrieval systems: approaches, isues, and methods”. *ARIST*, vol. 32 (1997).

Ingwersen, P. “Cognitive perspectives of information retrieval interaction: elements of a cognitive IR theory”. *Journal of documentation*, vol. 52, no. 1 (1992), p. 3-50.

Jacsó, Péter. “Content evaluation of databases”. *ARIST*, vol. 32 (1997), p. 231-267.

Jansen, B.J.; Pooch, U. “A review of web searching studies and a framework for future research”. *JASIS*, vol. 52, no. 3 (2001), p. 235-246.

Jansen, B.J.; Spink, A. Saracevic, T. “Real life, real users, and real needs: a study and analysis of user queries on the web”. *Information processing & management*, 36 (2000), p. 207-227.

Johnson, F.C. et al. *DEVISE: a framework for the evaluation of Internet search engines*. Resource: The council for Museums, Archives and Libraries, 2001. <http://www.mmu.ac.uk/h-ss/cerlim/projects/devise/devise-report.pdf> [Consulta: 25/08/03]

Jones, S; Cunningham, S.; McNab, R.; Boddie, S. “A transaction log analysis of a digital library”. <http://www.cs.waikato.ac.nz/~stevej/Research/PAPERS/ijodllogs.pdf> [Consulta: 20/05/02]

Juntunen, R. et al. “Quality requirements for databases: project for evaluating Finnish databases”. En: Online Information Meeting (15th: London, 10-12 December 1991). *Online information 91*. Ed. David Raitt. Oxford: Learned Information, 1991. p. 351-359.

Kemp, Alasdair. *Computer-based knowledge retrieval*. London: Aslib, 1988. Chap. 9 "Evaluation of systems and software. p. 210-227.

Kurth, M. "The limits and limitations of transaction log analysis". *Library Hi Tech*, No. 42 (1993), p. 98-104.

Lancaster, F.W. *Evaluación de la biblioteca*. Madrid: ANABAD, 1996. "Cap XI: Búsquedas en bases de datos". p. 199-238.

Lancaster, F.W.; Sandore, Beth. *Technology management in library and information services*. London: Library Association, 1997. Chap. 14: Evaluation of automated systems". p. 196-225.

Medawar, Katia. "Database quality: a literature review of the past and a plan for the future", *Program*, vol. 29, nº 3 (July 1995), p. 257, 272.

Notes, Greg R. "Tips for evaluating web databases", *Database*, (April-May 1998), p. 69-72.

Olvera, Dolores. "Evaluación de sistemas de recuperación de la información: aproximaciones y nuevas tendencias". *El profesional de la información*, vol 8, nº 11 (noviembre 1999), p. 4-14.

O'Neill, E.T.; Vizine-Goetz, D. "Quality control in online databases". *ARIST*, vol. 23 (1988), p. 125-156.

Ortego, M. del Pilar; Bonal, José Luis. "Indicadores para el control de calidad de bases de datos bibliográficas", En: Jornadas Españolas de Documentación Automatizada (5as: 1996: Cáceres). *V Jornadas Españolas de Documentación Automatizada: sistemas de información: balance de 12 años de jornadas y perspectivas de futuro.* Cáceres: Universidad de Extremadura: ABADMEX, 1996. p. 503-512.

Peters, T.A. "The history and development of transaction log analysis". *Library Hi Tech*, No. 42 (1993), p. 41-66.

Peters, T.a. et al. "An introduction to the special section on transaction log analysis". *Library Hi Tech*, No. 42 (1993), p. 38-40.

Puente, L.; Campo, C. del; Ruiz, M. "Inidcadores de rendimiento para la evaluación de un servicio de bases de datos en línea". *Scire*, vol 7, nº 1 (en.-jun. 2001), p. 89-114.

*The quality of electronic information products and services.* Imo Working paper 95/4. Luxembourg, 1995. 17 p.

Rodríguez Yunta, Luis. "Evaluación e indicadores de calidad en bases de datos", *Revista española de documentación científica*, Vol. 21, nº 1, 1998. p. 9-23.

Rittberger, M.; Rittberger, W. "Measuring quality in the production of databases". *Journal of information science*, vol 23, no. 1 (1997), p. 25-37.

Spinak, E. "Errores ortográficos en el ingreso en bases de datos", *Revista española de documentación científica*, vol. 18, nº 3, 1995, p. 307-319.

Spink, Amanda et al. "Searching the web: the public and their queries". *JASIS*, vol. 52, no. 3 (2001), p. 226-234.

Su, Louise T. "Evaluation measures for interactive information retrieval". *Information processing & management*, vol. 28, no. 4 (1992), p. 503-516.

Van Rijsbergen, C.J. *Information retrieval*. London: Butterworths, 1975. También accesible en: <http://www.dcs.gla.ac.uk/keith/Preface.html> [Consulta: 25/08/2003]

Vickery, B.; Vickery, A. *Information science in theory and practice*. London [etc.]: Bowker-Saur, 1987. 384 p.

Wang, H.; Xie, M.; Goh, T.N. "Service quality of Internet search engines". *Journal of information science*, vol. 25, no. 6 (1999), p. 499-507.

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

Xie, M.; Wang, H.; Goh, T.N. “Quality dimensions of Internet search engines”. *Journal of information science*, vol. 24, no. 5 (1998), p. 365-372.

## 7 Bibliografia global


Abad, Francisca. “Evaluación de las operaciones de análisis y difusión de la información”. En: *Manual de ciencias de la documentación*. Madrid: Síntesis, 2002. p. 671-690.

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

Abadal, E. “Diseño y creación de una base de datos en un medio de comunicación”. En: Fuentes, M.E. *Manual de documentación periodística*. Madrid: Síntesis, 1995. pp. 196-211.

Abadal, Ernest. “Elementos para la evaluación de interfaces de consulta de bases de datos”. *El profesional de la información*, vol. 11, núm. 5 (septiembre-octubre 2002), p. 349-360.

Abadal, E. *Sistemas y servicios de información digital*. Gijón: Trea, 2001, 147 pp.

Abadal, E.; Criach, D.; Cuadrado, M; Gascón, J.; Omella, E. “Centres de Documentació i Biblioteques de Sabadell en xarxa: una iniciativa per a incrementar els serveis de la biblioteca pública al municipi”. Ernest Abadal; [et al.] En: *7es. Jornades Catalanes de Documentació*. Barcelona: Col·legi Oficial de Bibliotecaris-Documentalistes de Catalunya, 1999. p. 127-135.

Abadal, E.; Cuadrado, M; Gascón, J.; Omella, E. “Disseny i creació de bases de dades bibliogràfiques amb CDS/ISIS: l’experiència de SABA-DOC”. *BiD: textos universitaris de biblioteconomia i documentació*. Núm 3 (desembre 1999). <http://www.ub.es/bid/03abadal.htm>

Abadal, Ernest; Martínez, Raúl. “Distribució de bases de dades en el web amb *Knosys Internet”*. *BiD: textos universitaris de biblioteconomia i documentació*, núm 4 (juny 2000). <http://www.ub.es/bid/04abadal.htm>

AENOR. *Norma UNE 50-106-90. Documentación. Directrices para el establecimiento y desarrollo de tesauros monolingües.* Madrid: AENOR, 1990, 47 pp.

Ahlberg, C.; Shneiderman, B. “Visual information seeking: tight coupling of dynamic query filters with starfield displays”. En: *Readings in information visualization: using vision to think*. Ed. by S.K. Card, J.D. Mackinlay, B. Shneiderman. San Francisco: Morgan Kaufmann, 1999. p. 244-250.

Altuna Esteibar, Belén. “Comportamientos de uso y estrategias de búsqueda de los usuarios de catálogos automatizados: breve revisión de la investigación”. En: *Miscelánea homenaje a Luis García Ejarque*. Madrid: Fesabid, 1992. p. 103-111

Arant, Wendi; Payne, Leila. “The common user interface in academic libraries: myth or reality”, *Library Hi Tech*, Vol. 19, No. 1 (2001), p. 63-76.

Archuby, Gustavo. *Wwwisis: manual de procedimientos para bibliotecarios*. [fitxer electrònic]. Versión 0. La Plata, septiembre 1998. 39 p.

Asensi, Viviana; Pastor, Juan Antonio. “Propuesta de un modelo de interfaz genérica para sistemas de recuperación de información”. *Scire*, vol. 4, nº 1 (ene-jun. 1998), p. 71-88.

Ashenfelter, J.P. *Choosing a database for your web site*. New York [etc.]: John Wiley & Sons, 1999. 443 p.

Auster, E. et al. “A system evaluation of the Educational System for Ontario”. *Journal of the ASIS*, vol. 30 (1979), p. 33-40.

Azorín, Virtudes; Fernández, Fco; Morillo, Matilde. “Evaluación de la calidad en la gestión de bases de datos iconográficas: las fotografías de historia del arte del Centro

de Estudios Históricos del CSIC". En: *Actas VI Jornadas Españolas de Documentación*. Valencia: FESABID, 1998. p .127-140.
Baeza-Yates, R.; Ribeiro-Neto, B. *Modern Information Retrieval*. New York: Addison-Wesley, 1999, 513 pp.
Baeza-Yates, Ricardo; Saint-Jean, Felipe. "Análisis de consultas a un buscador y su aplicación a la jerarquización de páginas web". *BiD: textos universitaris de Biblioteconomia i Documentació*, núm. 10 (juny 2003) <http://www2.ub.es/bid/consulta_articulos.php?fichero=10baeza.htm> [Consulta: 27/8/2003].
BAIGET, T. "La distribució de bases de dades a Espanya". En: *3es Jornades Catalanes de Documentació*. Barcelona: SOCADI; COBDC, 1989. p. 101-141.
Baiget, T. *Análisis de sistemas de información*. Barcelona: Institut Català de Tecnologia, 1986, 64 pp. (documento reprografiado).
BAIGET, Tomàs. "25 años de teledocumentación en España", *Revista Española de Documentación Científica*, vol. 21, núm. 4 (1998), p.373-387.
Banks, Julie. "Are transaction logs useful?: a ten year study". *Journal of Southern Academic and special librarianship*. Vol. 1, no. 3 (2000). <http://southernlibrarianship.icaap.org/content/v01n03/banks_jo1.html> [Consulta: 20/05/02]
Bash, Reva. "Decision points for databases". *Database* (August 1992), p. 46-50.
Bash, Reva. "Measuring the quality of data: report on the Fourth Annual SCOUG Retreat". *Database searcher*, vol. 6, no. 8 (october 1990), p. 18-23.
Bechini, Mònica; Burguillos, Ferran; Díaz, Albert. "Confección de categorías y recuperación de la información en Internet". En: Congreso ISKO-España (5º: 2001: Alcalá de Henares). *La representación y organización del conocimiento* [CD-ROM] *: metodologías, modelos y aplicaciones : actas del V Congreso ISKO-España: 25-27 de abril de 2001, Alcalá de Henares (Madrid).* Alcalá de Henares: Sociedad Internacional para la Organización del Conocimiento, Capítulo Español: Facultad de Documentación, Universidad de Alcalá, 2001. [p.404-414].
Belkin, Nicholas J.; Croft, W. Bruce. "Retrieval techniques". *Annual Review of Information Science and Technology*, n. 22, 1987
Beumala, Àngel et al. "Base de datos de recursos Internet científico-técnicos: Ep! (enlaces politécnicos). En: Jornadas Españolas de Documentación (6as: València, 29-21 octubre 1998). *6as Jornadas Españolas de Documentación: los sistemas de información al servicio de la sociedad: actas de las jornadas*. València: Fesabid; Avei, 1998. p. 149-156.
Bilal, Dania. "Children's use of the Yahooligans! web search engine (II): cognitive and physical behaviors on research tasks", *Journal of the American Society for Information Science and Technology*, vol 52, No. 2 (2001), p. 118-136.
Blair, D.C. "The challenge of commercial document retrieval, Part I: Major issues, and a framework based on search exhaustivity, determinacy of representation and document collection size". *Information Processign and Management*, v. 38, 2002, pp. 273-291
Blair, D.C. *Language and representation in information retrieval*. Amsterdam [etc] : Elsevier, 1990. 335 p.
Breeding, Marshall. "Strategies for measuring and implementing e-use". *Library technology reports* (May-June 2002).
Brisaboa, Nieves R., et al. "Sistema de consulta vía web para el Instituto Andaluz de Patrimonio Histórico". En: Jornadas de Bibliotecas Digitales (2ª: Almagro, 2001).

*JBIDI'2001: Jornadas de Bibliotecas Digitales*. [Ciudad Real]: Universidad de Castilla La Mancha, 2001. p. 99-116.
Brusilovsky, P. "Methods and techniques of adaptive hypermedia", *User Modeling and User Adapted Interaction. Special issue on adaptive hypertext and hypermedia*, Pittsburgh, 1996.
Buckland, M. *Information and information systems*. Westport: Greenwood Pres, 1991, 225 pp.
Card, S.K.; Mackinlay, J.D.; Shneiderman, B. *Readings in information visualization: using vision to think*. San Francisco: Morgan Kaufmann, 1999.
Casale, M. "CDROM database quality". *Online&CDROM review*, vol. 17, no. 5 (1993), p. 310-312.
Celma, Matilde; Casamayor, J.C.; Mota, L. *Bases de dades relacionals*. València: Universitat Politècnica de València, 1998. p. 1-20.
Cerezo, Eva; Alonso, B.; Gómez, Ana. "Evaluación de la calidad en la automatización de bibliotecas", *El profesional de la información*, vol. 11, nº 2 (marzo-abril 2002), p. 141-146.
Checkland, P. B. *Systems thinking, systems practice*. Chichester: Wiley, 1981.
Checkland, P. B.; Scholes, J. *Soft systems methodology in action*. Chichester: Wiley, 1990.
Chen, P.P-S. "The entity-relationship model: towards a unified view of data". *ACM transactions on databases systems*, v. 1, n. 1, 1976, pp. 9-36.
Choo, Chun Wei; Detlor, Brian; Turnbull, Don. "Information seeking on the web: an integrated model of browsing and searching", *First Monday*, Vol. 5, no. 2 (February 2000). <firstmonday.org/issues/issue5_2/choo/index.html>. [Consulta: 25/09/01]
Choo, Chun Wei; Detlor, Brian; Turnbull, Don. *Web work: information seeking and knowledege work on the world wide web*. Dordrecht [etc.]: Kluwer Academic, 2000. "Chapter 5 Models of information seeking on the world wide web", p. 133-158.
Chorafas, D. N. *Intelligent multimedia databases: from object orientation and fuzzy engineering to intentional database structures*. Englewood Cliffs, New Jersey: Prentice Hall, 1994, 360 pp.
Chowdhury, G.G. *Introduction to modern information retrieval*. London: Library Asociation, 1999, 451 pp.
Chowdhury, G.G.; Chowdury Sudatta. *Introduction to digital libraries*. London : Facet, 2003. Chap. 8: Information access and user interfaces, p. 152-177.
Chu, H.; Rosenthal, M. "Search engines for the World Wide Web: a comparative study and evaluation methodology". *ASIS 1996: Annual Conference Proceedings*. ASIS, 1996. <http://www.asis.org/annual-96/electronicproceedings/chu.html> [Consulta: 25/08/2003]
Cleverdon, Cyril W. "The Cranfield test of index language devices". *Aslib proceedings*, vol 19 (1967), p. 173-192.
Cleverdon, Cyril W. "User evaluation of information retrieval systems". *Journal of Documentation*, 30 ( June 1974), p. 170-180.
Codina, L. "Metodología de análisis de sistemas de información y diseño de bases de datos documentales: aspectos lógicos y funcionales". En: Baró, J.; Cid, P. (eds.). *Anuario SOCADI de Documentación e Información 1998*. Barcelona: SOCADI, 1998, pp. 195-210.
Codina, L. "Recuperación de información e hipertextos: sus bases lógicas y su aplicación a la documentación periodística". En: Fuentes, M. Eulália (ed.). *Manual de Documentación periodística*. Madrid: Síntesis, 1995, p. 213-230.

Codina, L. "Sistemas automáticos de recuperación de información textual". En: Gomez Guinovart, J. *Aplicaciones lingüísticas de la informático.* Santiago de Compostela: Tórculo, 1994, pp. 63-86.
Codina, L. "Teoría de recuperación de información: modelos fundamentales y aplicación a la gestión documental". *Information world en español*, n. 38, octubre 1995, p. 18-22.
Codina, L. “Metodología de creación de bases de datos documentales”. Parte I. *Information world en español*, n. 33, abril 1995; Parte II. *Information world en español*, n. 34, mayo 1995;
Codina, Lluís. "Evaluación de recursos digitales en línea: conceptos, indicadores y métodos", *Revista española de documentación científica*, vol. 23, nº 1, (enero-marzo 2000), p. 9-44.
Codina, Lluís. "Parámetros e indicadores de calidad para la evaluación de recursos digitales". En: *VII Jornadas Españolas de Documentación. La gestión del conocimiento: retos y soluciones de los profesionales de la información*. Bilbao: Universidad del País Vasco, 2000. p. 135-144.
Codina, Lluís. “Cómo funcionan los servicios en Internet: un informe especial para navegantes y creadores de información (I)”, *El profesional de la información*, vol 7, nº 5, mayo 1997, p. 22-27.
Codina, Lluís. *Sistemes d’informació documental: concepció, anàlisi i disseny de sistemes de gestió documental amb microordinadors*. Barcelona: Pòrtic, 1993.
Codina, Lluís; Abadal, Ernest. “Gestió documental amb microordinadors: característiques, estructura i tecnologia dels sistemes de gestió documental”. *Item*, núm. 11, 1992, p. 72-100.
Connolly, T.M.; et al. *Database systems: a practical approach to design, implementation and management*. Wokingham [etc.]: Addison-Wesley, 1995. .
Cooper, Michael D. “Design considerations in instrumenting and monitoring web-based information retrieval systems”, *Journal of the ASIS*, vol. 49, no. 10 (1998), p. 903-919.
Cooper, Michael D. “User patterns of web-based library catalog”. *JASIS*, vol. 52, no. 2 (2001), p. 137-148.
Crestani, Fabio, Funte, P. de ; Vegas, J**.** “Diseño de una interfaz de consulta para la recuperación de documentos estructurados”. En: Jornadas de Bibliotecas Digitales (2ª: Almagro, 2001). *JBIDI’2001: Jornadas de Bibliotecas Digitales*. [Ciudad Real]: Universidad de Castilla La Mancha, 2001. p. 85-97.
Curras, E. *La información en sus nuevos aspectos.* Madrid: Paraninfo, 1988, 307 p.
De Groote, Sandy. “PubMed, Internet Grateful Med, and Ovid: a comparison of three Medline Internet interfaces”, *Medical reference services quarterly*, vol. 19, no. 4, winter 2000, p. 1-13.
Dervin, B.; Nilan, M. “Information needs and uses”. *Annual review of information scince and technology*, vol. 21 (1986), p. 5-33.
*DESIRE information gateways handbook* [en línea]. DESIRE, c1999-2000, last updated 26 April 00. <http://www.desire.org/handbook/>. [Consulta: 27/04/2001].
Duflos, Annick. *Las criteres de évaluation des banques de données: le démarche qualité chez las professionnels de l’information électronique*. Paris: ADBS, 1995. 146 p.
Eíto, Ricardo. “Sistemas GED e indizadores intranet: ¿alternativas excluyentes o tecnologías complementarias?”. *El profesional de la información*, vol 7, nº 9, septiembre 1998, p. 5-9.

Ellis, D. “The physical and cognitive paradigm in information retrieval research”. *Journal of documentation*, vol. 48 no. 1 (1992), p. 45-64.

Ellis, D. *New horizons in information retrieval*. London: The Library Asociation, 1990, 138 pp.

Ellis, David; Ford, Nigel; Furner, Jonathan. "In search of the unknown user: indexing, hypertext and the world wide web", *Journal of documentation*, January 1998, 54 (1), 28-47.

Espelt, Constança. “Improving subject retrieval: user-friendly interfaces and effectiveness”, *BiD: textos universitaris de biblioteconomia i documentació*, núm 1 (juny 1998). <http://www.ub.es/bid/01espel1.htm> [Consulta: 27/09/01]

Extremeño, Ana. “Análisis cualitativo de la base de datos Ecosoc”, *El profesional de la información*, vol. 7, nº 10, (octubre 1998), p. 4-11.

Fernández, Mª Jesús; Angós, José Mª; Salvador, José A. “Interfaces de usuario: diseño de la visualización de la información como medio para mejorar la gestión del conocimiento y los resultados obtenidos por el usuario”. En: Congreso ISKO-España (5º: 2001: Alcalá de Henares). *La representación y organización del conocimiento* [CD-ROM]*: metodologías, modelos y aplicaciones: actas del V Congreso ISKO-España: 25-27 de abril de 2001, Alcalá de Henares (Madrid).* Alcalá de Henares: Sociedad Internacional para la Organización del Conocimiento, Capítulo Español: Facultad de Documentación, Universidad de Alcalá, 2001. [p.506-517].

Fidel, Raya. “Towards expert systems for the selection of search keys”. *Journal of the American Society for Information Science*, vol. 37, no. 1 (1986), p. 37-44.

Fidel, Raya. *Database design for information retrieval: a conceptual approach*. New York [etc.]: John Wiley & Sons, 1987. 232 p.

Figuerola, Carlos G.; Alonso, José L.; Zazo, Angel F. "Diseño de un motor de recuperación de la información para uso experimental y educativo", BiD, 4, juny 2000. <http://www.ub.es/bid/04figue.htm>

Fox, Edward A. (1987). “Recuperación de información: investigación de nuevas posibilidades”. En: *CD-ROM: el nuevo papiro*. Madrid: Anaya, 1987.
Frakes, W. B.; Baeza-Yates, R. (eds). *Information retrieval: data structures & algorithms*. Englewod Cliffs: Prentice Hall, 1992, 504 p.

Fox, Edward A. et al. “Users, user interface, and objects: Envision, a digital library”. *JASIS*, vol. 44 no. 8 (1993), p. 480-491.

Frants, Valery I.; Shapiro, J.; Voiskunskii, Vladimir. *Automated information retrieval: theory and methods.* San Diego [etc.]: Academic Press, 1997. Cap 6. Automatic indexing of documents, p. 136-165.

Frías, José Antonio; Martín, Fernando. “El análisis transaccional como técnica de recogida de datos para el estudio del comportamiento de los usuarios del catálogo en línea”. En: Congreso ISKO-España EOCONSID'99 (4º. 1999. Granada). *Actas de las VI Congreso ISKO-España EOCONSID'99 : Representación y Organización del Conocimiento en sus distintas perspectivas : su influencia en la recuperación de información*. Granada: ISKO, 1999. p. 427-434.

García Figuerola, Carlos. "La recuperación de información en colecciones documentales multilingües". En: Pinto, María; Cordón, José A. *Técnicas documentales aplicadas a la traducción*. Madrid: Síntesis, 1999. p. 129-142

García Marco, Francisco J. “De la consulta de catálogos a la gestión de información: tensiones hacia el cambio en el diseño de OPACS”, *Boletín de la ANABAD* (1991), p. 325-334.

García Marco, Francisco J. ”Interfaces amigables para la recuperación de la información bibliográfica”, *Scire*, vol. 1, nº 1 (ener.-jun. 1995), p. 127-148.

Gil Leiva, Isidoro. *La automatización de la indización de documentos*. Gijón: Trea, 1999. 221 p.

Gillman, Peter (ed.). *Text retrieval: the state of the art*. London: Taylor Graham, 1990, 208 p.

*Guidelines for OPAC displays*. Prepared for the IFLA Task Force on Guidelines for

Gutiérrez de Mesa, José A.; Hilera, José R. "Generación de documentación hipermedia en Internet a partir de información multimedia en bases de datos", *Cuadernos de documentación multimedia*, núms.6-7, 1997-1998, p. 135-140.

Harter, Stephen; Hert, Carol. "Evaluation of information retrieval systems: approaches, isues, and methods". *ARIST*, vol. 32 (1997).

Head, Alison J. *Design wise: a guide for evaluating the interface design of information resources.* Medford: CiberAge Books, 1999. 196 p.

Hearst, Marti A. "Sistemas para consultar la red", *Investigación y ciencia* (mayo 1997), p. 44-49.

Hearst, Marti A. "User interfaces and visualization". En: Baeza-Yates, Ricardo; Ribeiro-Neto, B. *Modern information retrieval*. New York: ACM; Harlow: Addison-Wesley, 1999. p. 257-323.

Herrero, Víctor. "La conexión con bases de datos Microisis a través del World Wide Web", *Bol. Anabad*, 2, abril-junio 1998, p. 309-316.

Hildreth, Charles R. "Beyond boolean: designing the next generation of online catalogs", *Library trends*, (Spring 1987), p. 647- 667.

Information systems on-line course. INSY312 Introduction to database design: Lesson 1: Overview of database management systems. Mercer University. <http://mumc.mercer.edu/etris/dbless1.htm> [Consulta: 25/01/1999]

Ingwersen, P. "Cognitive perspectives of information retrieval interaction: elements of a cognitive IR theory". *Journal of documentation*, vol. 52, no. 1 (1992), p. 3-50.

Jackson, G. A. *Introducción al diseño de bases de datos relacionales*. Madrid: Anaya, 1990, 203 pp.

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

Jansen, B.J.; Pooch, U. "A review of web searching studies and a framework for future research". *JASIS*, vol. 52, no. 3 (2001), p. 235-246.

Jansen, B.J.; Spink, A. Saracevic, T. "Real life, real users, and real needs: a study and analysis of user queries on the web". *Information processing & management* , 36 (2000), p. 207-227.

Johnson, F.C. et al. *DEVISE: a framework for the evaluation of Internet search engines*. Resource: The council for Museums, Archives and Libraries, 2001. <http://www.mmu.ac.uk/h-ss/cerlim/projects/devise/devise-report.pdf> [Consulta: 25/08/03]

Joint, Nicholas. "Designing interfaces for distributed electronic collections: the lesons of traditional librarianship". *Libri*, vol 51 (2001), p. 148-156.

Jones, S; Cunningham, S.; McNab, R.; Boddie, S. "A transaction log analysis of a digital library". <http://www.cs.waikato.ac.nz/~stevej/Research/PAPERS/ijodllogs.pdf> [Consulta: 20/05/02]

Juntunen, R. et al. "Quality requirements for databases: project for evaluating Finnish databases". En: Online Information Meeting (15th: London, 10-12 December 1991). *Online information 91*. Ed. David Raitt. Oxford: Learned Information, 1991. p. 351-359.

Kemp, A. *Computer-based knowledge retrieval*. London: Aslib, 1988. 399 p.

Kemp, Alasdair. *Computer-based knowledge retrieval*. London: Aslib, 1988. Chap. 9 "Evaluation of systems and software. p. 210-227.
Kowalski, G. *Information retrieval systems: theory and implementation.* Boston: Kluwer, 1997, 282 pp.
Kurth, M. "The limits and limitations of transaction log analysis". *Library Hi Tech*, No. 42 (1993), p. 98-104.
Lancaster, F. W. *Indexing and abstracting in theory and practice*. Champaing (IL): University of Illinois, 1998, 412 pp.
Lancaster, F.W. *Evaluación de la biblioteca*. Madrid: ANABAD, 1996. "Cap XI: Búsquedas en bases de datos". p. 199-238.
Lancaster, F.W.; Sandore, Beth. *Technology management in library and information services*. London: Library Association, 1997. Chap. 14: Evaluation of automated systems". p. 196-225.
Leloup, Catherine. *Motores de búsqueda e indización*. Barcelona: Gestión 2000, 1998. 287 p.
Lewis, P. *Information systems development*. London: Pitman,1994, 260 pp.
Lim, Edward. "Pasarelas temáticas del Sudeste Asiático: análisis de sus métodos de clasificación [en línia]". En: IFLA Council and General Conference (65a: Bangkok: 1999). *Conference proceedings.* <http://ifla.inist.fr/IV/ifla65/papers/011-117s.htm>. [Consulta: 16/07/00].
Losee Jr., R.M. *The science of information*. San Diego: Academic Pres, 1990, 293 pp.
Penrose, R. *La nueva mente del emperador*. Madrid: Mondadori, 1991, 597 p.
Marchionini, Gary. *Information seeking in electronic environments*. Cambridge: Cambridge University, 1995. 224 p.
Marchionini, Gary; Komlodi, A. "Design of interfaces for information seeking", *Annual review of information science and technology*, Vol. 33, (1998), p. 89-130.
Marchionini, Gary; Plaisant, C.; Komlodi, A. "Interfaces and tools for the Library of Congress National Digital Library Program", *Information processing & management*, Vol. 34, No. 5 (1998), p. 535-555.
Marcos Mora, Mari Carmen. "Motores de recuperación de la información: un análisis comparativo (parte II)", *El profesional de la información*, vol 7, nº 3, marzo 1998, p. 13-19.
Marcos, Mari-Carmen. "HCI (human computer interaction): concepto y desarrollo", *El profesional de la información*, vol. 10, nº 6 (junio 2001), p. 4-16.
Marcos, Mari-Carmen. "Interacción persona-ordenador en las interfaces de rcuperación de información". En: Jornadas Españolas de Documentación (8as: Barcelona, 6-8 febrero 2003). *Fesabid 2003: los sistemas de información en las organizaciones: eficacia y transparencia*. Barcelona: Fesabid, 2003. p. 463-476.
Martínez, Victoria. "Un modelo para el uso de internet en los centros de información juvenil", *El profesional de la información*, vol 8, nº 7-8, julio-agosto 1999, p. 22-33.
Medawar, Katia. "Database quality: a literature review of the past and a plan for the future", *Program*, vol. 29, nº 3 (July 1995), p. 257, 272.
Miguel, Adoración de; Piattini, Mario. *Fundamentos y modelos de bases de datos*. Madrid: RA-MA, 1997.
Miguel, Adoración de; Piattini, Mario. *Fundamentos y modelos de bases de datos*. Madrid: Ra-Ma, 1997.
Moya, Félix de. *Los sistemas integrados de gestión bibliotecaria: estructuras de datos y recuperación de información*. Madrid: Anabad, 1995. p. 113-132.

Moya, Félix. “Técnicas avanzadas de recuperación documental”. En: López Yepes, J. *Manual de ciencias de la documentación*. Madrid: Pirámide, 2002.

Moya, Félix; Herrero, Víctor. “Investigaciones en curso sobre interfaces gráficos en dos y tres dimensiones para el acceso a la información electrónica”, *Cuadernos de documentación multimedia*, nº 8, (1999) <www.ucm.es/info/multidoc/multidoc/revista/num8/moya.html>. [Consulta: 25/09/01]

Muñoz, Jesús E. “Bancos de imágenes: evaluación y análisis de los mecanismos de recuperación de imágenes”, *El profesional de la información*, vol. 10, nº 3 (marzo 2001), p. 4-18.

Nackerud, Shane A. “The potential of CGI: using pre-built CGI scripts to make interactive web pages”, *Information technology and libraries*, december 1998, p. 222-229.

Nielsen, Jakob**.** *Usabilidad: diseño de sitios web*. Madrid [etc.]: Prentice Hall, 2000. “Opciones de búsqueda”, p. 224-245.

Nieuwenhuysen, P. “Criteria for the evaluation of text storage and retrieval software”. *The electronic library*, vol 6, no.3 (june 1980), p. 160-166.

Notes, Greg R. "Tips for evaluating web databases", *Database*, (April-May 1998), p. 69-72.

O’Neill, E.T.; Vizine-Goetz, D. “Quality control in online databases”. *ARIST*, vol. 23 (1988), p. 125-156.

Olvera, Dolores. “Evaluación de sistemas de recuperación de la información: aproximaciones y nuevas tendencias”. *El profesional de la información*, vol 8, nº 11 (noviembre 1999), p. 4-14.

OPAC Displays by Martha Yee. Draft version. November 24, 1998.

Ortego, M. del Pilar; Bonal, José Luis. “Indicadores para el control de calidad de bases de datos bibliográficas”, En: Jornadas Españolas de Documentación Automatizada (5as: 1996: Cáceres). *V Jornadas Españolas de Documentación Automatizada: sistemas de información: balance de 12 años de jornadas y perspectivas de futuro*. Cáceres: Universidad de Extremadura: ABADMEX, 1996. p. 503-512.

Osborne, L.; Nakamura, M. *Systems analysis for librarians and information professionals*. 2nd ed. Englewood, CO: Libraries Unlimited, 2000. 261 pp.

Pastor, Juan A. et al. “Proyecto SABIO: Sistema de Acceso a Bases de Información Organizada”. En: Jornadas Españolas de Documentación (6as: València, 29-21 octubre 1998). *6as Jornadas Españolas de Documentación: los sistemas de información al servicio de la sociedad: actas de las jornadas*. València: Fesabid; Avei, 1998. p. 695-702.

Peña, Rosalía. *Gestión digital de la información: de bits a bibliotecas digitales y la web*. Madrid: Ra-ma, 2002. p. 105-115.

Peters, T.A. “The history and development of transaction log analysis”. *Library Hi Tech*, No. 42 (1993), p. 41-66.

Peters, T.a. et al. “An introduction to the special section on transaction log analysis”. *Library Hi Tech*, No. 42 (1993), p. 38-40.

Puente, L.; Campo, C. del; Ruiz, M. “Inidcadores de rendimiento para la evaluación de un servicio de bases de datos en línea”. *Scire*, vol 7, nº 1 (en.-jun. 2001), p. 89-114.

Puig Torné, J. *Proyectos informáticos: planificación, desarrollo y control*. Madrid: Paraninfo, 1994.

*Readings in information visualization: using vision to think*. Ed. by S.K. Card, J.D. Mackinlay, B. Shneiderman. San Francisco: Morgan Kaufmann, 1999.

Rijsbergen, C. J. van (1981). *Information Retrieval*. Londres: Butterworths, 1981

Rittberger, M.; Rittberger, W. “Measuring quality in the production of databases”. *Journal of information science*, vol 23, no. 1 (1997), p. 25-37.

Rodríguez Muñoz, José V.; Saorín, Tomàs. “Modelado documental de servicios de información en web”, *El profesional de la información*, vol. 7, nº 9 (septiembre 1998), p. 10-18.

Rodríguez Yunta, Luis. “Evaluación e indicadores de calidad en bases de datos”, *Revista española de documentación científica*, Vol. 21, nº 1, 1998. p. 9-23.

Rowley, Jennifer. “The controlled versus natural indexing languages debate revisited: a perspective on information retrieval practice and research”, *Journal of information science*, 20, 2, 1994, p. 108-119.

Sabin-Kildiss, Luisa; Cool, C.; Xie, H. “Assessing the functionality of web-based versions of traditional search engines”, *Online* (march-april 2001), p. 18-26.

Saffady, William. “Text retrieval products for libraries”, *Library technology reports*, vol. 36, no. 2 (march-april 2000), p. 7-16.

Salton, G. *Automatic text procesing: the transformation, analysis, and retrieval of information by computer*. Reading (MA): Addison-Wesley, 1989 , 530 p.

Salton, G.; McGill, M. J. *Introduction to modern information retrieval.* New York: McGraw-Hill, 1983 , 448 pp.

*Search engine watch*. Danny Sullivan, editor. Internet.com, 1996-1999. <http://searchenginewatch.com>. [Consultat: gener 1999].

Searle, John R. (1990). “¿Es la mente un programa informático?”. *Investigación y Ciencia*, nº 162, marzo 1990.

Shneiderman, B. “Dynamic queries for visual information seeking”. En: *Readings in information visualization: using vision to think.* Ed. by S.K. Card, J.D. Mackinlay, B. Shneiderman. San Francisco: Morgan Kaufmann, 1999. p. 236-243.

Shneiderman, Ben; Don Byrd and W. Bruce Croft. "Clarifying search: a user-interface framework for text searches", *D-Lib Magazine*, January 1997. <www.dlib.org/dlib/january97/retrieval/01shneiderman.html> [Consulta: 25/09/01]

Sieverts, E.G. et al. “Software for information storage and retrieval tested, evaluated anc compared: Part 1 – General introduction”. *The electronic library*, vol 9, no.3 (1991), p. 145-154.

Sieverts, E.G. et al. “Software for information storage and retrieval tested, evaluated anc compared: Part 2 – Classical retrieval systems”. *The electronic library*, vol 9, no.6 (1991), p. 301-316.

Sieverts, E.G. et al. “Software for information storage and retrieval tested, evaluated anc compared: Part 4 – Indexing and full-text retrieval programs”. *The electronic library*, vol 10, no.4 (1992), p. 195-206.

Sinclair, J.; McCullough. *Creación de bases de datos en Internet*. Madrid: Anaya Multimedia, 1997. 504 p.

Soergel, D. *Organizing information: principles of data base and retrieval systems.* Orlando: Academic Pres, 1985, 450 pp.

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

Sparck Jones, K; Willett, P. *Readings in information retrieval*. San Francisco: Morgan Kaufmann, 1997.

Spinak, E. “Errores ortográficos en el ingreso en bases de datos*”, Revista española de documentación científica*, vol. 18, nº 3, 1995, p. 307-319.

Spink, Amanda et al. “Searching the web: the public and their queries”. *JASIS*, vol. 52, no. 3 (2001), p. 226-234.

Stein, Lincoln D. *How to set up and maintain a web site*. Reading [etc.]: Addison-Wesley, 1997. 793 p.
Su, Louise T. "Evaluation measures for interactive information retrieval". *Information processing & management*, vol. 28, no. 4 (1992), p. 503-516.
Tenopir, Carol. "Full text database retrieval performance", *Online review*, 1985, vol. 9, No. 2, p. 149-163.
Tenopir, Carol; Lundeen, Gerald. *Managing your information: how to design and create a textual database on your microcomputer*. New York: Neal-Schuman, 1988. 226 p.
*The quality of electronic information products and services.* Imo Working paper 95/4. Luxembourg, 1995. 17 p.
Tittel, Ed et al. *La biblia de la programación CGI*. Madrid: Anaya Multimedia, 1997. 734 p.
Underwood, P. G. *Soft systems analysis and the management of libraries, information services and resource centres*. London: Library Association, 1996, 198 pp.
Van Rijsbergen, C.J. *Information retrieval*. London: Butterworths, 1975. También accesible en: <http://www.dcs.gla.ac.uk/keith/Preface.html> [Consulta: 25/08/2003]
Van Slype, G. *Los lenguajes de indización: concepción, construcción y utilización en los sistemas documentales.* Madrid: Fundación Germán Sánchez Ruipérez, 1991, 198 p.
Vickery, B.; Vickery, A. *Information science in theory and practice*. London [etc.]: Bowker-Saur, 1987. 384 p.
Vizine-Goetz, Diane. "Using library classification schemes for Internet resources [en línia]". En: OCLC Internet Cataloging Project Colloquium (1996: San Antonio, Texas). *Proceedings of the OCLC Internet Cataloging Colloquium*. [Dublin, Ohio]: OCLC, 1996. <http://www.oclc.org/oclc/man/colloq/v-g.htm>. [Consulta: 23/07/00].
Walker, D.W. *Sistemas de información basados en ordenador.* Barcelona: Marcombo, 1991
Wang, H.; Xie, M.; Goh, T.N. "Service quality of Internet search engines". *Journal of information science*, vol. 25, no. 6 (1999), p. 499-507.
Warren, Scott. "Visual displays of information: a conceptual taxonomy". *Libri*, vol. 51 (2001), p. 135-147.
Willitts, John. *Database design and construction: an open learning course for students and information managers*. London: Library Association, 1992. XXII, 425 p.
Wilson, T.D. "EQUIP: a european survey of quality criteria for the evaluation of databases, *Journal of information science*, 24 (5), 1998, p. 345-357.
Xie, M.; Wang, H.; Goh, T.N. "Quality dimensions of Internet search engines". *Journal of information science*, vol. 24, no. 5 (1998), p. 365-372.
Yourdon, E. *Análisis estructurado moderno*. México: Prentice-Hall Hispanoamericana, 1993, 735 pp.