23 Sistemas de información documental: concepto, modelo, estructura y organización

Purificación Moscoso

23.1. CONCEPTO DE SISTEMA DE INFORMACIÓN DOCUMENTAL

El estudio y análisis del concepto «sistema de información documental» ha desempeñado un papel fundamental en el desarrollo de nuestra disciplina. Se trata, además, de un concepto clave de los fundamentos teóricos en los que se sustenta la materia objeto de esta parte cuarta del manual, ya que para que la información documental pueda recuperarse según las necesidades de los distintos tipos de usuarios debe almacenarse y organizarse en sistemas cuya estructura se adecue a las características de esta clase específica de información, y cuyo motor de búsqueda posibilite obtener medidas de eficacia satisfactorias.

Ahora bien, este concepto entronca directamente con la noción de sistema desarrollada desde la Teoría General de Sistemas. Quien se considera en la actualidad el padre de esta teoría, el profesor de biología Ludwig von Bertalanffy, que la presentó en 1937 en el Seminario Filosófico de Charles Morris de la Universidad de Chicago, definió sistema como:

> Un conjunto de elementos interrelacionados entre ellos y con su entorno. El aspecto realmente importante es la interacción entre los elementos para crear un todo, un sistema dinámico. Dicho sistema, si se trata de un sistema abierto, interactúa con su entorno[1].

En realidad, un sistema no es un simple conjunto de elementos, ni la suma de todos ellos. Lo que lo caracteriza, realmente, son las relaciones que enlazan los ele-

[1] Bertalanffy, L. (1972): «The history and status of General Systems Theory», *Academy of Management Journal*, núm. 15, p. 417.

mentos entre sí, de manera que lo que ocurre para un elemento repercute en los demás. De esta forma, a la suma de los elementos, el sistema añade las múltiples relaciones que los enlazan, así como las acciones y reacciones de unos elementos sobre otros. Además, la relación con el entorno influye de manera decisiva en el comportamiento y la actividad del sistema.

A este respecto, se habla de entornos generales y específicos. Los primeros afectan a todas las organizaciones de una sociedad determinada, y se caracterizan por valores tales como el estado del desarrollo de la tecnología en un momento dado, o los sistemas políticos, sociales, culturales y educativos, por ejemplo. Los entornos específicos son los que influyen sobre sistemas u organizaciones concretas, y en ellos cabría considerar, por ejemplo, los usuarios o la tecnología utilizada.

Los sistemas de información son un tipo específico de sistemas cuya estructura, organización y función hacen posible transformar datos en información e información en datos, entendiendo por información el dato que altera el estado del sistema que lo percibe. Así, el concepto de sistema de información implica la noción de transformación, ya que se trata de estructuras que llevan a cabo una actividad, función u operación y son capaces de transformar elementos de entrada en elementos de salida.

Según Debons, Horne y Cronenweth, cabría definir el término sistema de información como:

> «Un conjunto de personas, máquinas y procedimientos que aumentan el potencial biológico humano para adquirir, procesar y actuar sobre los datos[2].»

Los sistemas de información se encuentran condicionados, y condicionan, a su vez, los entornos generales y específicos en los que se hallan inmersos. Y una vez más, es preciso incidir en el análisis de esta clase de sistemas dentro del contexto global de la sociedad de la información, sociedad en la que la información ha adquirido una nueva dimensión, al tratarse del principal recurso intangible de todo tipo de organizaciones.

Los sistemas de información son, además, sistemas complejos, dada la cantidad y diversidad de elementos implicados, así como de relaciones generadas entre éstos. Su capacidad de cambio, característica imprescindible en estos tipos de estructuras, se la confiere su capacidad de respuesta y de adaptación al entorno. Por ello, el diseño de un sistema de información no sólo requiere el análisis de los datos, sino también el del entorno y el de sus elementos.

Lo explicado hasta el momento nos permite delimitar el sintagma «sistema de información» a partir de las siguientes premisas. En primer lugar, en tanto que sistemas, se compone de elementos unidos mediante alguna forma de interacción regulada, de manera que se estructuran en un todo organizado. Asimismo, tienen una

---

[2] Debons, A.; Horne, E., y Cronenweth, S. (1988): *Information Science: An Integrated View*, Boston, MA: Hall & Co. p. 14.

finalidad, que consiste en organizar recursos informativos y en resolver necesidades concretas de información. Para ello, es preciso que exista un conjunto de instrucciones que dirijan la actividad o función que posibilita alcanzar el objetivo para el que han sido diseñados. Y, por último, este conjunto de instrucciones permite llevar a cabo toda una serie de operaciones y procedimientos.

Ahora bien, el estudio del concepto de sistema de información se ha abordado desde perspectivas muy diversas, ya que se trata de un término que hace referencia a modelos de distintas clases. Así, por ejemplo, en el mundo anglosajón los acrónimos MIS *(Management Information Systems)* e IMS *(Information Management Systems)* aluden a los sistemas de gestión de la información utilizados en los procesos de toma de decisiones dentro de las organizaciones, y a los sistemas de gestión de la información cuya finalidad es gestionar los recursos informativos, tanto los generados en las propias organizaciones como los procedentes de los entornos externos. Por otra parte, se habla también de sistemas de información para referirse a las redes y centros de información en el marco de las políticas nacionales e internacionales de información, aspecto éste al que se dedica el capítulo 10 de este manual.

El objeto de este capítulo es un modelo determinado de sistemas de información, los sistemas de información documental, cuyo concepto, obviamente, se desarrolla a partir de la noción de sistema y, en concreto, de sistema de información, pero cuyas características estructurales y funcionales le confieren una identidad concreta y diferenciada de otros tipos de sistemas.

En el contexto de la Documentación, información es un concepto que hace referencia, por una parte, al proceso, al acto de comunicación, que modifica el estado del conocimiento humano. Por otra, información es la representación concreta y tangible del conocimiento. Por ello, en nuestra disciplina, los sistemas de información documental se consideran sistemas cognitivos, ya que tienen la capacidad de transformar el estado del conocimiento, tanto desde un punto de vista general como individual.

Así pues, los sistemas de información documental se ocupan de la representación concreta y tangible de la información, entendida ésta como proceso de acceso y adquisición de conocimiento. Son, además, sistemas de símbolos mediante los cuales se representa lo que en terminología informática se denomina «mundo real» o «mundo objeto», y para lo cual hacen uso de un aparato conceptual que permite expresar las características de los elementos que conforman el mundo al que hacen referencia.

23.2. APARATO CONCEPTUAL DE LOS SISTEMAS DE INFORMACIÓN DOCUMENTAL

Los sistemas mediante los cuales se gestiona la información que generan, producen o captan los distintos tipos de organizaciones se desarrollan siguiendo, básí-

camente, dos modelos: el relacional y el documental. El desarrollo de uno u otro va a depender del objetivo para el cual se implemente el sistema, del tipo de información de que se trate, así como de las necesidades específicas de quienes harán uso del sistema.

En ambos casos se trata de estructuras organizadas cuya finalidad es transformar datos en información e información en datos. Ahora bien, estos datos pueden ser de naturaleza muy diversa. Pueden referirse, por ejemplo, a los clientes de una distribuidora de libros, o a las ventas realizadas por una tienda de discos, o al consumo de la energía eléctrica suministrada por una compañía. Existen también datos que hacen referencia a lo que se ha publicado sobre una materia concreta, por ejemplo, el cambio climático como consecuencia de la emisión de gases de efecto invernadero, o aquellos que describen las principales características de distintas páginas web. Es obvio que se trata de información de características muy distintas, y cuyo tratamiento, por consiguiente, exige el uso de procedimientos diferentes.

Ahora bien, aunque se trate de modelos claramente diferenciados en cuanto a la forma de estructurar y organizar los datos, así como a los mecanismos para acceder a ellos, ambos comparten un aparato conceptual común que permite representar el mundo o el entorno al que hacen referencia. Los objetos materiales o conceptuales del mundo real que se representa son las entidades, que tienen necesariamente que ser identificables: personas, organizaciones, artículos de publicaciones científicas, noticias de prensa o páginas web, por ejemplo. Las características que poseen las entidades en el mundo real se denominan atributos, de forma que cada entidad en particular está representada por «el valor de sus atributos», que equivale al contenido de un campo concreto. Cada una de las entidades representadas es distinta de las otras, diferencia que se plasma por medio de los valores de los atributos.

Así, por ejemplo, si el sistema de información representa los recursos electrónicos accesibles en Internet sobre una determinada materia, las entidades se corresponderían con páginas web, artículos de carácter científico, revistas electrónicas, etc. Los atributos serían aquellos elementos que permiten caracterizarlos, describirlos y analizarlos para su posterior identificación y recuperación. En este caso, por ejemplo, los atributos podrían ser: el título, el autor o creador, la materia, la fecha, el tipo de recurso, el identificador y la lengua, por ejemplo.

Entidades y atributos son conceptos, que se representan con datos. Cada una de las entidades del mundo real se corresponde con un registro, que es la unidad de información básica de los sistemas de información. Los registros se componen de campos, que representan los atributos que caracterizan las entidades. El contenido de cada campo es su valor, por lo que cada entidad se representa por el valor de sus atributos.

Y siguiendo con el ejemplo anterior, cada recurso electrónico quedaría representado e identificado por el valor de sus atributos, como se ejemplifica en la figura 23.1.

Cada uno de los valores de un atributo se denomina campo. Para una entidad particular el conjunto de campos se denomina registro, que, como ya se ha explicado, puede referirse a un recurso electrónico, a un paquete de software, a un artículo científico, dependiendo del mundo real que represente el sistema de información.

Atributos = Campos	Valor de los atributos = Valor de los campos
Título: Nombre dado al recurso por el creador o editor	DC Metadata User Guidelines
Autor o creador: Persona(s) u organización(es) principales responsables del contenido intelectual del recurso	Hansen, Preben
Materia o palabras clave: Tema del recurso, normalmente expresado con palabras clave o frases que describen el contenido del recurso	Dublin Core, metadatos, guía de uso
Fecha: Fecha asociada a la creación o disponibilidad del recurso. Puede ser una fecha concreta o un rango.	1998-01-16
Tipo de recurso: Categoría del recurso: página web, informe, artículo, revista, etc.	Guía de uso
Identificador del recurso: Número utilizado para identificar de forma única y exclusiva el recurso: URL o ISBN, por ejemplo	*http://www.sics.se/~preben/DC/DC_guide.html*
Lengua: Lengua(s) del contenido intelectual del recurso	Inglés

Figura 23.1. Atributos y valores de atributos para la entidad *DC Metadata User Guidelines.*

Los datos se pueden estructurar siguiendo una jerarquía, que iría del bit, al byte, al subcampo, al registro, al fichero y a la base de datos (figura 23.2). La unidad menor reconocible es un bit, que se representa por 0 y 1. Ocho bits forman un byte,

Bit	0.1
Byte	8 bits = 1 carácter escrito
Subcampo	Conjunto de bytes
Campo	Conjunto de subcampos
Registro	Conjunto de campos
Fichero	Conjunto de registros
Base de datos	Conjunto de ficheros

Figura 23.2. Jerarquía de los datos.

que representa un carácter escrito. Un campo o un subcampo es el conjunto de caracteres que representa el valor de un atributo para la entidad considerada. Un campo puede estar compuesto por dos o más subcampos, o por ninguno. Por ejemplo, en el registro MARC, la mención de publicación de una monografía se considera un campo donde existen cinco subcampos: lugar de publicación, editorial, fecha de publicación, lugar de impresión e imprenta. La totalidad de campos y subcampos que describen los atributos de una entidad es el registro, que representa dicha entidad.

La estructura de los registros de una base de datos varía según la información tratada y las necesidades de los usuarios que van a acceder a ella. Así, la información relacionada con la biomedicina difiere de la que recoge una base de datos sobre patentes. Por consiguiente, los campos de los que se componen los registros no tienen por qué coincidir ni total ni parcialmente.

De esta forma, como se muestra en la figura 23.3, se establece una equivalencia entre mundo real y simbólico, entidades y registros, atributos y campos, y valor de los atributos y de los campos.

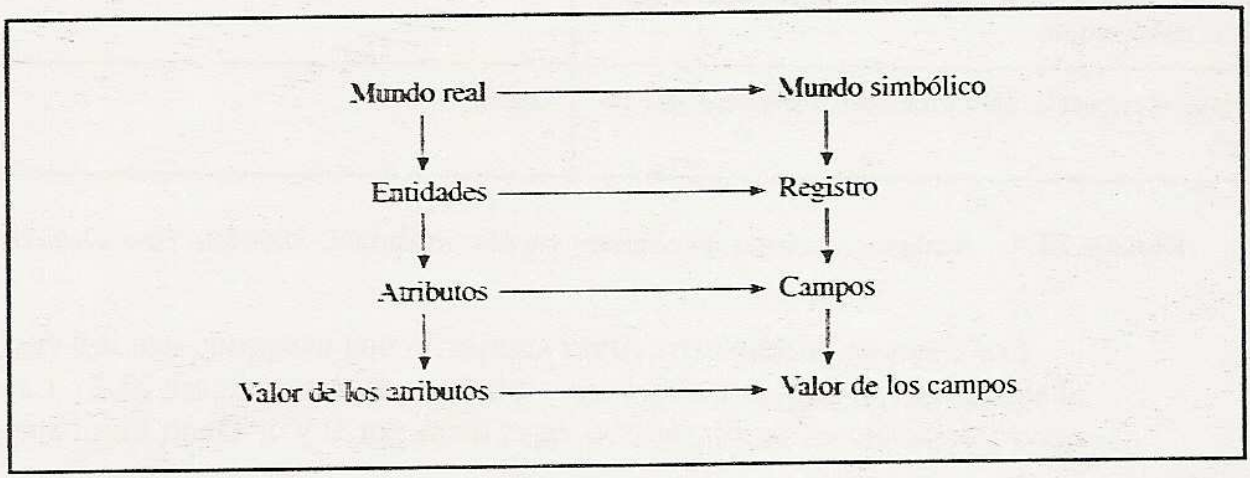
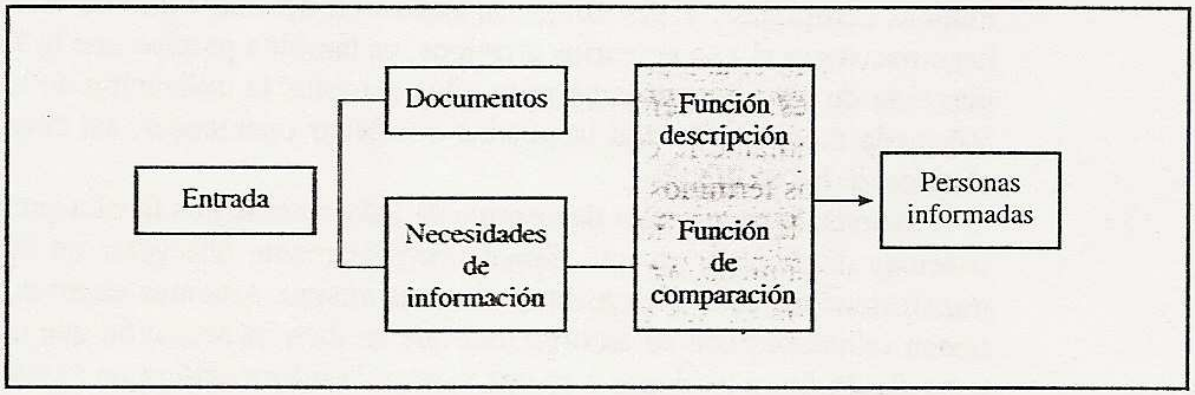


Figura 23.3. Equivalencia entre mundo real y mundo simbólico.

23.3. MODELO CONCEPTUAL DE LOS SISTEMAS DE INFORMACIÓN DOCUMENTAL

Siguiendo a Willitts, un sistema de información documental es un sistema que acepta como entradas documentos cognitivos y necesidades de información, y que produce como salida personas informadas. Como sistema de información que es, en su interior tiene lugar un proceso de transformación de las entradas en salidas, que consiste en una doble función de descripción y comparación[3] (figura 23.4).

[3] Recogido en Codina, L.: *Sistemas de gestión documentales: estado del arte y estrategias de utilización (I)*, *Binary*, junio 1994, núm. 62, pp. 114-119.



FUENTE: Adaptado de Willits, J. (1992): *Database Design and Construction,* Londres: Library Association.

Figura 23.4. Modelo conceptual de un sistema de información documental.

La función de descripción es la que permite identificar las entidades representadas en el sistema, y, por consiguiente, acceder a los documentos conforme a los elementos utilizados en dicha descripción. Para ello, es preciso que cada documento se represente de acuerdo con aquellas características que, por una parte, lo diferencian de los demás, y, por otra parte, lo relacionan con el resto. De esta forma, los usuarios pueden recuperar no sólo un documento concreto, sino también conjuntos de documentos. Así pues, los campos en los que se estructuran los registros tienen que corresponderse con atributos de carácter descriptivo, título, autor o fecha, por ejemplo, así como de índole analítica, que serían aquellos cuyos valores representan el contenido de los documentos.

Por otra parte, la función de comparación es la que hace posible relacionar las necesidades de información de los usuarios con un documento concreto o con un conjunto de ellos. Y para ello, es necesario que los sistemas de información documental cuenten con motores de búsqueda capaces de comparar los términos o frases a través de los cuales se expresa la demanda informativa con los utilizados para describir y/o analizar los documentos, mediante mecanismos de confrontación, exacta o parcial. Es necesario, igualmente, que estos sistemas posibiliten a los usuarios unir, combinar y delimitar términos y frases de búsqueda a través de operadores y mecanismos específicos, como se explicará en el apartado quinto de este capítulo.

Es claro, por consiguiente, que como sistemas de información que son llevan a cabo una función que permite alcanzar el objetivo para el que han sido diseñados, que en este caso consiste en satisfacer las necesidades de información de los usuarios identificando los documentos susceptibles de cubrir las demandas expresadas. Y como afirma Codina, el factor que marca el auténtico rendimiento de un sistema de información documental es su capacidad para recuperar documentos de acuerdo con sus atributos semánticos, ya que los documentos cognitivos se caracterizan precisamente porque son semánticamente no triviales, ya que presentan estructuras te-

máticas complejas[4]. Y puesto que la expresión del contenido de estos documentos hace necesario el uso de varios términos, es también preciso que la lógica de recuperación de estos sistemas permita a los usuarios la utilización de una sintaxis de búsqueda mediante la cual se puedan combinar operadores, así como delimitar el alcance de los términos.

Además, la propia idea de sistema de información nos lleva a entenderlos como sistemas dinámicos, ya que tienen necesariamente que estar en armonía con la transformación de sus elementos. Son, asimismo, sistemas abiertos, ya que mantienen relaciones con su entorno, del que reciben información que alimenta al sistema desde fuera, y al que a su vez vierten la información que se transforma y genera dentro del propio sistema. La normalización de formatos de almacenamiento, así como los distintos softwares de conversión de formatos, hacen posible la comunicación entre los distintos sistemas a través de la importación y exportación de registros. Asimismo, la actual infraestructura de redes permite el acceso desde cualquier sistema a cualquier sistema con independencia del hardware y software del que se trate.

De la misma manera, no cabe entender un sistema de estas características como sistema cerrado, ya que ninguno puede considerarse, en ningún caso, autosuficiente ni tampoco deslindado del entorno al que pertenece. A este respecto Debons y Montgomery hablan de entornos compuestos por personas, equipos y procedimientos organizados para mejorar los objetivos específicos de la información[5].

En un sistema de información documental la información se estructura en una base de datos, que consiste en un conjunto de datos almacenados en soporte informático y organizados de tal forma que puedan recuperarse de determinadas maneras, de acuerdo con las necesidades expresadas en la estrategia de búsqueda. El proceso de recuperación de estos datos se lleva a cabo por sistemas informáticos, y para que la información sea accesible conforme a los fines estipulados, es necesario describir y analizar los documentos según sus características específicas. Lo que condiciona la organización de los datos que se van a almacenar y procesar, así como las formas en las que éstos se podrán recuperar, es, por un lado, la naturaleza de la información y, por otro lado, las características y necesidades del colectivo que va a hacer uso de estos datos.

En el diseño, mantenimiento y explotación de una base de datos documental, desempeña un papel fundamental el denominado «diccionario de datos», que consiste en una lista de todos los campos de una base de datos con una declaración de su dominio y de cualquier otra circunstancia que afecte al comportamiento del campo. A este respecto, el concepto de campo indizado es especialmente importante en el modelo documental, ya que una de las características de estos sistemas es su ca-

---

[4] Codina, L.: *Sistemas de gestión documentales: estado del arte y estrategias de utilización (I)*, p. 116.

[5] Debons, A., y Montgomery, K. L. (1974): «Design and evaluation of information systems», *Annual Review of Information Science Technology*, vol. 9, p. 66.

pacidad de generar índices, característica que, a su vez, hace posible la recuperación de información de manera precisa y exhaustiva según requiera la demanda informativa. Un campo indizado es aquel cuyo valor o valores forman parte de un índice. En el modelo documental, en principio, todos los valores de un campo pueden utilizarse como entradas de un índice.

Estos índices, llamados ficheros o índices inversos, consisten en una tabla de atributos o características acompañada de una colección que enumera consecutivamente toda entidad asociada a los atributos de la tabla. De esta forma, un índice o fichero inverso consta de un conjunto de anotaciones, una para cada registro, que expresan los valores de los campos por el cual se puede recuperar la información así como el lugar que ocupa cada uno de los términos o frases dentro del campo en cuestión, y un puntero, que permite el acceso inmediato a dicho registro. Se trata, por consiguiente, de un fichero en el que cada registro se corresponde con cada una de las palabras o frases que han sido objeto de indización, con un campo que recoge información sobre la localización de término, el tipo de campo, el lugar que ocupa dentro de este campo, etc.

Como ya se ha explicado, cualquier campo puede servir de base para construir un índice. Los softwares utilizados para el diseño, gestión y mantenimiento de sistemas de información documental difieren entre sí según estén capacitados para generar un único índice o varios. Algunos tienen la capacidad de crear índices independientes para los distintos campos. Y una de las decisiones fundamentales consiste en determinar de qué forma se va a procesar la información de cada campo en el fichero inverso, ya que ello determinará las posibilidades de búsqueda.

Los ficheros inversos interactúan con lo que se denomina índice de palabras vacías, que se genera automáticamente y cuya función consiste en recoger todas aquellas palabras carentes de significado relevante de cara a la recuperación de documentos, tales como artículos, preposiciones o conjunciones, y asociarlas al lugar que ocupan en cada uno de los registros.

Para entender el concepto de índice o fichero inverso basta pensar en el índice de materias de los libros impresos, por ejemplo, en los que cada una de las materias va acompañada de la página o páginas en las que se trata o menciona.

En el diseño de un sistema de información documental es fundamental decidir, primero, qué campos son susceptibles de que sus valores formen parte de un índice, y, segundo, la forma en la que dichos campos se van a indizar, ya que ésta determina las posibilidades de búsqueda en la base de datos.

Básicamente existen dos formas de indización: por palabras y por frases. En la primera, cada palabra individual se considera como una cadena de caracteres separadas por blancos y/o signos de puntuación. La segunda forma de indización se utiliza para los campos cuyos datos están sujetos a algún tipo de normalización. Este procedimiento implica que una secuencia de dos o más palabras se incluyen, tal y como aparecen en el campo, en el índice pertinente, respetando los espacios y signos de puntuación de dicha secuencia. La indización por palabras permite el acceso a los datos por lenguaje libre, mientras que la indización por frases requiere el

uso de lenguajes controlados; esto es, la utilización de la forma aceptada para una autor, una fecha o una materia, por ejemplo.

La combinación de ambas formas de indización es la que en la actualidad se utiliza en mayor medida, ya que posibilita el acceso a la información tanto por lenguaje libre como por lenguajes controlados. Recordamos que si bien el acceso por lenguaje libre permite una mayor versatilidad a la hora de recuperar documentos, tiene como consecuencia un mayor porcentaje de ruido en los resultados obtenidos. Por el contrario, el uso de lenguajes controlados, que por lo general hace posible obtener una mayor pertinencia en las búsquedas, puede conllevar silencio en los resultados.

Por último, cabe decir que un sistema de información documental está formado por tres elementos. En primer lugar, por la propia información, estructurada en forma de base de datos. En segundo lugar, por el software que permite el diseño, gestión y mantenimiento de la base de datos, así como la recuperación de los documentos descritos y analizados. En tercer lugar, por el software de interfaz, que es el que determina y condiciona la comunicación entre los usuarios y el sistema.

Ahora bien, se trata de tres elementos que son también componentes básicos de todo sistema de información. Sin embargo, el objeto de los sistemas de información documental es un tipo de información muy concreta, que cabría caracterizarla como bibliográfica documental y que agrupa a lo que hemos denominado documentos cognitivos, cuyas propiedades, así como su ritmo exponencial de producción, hacen necesario contar con softwares de gestión y recuperación específicos. Se trata de softwares cuyas prestaciones y funcionalidades permiten diferenciarlos claramente de los motores que subyacen en sistemas no documentales. En cuanto a las características del tercero de los elementos señalados, la interfaz, éstas deben estar íntimamente relacionadas con las de los usuarios, profesionales o finales, para quienes se diseña el sistema.

A continuación se van a dedicar sendos apartados a explicar las características de las bases de datos y de los motores de recuperación de los sistemas de información documental.

23.4. CARACTERÍSTICAS DE LAS BASES DE DATOS DOCUMENTALES

Los documentos cognitivos se componen, fundamentalmente, de caracteres alfabéticos, al igual que los elementos que se utilizan para describirlos y analizarlos. Los valores de los atributos de las entidades representadas en los sistemas de información documental son, básicamente, palabras. Por consiguiente, los motores de recuperación de información documental deben tener la capacidad y habilidad para manipular grandes cadenas de caracteres, así como para proporcionar flexibles y sofisticados mecanismos de búsqueda.

Es usual, no obstante, incluir datos numéricos para referirse a la fecha de creación o publicación de un recurso, aunque las necesidades de información de los usua-

rios de estos sistemas no requieren operaciones matemáticas complejas, más allá de la simple función de comparación para conocer, por ejemplo, qué existe publicado sobre una materia entre dos fechas concretas, a partir de una fecha determinada, en un año concreto o con anterioridad a éste.

Una de las características fundamentales de la información documental es el ritmo vertiginoso que dirige su producción. Así, por ejemplo, el crecimiento exponencial de la producción de información científico-técnica precisa desarrollar sistemas de información capaces de gestionar nuevos documentos a un ritmo mucho mayor del que se necesita en entornos no documentales. Por tanto, en comparación con otros tipos de bases de datos, las documentales contienen gran cantidad de registros. El objetivo de muchos de estos sistemas es controlar lo accesible sobre un determinado campo de conocimiento: ciencias de la salud, educación o sociología, por ejemplo.

Asimismo, el proceso de descripción y análisis de los documentos cognitivos requiere la generación de registros con múltiples campos, puesto que este tipo de entidades necesitan de varios elementos que representen sus características. En ocasiones, para identificar un recurso pueden ser suficientes los datos sobre su autor o creador, el título o el localizador. Sin embargo, la utilidad de estos sistemas reside, fundamentalmente, en proporcionar otros tipos de datos que permitan al usuario discernir sobre su adecuación a la demanda que expresa. Por ejemplo, las palabras o frases que representan su contenido o un resumen sobre el mismo. Por otra parte, las demandas de información, en general, imponen condiciones del tipo fecha, clase o lengua del recurso. En consecuencia, para facilitar el acceso a la información a través de búsquedas exhaustivas y/o pertinentes, es necesario que los registros de estas bases de datos consten de múltiples campos.

Se precisa, igualmente, que el software permita una longitud variable de los campos, así como la repetición de valores. Cuando se representan otro tipo de entidades es posible prever la longitud de los campos, o al menos la máxima. Por ejemplo: un año se compone de cuatro o dos dígitos, según se represente; el DNI de una persona consta de ocho dígitos y una letra, el código postal, de cinco. Por el contrario, el título de un documento puede variar entre ocho y doscientos caracteres.

Asimismo, otra de las características de la información documental es que varios de los elementos que sirven para describirla e identificarla se repiten. Por ejemplo, es muy común que una monografía tenga más de un autor, o que para representar su contenido tengamos que utilizar más de un descriptor. Esto hace que sea necesario que los campos que componen los registros admitan lo que se conoce con el nombre de valores repetibles, de forma que cada uno de estos valores tenga la misma importancia a la hora de recuperar el documento, por lo que debe ser susceptible de formar parte de un índice.

Del mismo modo, la multiplicidad de puntos de acceso es otra característica de estos sistemas, así como la capacidad de generar índices inversos, que como se ha explicado en el punto anterior consiste en una tabla de atributos o características acompañada de una colección que enumera consecutivamente toda entidad asociada a los atributos de la tabla.

La función principal de los sistemas de información documental no es permitir acceder a un documento del que se conocen sus datos, sino posibilitar recuperar conjuntos de documentos de acuerdo con determinadas características que satisfagan demandas informativas concretas. Tales demandas se expresan por medio de estrategias que combinan elementos de descripción y análisis, y/o conceptos relevantes del tema en cuestión. Los usuarios de estos sistemas conforman un amplio abanico de características muy diversas, por lo que es preciso que el sistema permita diferentes puntos de acceso. En realidad, se hace cada vez más deseable que los usuarios puedan buscar por cualquiera de las palabras de cada uno de los campos del registro, si bien esta posibilidad produce ruido en los resultados de las búsquedas. Una de las mayores ventajas de la información electrónica es la posibilidad de acceder a ella de manera no lineal, y para ello es preciso dotar a los documentos de múltiples puntos de acceso.

En una base de datos documental los registros se introducen de manera secuencial, y forman lo que se denomina «fichero lineal», que es el conjunto de registros de la base de datos. Ahora bien, puesto que el acceso a estos registros se realiza a través de distintos puntos, el orden de introducción de los mismos carece de toda importancia a la hora de recuperar la información. El acceso a la información se realiza a través de los índices inversos que, como ya se ha explicado, son característicos de este tipo de sistemas.

Por último, cabe señalar que la uniformidad de los registros de las bases de datos es también propio de estos sistemas, puesto que la descripción de las entidades se puede llevar a cabo con los mismos atributos. Así, todos los documentos tienen un título, la mayoría un autor o creador, de la gran mayoría conocemos también la fecha, y su contenido se puede resumir y/o describir utilizando descriptores.

23.5. LOS MOTORES DE BÚSQUEDA DE LOS SISTEMAS DE INFORMACIÓN DOCUMENTAL

Los motores de búsqueda que subyacen en un modelo documental son los llamados sistemas de almacenamiento y recuperación de información, conocidos bajo el acrónimo anglosajón IRS, que responde a la expresión *Information Retrieval System*. Estos sistemas tienen unas propiedades, funcionalidades y prestaciones determinadas que los diferencian de otros tipos de software de gestión de la información, y como afirma el profesor Moya, su concepción:

> Parte del principio de que la información procesable por un sistema informático se organiza basándose en documentos. La consecuencia inmediata de esta idea es que los IRS se desarrollan con el múltiple objetivo de almacenar, recuperar y mostrar grandes cantidades de documentos, entendidos éstos como secuencias más o menos extensas de caracteres que se agrupan formando palabras, que se agrupan en frases, que a su vez conforman párrafos y que, por fin, en número variable, componen los documentos[6].

Según Ashford, presentan como funcionalidades básicas las siguientes[7]:

— Diseño y modificación de las estructuras de los documentos que formarán la base de datos.
— Recuperación de conjuntos de documentos.
— Salida formateada de documentos recuperados.
— Control de los términos indizados.
— Mantenimiento de los documentos en la base de datos.
— Sistemas de recuperación y autoarranque.
— Gestión de tipos de usuarios mediante niveles de autorización de los mismos.
— Monitorización del sistema.

Son apropiados para la gestión y recuperación de información documental porque su estructura física de datos permite la definición de estructuras lógicas complejas, sin limitación en la extensión de sus elementos y con múltiples repeticiones en cada uno de ellos. Además, la lógica que rige el acceso al contenido de la base de datos se basa en operadores que permiten combinar conceptos expresados mediante palabras o frases, por lo que son capaces de recuperar documentos de acuerdo con sus atributos semánticos.

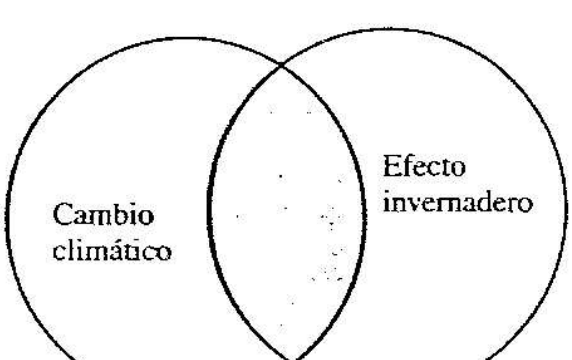
Se trata, fundamentalmente, de dos tipos de operadores: los denominados operadores booleanos y los de proximidad. Los primeros deben su nombre al matemático George Boole, precursor de la lógica simbólica y del álgebra de conjuntos. Los segundos, también llamados operadores de adyacencia, permiten paliar algunas de las limitaciones que impone el álgebra de Boole en la recuperación de información.

Los operadores booleanos se utilizan para representar relaciones entre conceptos, relaciones que se expresan, básicamente, a través de tres operadores: intersección (AND o Y, según se utilice la nomenclatura inglesa o española), unión (OR u O) y exclusión (NOT o NO). El uso de este tipo de operadores se basa en los principios del álgebra de Boole, de manera que las relaciones entre conceptos se expresan como relaciones entre conjuntos, dando como resultado un conjunto de documentos que, en principio, reúnen las condiciones impuestas en la estrategia de búsqueda.

El operador de intersección AND se utiliza cuando se requiere recuperar registros que contengan los términos expresados en la demanda informativa. Así, por ejemplo, si se desea encontrar documentos que traten del cambio climático causado por la emisión de gases de efecto invernadero, la estrategia de búsqueda se expresaría de la siguiente forma: «Cambio climático» AND «Efecto invernadero». El resultado de la búsqueda sería la intersección del conjunto A (cambio climático) con el B (efecto invernadero), que equivaldría a los registros indizados mediante los dos descriptores «cambio climático» y «efecto invernadero» (figura 23.5).

---

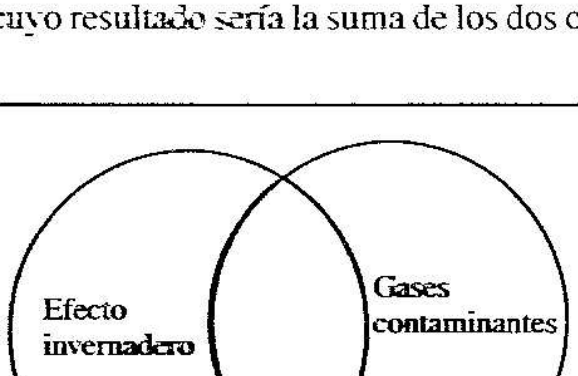
[7] Moya, F. de: *Los sistemas integrados de gestión bibliotecaria*, p. 113.
[7] Ashford, J. H. (1984): «Information storage and retrieval systems on mainframe and minicomputers». *Program*, vol. 18, pp. 124-146.



Conjunto A: Documentos indizados con el descriptor «Cambio climático».
Conjunto B: Documentos indizados con el descriptor «Efecto invernadero».
Conjunto C: Documentos indizados con los descriptores «Cambio climático» y «Efecto invernadero».

Figura 23.5. Conjunto de documentos resultado de la búsqueda «Cambio climático» AND «Efecto invernadero».

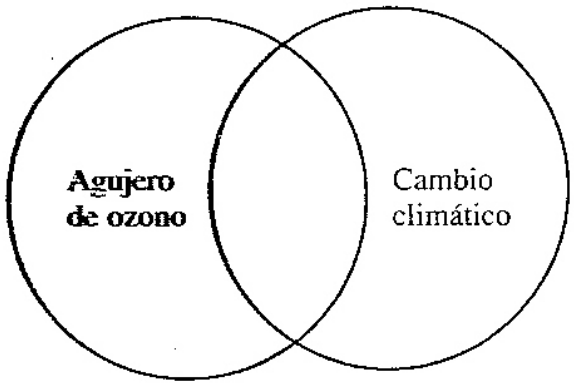
El operador de unión o suma, OR, se utiliza para recuperar el conjunto de registros que contengan cualquiera de los términos expresados. Por ejemplo, aquellos que se refieran a documentos que aborden el tema del efecto invernadero o de la emisión de gases contaminantes; búsqueda que se expresaría «Efecto invernadero» OR «Gases contaminantes», y cuyo resultado sería la suma de los dos conjuntos (figura 23.6).



Conjunto A: Documentos indizados con el descriptor «Efecto invernadero».
Conjunto B: Documentos indizados con el descriptor «Gases contaminantes».
Conjunto C = A + B: Documentos indizados con cualquiera de los dos descriptores.

Figura 23.6. Conjunto de documentos resultado de la búsqueda «Efecto invernadero» OR «Gases contaminantes».

En cuanto al operador de exclusión o resta, NOT, su uso excluye los registros en los que aparece el término o frase precedido por NOT. Por ejemplo, registros referidos a documentos que traten el fenómeno del agujero en la capa de ozono, pero que no hablen también del cambio climático: «Agujero de ozono» NOT «Cambio climático» (figura 23.7). Se trata, no obstante, de un operador complejo de utilizar, ya que con frecuencia excluye documentos relevantes porque contienen el término que en la estrategia trata de excluirse.



Conjunto A: Documentos indizados con el descriptor «Agujero de ozono».
Conjunto B: Documentos indizados con el descriptor «Cambio climático».
Conjunto C = A – B: Documentos indizados con el descriptor «Agujero de ozono» pero no con «Cambio climático».

Figura 23.7. Conjunto de documentos resultado de la búsqueda «Agujero de ozono» NOT «Cambio climático».

Obviamente, los operadores pueden combinarse entre sí, dando lugar a estrategias más complejas. Por ejemplo («Efecto invernadero» OR «Gases contaminantes») AND («Cambio climático» OR «Calentamiento del planeta» OR «Agujero de ozono»).

Ahora bien, aunque se trata de los operadores más utilizados en las búsquedas documentales, éstos presentan problemas importantes a la hora de obtener medidas de eficacia satisfactorias: el uso del operador OR con frecuencia da lugar a ruido, mientras que la utilización de AND puede generar silencio. Por ello, los motores de búsqueda de los sistemas de información documental permiten la utilización de operadores de proximidad, que tienen en cuenta el lugar que ocupan las palabras empleadas en la estrategia de búsqueda dentro del contexto, dado que la cercanía de los términos es relevante a la hora de expresar un determinado concepto. Asimismo, el uso de estos operadores puede favorecer la obtención de una mayor pertinencia, ya que provocan, en general, menos ruido.

Los operadores de proximidad permiten al usuario expresar la cercanía entre los términos, así como el orden en el que éstos deben aparecer. De esta forma, la utilización del operador de proximidad «con»: efecto(w)invernadero, permitiría recuperar todos aquellos registros en cuyos campos aparecieran ambos términos unidos. Asimismo, el uso del operador «cerca»: cambio(n) clima?, resultaría en el conjunto de documentos en los que aparecieran ambos términos, sin que importe el orden. Por ejemplo, documentos titulados *Los estudios sobre el clima reflejan que el cambio es cada vez más evidente*. Además, algunos sistemas de recuperación permiten también especificar el número de caracteres máximo que debe mediar entre los términos.

Estos sistemas de recuperación hacen posible, asimismo, especificar el campo que debe contener los términos utilizados en la estrategia de búsqueda: cambio(n)clima/título, de forma que se recuperarían todos aquellos documentos cuyo título contenga el término cambio y clima, sin que importe el orden en el que éstos aparezcan. Por ejemplo: *Clima y vegetación: los cambios originados como consecuencia del efecto invernadero*. o *El cambio en el clima occidental, una realidad cada vez más preocupante*.

Los operadores de proximidad son de gran valor para acceder al contenido de la base de datos utilizando palabras del lenguaje natural, lo que permite llevar a cabo búsquedas en los campos de título y resumen, así como en el de materias cuando se desconocen las formas normalizadas.

Los motores de búsqueda de los sistemas documentales brindan también la posibilidad de utilizar los llamados operadores de cálculo o de rango, si bien las operaciones que pueden llevar a cabo sobre campos como el de fecha son muy limitadas. Además, permiten el uso de técnicas de truncamiento, a derecha, izquierda o en el medio, dependiendo de la sofisticación del programa, lo que hace posible obtener una mayor exhaustividad en las búsquedas realizadas. Así, por ejemplo, truncando clima?, el usuario recuperaría registros con los términos clima, climas, climático y climáticos.

Y una de las prestaciones fundamentales de los motores de búsqueda de estos sistemas es que permiten redefinir las estrategias de búsqueda, reutilizando y/o modificando los operadores utilizados.

Por otra parte, en la actualidad cada vez es más frecuente que los sistemas de recuperación de información incorporen capacidades y prestaciones propias de los entornos hipertextuales, así como mecanismos de confrontación parcial *(best match)* y de relevancia *(relevance ranking)*.

El hipertexto es un modelo basado en la idea de que el pensamiento humano funciona mediante asociaciones, aprovechando las ventajas que proporciona el almacenamiento electrónico de información para solventar las limitaciones impuestas por la naturaleza del texto impreso. El hipertexto es, en definitiva, una forma de organizar y gestionar la información, textual, sonora y gráfica, de forma no lineal. Esto es, basándose en asociaciones. Esta capacidad de asociación permite crear enlaces entre distintos nodos de información, de forma que un do-

cumento puede estructurarse por medio de referencias a él mismo o a otros documentos.

El mecanismo de *best match* y *relevance ranking* no requiere de combinaciones lógicas de términos de búsqueda. Se basan en la analogía o en el proceso de *best-match*. El usuario sólo tiene que introducir un número de términos que sea relevante para la materia en cuestión, y pueden ser tantos como considere oportuno. De esta forma, el sistema busca aquellos registros que se ajustan de alguna manera a la lista de términos. Así, al usuario se le muestran, primero, los registros que más se ajustan a la lista *(best match)*.

En los sistemas más avanzados se les asigna un peso según su importancia en la base de datos (relevancia). El número de ocurrencias de un término dado en un registro desempeña un papel importante en el algoritmo de ranking. De esta forma, la relevancia es inversamente proporcional a la frecuencia de aparición del término en la base de datos. Este sistema de búsqueda puede reemplazar completamente el uso de combinaciones booleanas. De hecho, la llamada «superación de Boole» consiste en sustituir el álgebra de conjuntos por sistemas de ponderación, relevancia y comparación.

Para finalizar, cabe decir que los motores de búsqueda de los sistemas de información documental están específicamente diseñados para almacenar y recuperar texto, o, en otras palabras, información documental. Por consiguiente, la información se puede almacenar en registros estructurados en campos de longitud variable que admiten la repetición de valores. Además, son capaces de manejar conjuntos de registros y redefinir las estrategias a partir de los resultados obtenidos. Asimismo, cada una de las palabras de los campos puede hacerse recuperable a través de los índices inversos.

En líneas generales, los motores de búsqueda de los sistemas de información documental presentan, como características más comunes, las siguientes:

— Flexibilidad de las estructuras de registros.

— Campos de longitud variable.

— Gestión de valores repetibles.

— Estructura de ficheros inversos: búsqueda por índices y no secuencial.

— Acceso al contenido de la base de datos a través de operadores booleanos y de proximidad o adyacencia.

— Técnicas de truncamiento y limitación por campos.

— Operadores de comparación.

— Generación de conjuntos.

— Posibilidad de importar y exportar registros.

— Flexibilidad en la salida de datos.