Sistemas de Gestión de Bases de Datos Documentales

Características Principales y Metodología de diseño

Por **Lluís Codina**

Julio 2015

Departamento de Comunicación
Universitat Pompeu Fabra
Grupo de Investigación en
Documentación Digital y Comunicación Interactiva (DigiDoc)
Máster Universitario en Comunicación Social

Sistemas de Gestión de Bases de Datos Documentales:
Características Principales
y Metodología de diseño

Por **Lluís Codina**
Universitat Pompeu Fabra
lluis.codina@upf.edu
Julio 2015

Resumen

En esta publicación se describe la naturaleza, funciones y características principales de las bases de datos documentales y de los sistemas de gestión de bases de datos en general. Se presentan los componentes de un base de datos, las diferencias entre sistemas relacionales y documentales y se introducen las herramientas y fases principales para el diseño de una base de datos con cualquier propósito, pero en especial para situaciones donde es necesario describir y representar objetos o conceptos multidimensionales, y por tanto con diversas características o atributos, como es típico en algunas de investigaciones académicas. Incluye sugerencia de actividades para quienes deseen profundizar en los temas expuestos.

1. Introducción

Las bases de datos son una de las tecnologías para la organización de la información más eficientes y poderosas de que disponemos en la actualidad. De hecho, las bases de datos están en el núcleo de los sistemas y de los servicios de información que poseen mayor significación tanto desde un punto de vista económico como social.

Por ejemplo, la mayor parte de los sistemas de información empresarial, desde el control de existencias de almacén hasta el sistema de ventas, tiene en su núcleo una base de datos. En el seno de muchas actividades sociales vinculadas con la gestión del conocimiento sucede otro tanto: desde la gestión de archivos y museos hasta la automatización de centros de documentación, pasando por los mejores servicios de información en Internet o por la gestión de investigaciones académicas, todos estos sistemas tienen una base de datos en su núcleo.

De hecho, solamente es posible gestionar volúmenes de información de importancia crítica de manera eficiente mediante el uso de bases de datos, lo que explica la gran penetración de esa tecnología.

Por esa razón, muchos profesionales de la información y la comunicación, desde documentalistas hasta comunicadores y comunicólogos, acaban involucrados de alguna manera en la utilización de bases de datos. Aquellos profesionales que, ahora o en el futuro, se vean involucrados en la creación y el mantenimiento de bases de datos podrán conocer con este documento en qué consiste un sistema de gestión de bases de datos y cómo se desarrollan bases de datos documentales.

2. Bases de datos *versus* sistemas de gestión de bases de datos

La primera distinción que corresponde establecer es la que existe entre la pareja de conceptos [base de datos], [sistema de gestión de bases de datos], para lo cual proponemos las siguientes definiciones:

Base de datos: una colección de datos relativos a un dominio del conocimiento, agrupados en unidades lógicas denominadas registros y que pueden ser manipulados por un programa informático.

Algunos ejemplos:

- ***Eric***, base de datos sobre educación y ciencias sociales http://www.eric.ed.gov
- ***Getty Images***, banco de imágenes orientado al público profesional http://www.gettyimages.com
- ***Index Translationum***, base de datos de libros traducidos en todo el mundo http://www.unesco.org/xtrans
- ***Prensa Histórica***, base de datos de prensa histórica española. http://prensahistorica.mcu.es/
- ***IMDB,*** base de datos de cinematografía mundial http://www.imdb.com/

Sugerencia de Actividad: Entre en las anteriores bases de datos. Una vez vistas las cinco, seleccione dos de ellas. Intente entender cuál es el contenido sobre el que ofrecen información. Haga algunas pruebas de búsqueda. Intente determinar si hay alguna estructura común en la forma que presentan los datos una vez hecha una búsqueda. Ejemplo: en IMDB cuando examinamos los resultados de búsqueda para un título de película, ¿vemos algunos datos comunes, como título, año de producción, etc.? Vea si en los dos casos que usted ha seleccionado es posible detectar esta clase de estructuras. Documente la práctica con capturas de pantalla y una breve explicación.

Sistema de gestión de bases de datos: un programa informático que permite la creación, mantenimiento y explotación de bases de datos.

Algunos ejemplos:

- ***Access,*** sistema de gestión de base de datos relacionales de la empresa Microsoft, parte del paquete Office.
  http://office.microsoft.com/es-es/access/
- ***Base***, sistema de gestión de bases de datos relacionales del paquete Libre Office, de dominio público.
  https://es.libreoffice.org/descubre/base/
- ***Inmagic DB/Text Works***, sistema de gestión de base de datos documentales de la empresa InMagic.
  http://www.inmagic.com/
- ***FileMaker***, sistema de gestión de bases de datos con características documentales y relacionales al mismo tiempo, de la empresa del mismo nombre (propiedad de Apple).
  http://www.filemaker.com
- ***Knosys,*** sistema de gestión de bases de datos documentales de la empresa Micronet.
  http://www.knosys.net/
- ***WinISIS***, sistema de gestión de bases de datos documentales de la Unesco y de distribución libre a instituciones sin ánimo de lucro.
  http://portal.unesco.org/ci/en/ev.php-URL_ID=2071&URL_DO=DO_TOPIC&URL_SECTION=201.html

Sugerencia de Actividad: Entre al menos en uno de los sitios web de los productos o empresas anteriores de tipo relacional y en otro de tipo documental. Haga capturas de pantalla. Añada una breve descripción del tipo de soluciones que parecen ofrecer o que destacan más en cada caso (soluciones empresariales, personales, de gestión, de tratamiento de documentos, de tratamiento de datos, etc.).

En resumen: una base de datos, en sentido estricto, es un conjunto de datos, mientras que un sistema de gestión de bases de datos es el programa informático que los manipula. En cambio, por economía de lenguaje se suele utilizar la expresión "base de datos" para referirse a cualquiera de los dos conceptos, pero como vemos, en cada caso, y según el contexto, la expresión se puede referir a cosas bien distintas.

3. Estructura de una base de datos

Hemos explicado antes que una base de datos es una colección de datos relativos a un dominio del conocimiento. La definición precedente era necesaria para explicar qué es una base de datos desde el punto de vista tecnológico.

Ahora bien, nosotros nos proponemos no solamente saber *qué es* una base de datos, sino también saber *crear* bases de datos. Para ello, debemos familiarizarnos también con las definiciones (basadas en teoría de sistemas) de los siguientes cuadros. Este conjunto de definiciones proporciona un modo de ver las bases de datos que, como podremos comprobar más adelante, facilita el diseño de sistemas de información basados en bases de datos. Son las siguientes definiciones:

Cuadro n.1: Definición de Base de datos

Una base de datos es un sistema de información que mantiene registros sobre las características o sobre las actividades de alguna parte del mundo real.

Cuadro n. 2: Definición de entidad

Las cosas del mundo real sobre las cuales mantiene registros una base de datos se denominan entidades. Las entidades pueden consistir en objetos físicos, como libros, en seres animados, como personas, o en conceptos abstractos, como ideas y teorías científicas.

Cuadro n. 3: Relación entre bases de dato y entidades

De acuerdo con las definiciones establecidas en los cuadros 1 y 2, ahora podemos decir que una base de datos es un conjunto de registros que describen entidades del mundo real.

Cuadro n. 4: Entidades y atributos

Por la definición del cuadro n. 3, podemos decir que, si las entidades pueden ser descritas o representadas, ello implica que las entidades deben poseer algunas propiedades reconocibles, que denominamos atributos.

Cuadro n. 5: Atributos y campos

Para representar entidades del mundo real en registros es necesario representar sus atributos. Los atributos de una entidad en un registro se denominan campos.

Volvamos ahora, por un momento, a adoptar una perspectiva tecnológica para ver que un registro es un grupo de campos, que a su vez se componen de bytes, y que cada byte sirve para

representar un número o de una letra. En resumen:

Cuadro n. 6: Base de Datos

Una base de datos = Un conjunto de registros Un registro = Un conjunto de campos Un campo = Un conjunto de bytes

Cuadro n. 7: Relación BDD / Mundo real

Cuando en la base de datos hablamos de:	En el mundo real hablamos de:
Registros o Tablas	Entidades
Campos o Columnas de una tabla	Atributos

3.1. Registros y campos

Ahora vamos a volver sobre las dos parejas formadas por:

- Entidades y Atributos
- Registros y Campos

Entidad

Hemos dicho que las bases de datos contienen información sobre cosas del mundo real, es decir, tanto del mundo material como del mundo conceptual. A esas cosas del mundo real sobre las que una base de datos almacena información se las denomina entidades y pueden ser cosas materiales (libros, personas, etc.) o cosas intangibles (ideas, conceptos, etc.). Para representar tipos de entidades se utilizan registros o tablas como veremos más adelante, de modo que cada clase en entidad es un modelo de registro o una tabla.

Atributo

Los parámetros o rasgos que caracterizan a una entidad, como por ejemplo los atributos de un libro pueden ser: el nombre del autor, el título del libro, la fecha de publicación, etc. Los conceptos Entidad y Atributo, que son los que nos conviene utilizar cuando pensamos en cosas del mundo real, pasan a ser Registros y Campos cuando pensamos en términos de bases de datos:

Registros

Los registros son representaciones de entidades. Al mismo tiempo, son la unidad principal de información que se utiliza en las bases de datos. Cada registro se refiere a una entidad en una relación 1:1, es decir, una entidad, un registro. Por ejemplo, un libro (entidad), un registro. El registro se corresponde con el concepto de ficha que se utiliza en los ficheros manuales con los

que todos estamos familiarizados.
En las bases de datos relacionales, los registros son las filas de una tabla. Por tanto, un modelo de registro es también una tabla.

Campos

Los campos son las partes en que se articula un registro. Cada campo corresponde a un atributo de la entidad. Los campos, por tanto, son zonas de información que ayudan a estructurar los datos relativos a la entidad. En una base de datos bibliográfica, por ejemplo, los registros se estructuran, típicamente, en campos como: título, autor, fuente, etc. En una tabla, los campos son las columnas de la misma. Las siguientes figuras resumen de nuevo las definiciones y las relaciones precedentes:

Figura 1a: **Entidad libro** (datos tomados de la cubierta)

Figura 1b: Libro: (datos de la página de créditos)

Y. Bar-Hillel, M. Bunge,
A. Mostowski, J. Piaget,
A. Salam, L. Tondl, S. Wanatabe

El pensamiento científico
CONCEPTOS, AVANCES, METODOS

Publicado conjuntamente por:
Editorial Tecnos, Madrid y UNESCO, País, Francia.

8 UNESCO
ISBN: 84-309-10220-0
Depósito legal: M. 40.335-1983
tecnos / unesco

Figura 1c: Un Registro en forma de ficha (representación del libro anterior)

Título	El pensamiento científico: conceptos, avances y métodos
Autor	BAR-HILLEL, et al.
Fuente	Madrid, Paris: Tecnos; Unesco, 1983, 265 pp. (ISBN: 84-309-10220-0)
Descriptores	Filosofía de la ciencia, Conceptos científicos, Teoría de conjuntos, Teoría de sistemas, Lógica, Psicología, Física, Semiótica

Figura 1d: Un Registro en forma de tabla

Título	***Autor***	***Fuente***	***Descriptores***
El pensamiento científico: conceptos, avances y métodos	BAR-HILLEL, et al.	Madrid, Paris: Tecnos; Unesco, 1983, 265 pp. (ISBN: 84-309-10220-0)	Filosofía de la ciencia, Conceptos científicos, Teoría de conjuntos, Teoría de sistemas, Lógica, Psicología, Física, Semiótica

Aunque los registros y los campos son la parte más evidente de una base de datos, la estructura de esta no se completa sin un tercer elemento: los índices. En el siguiente punto nos ocupamos de los índices, un elemento menos visible que los anteriores, pero sin el cual las bases de datos no podrían proporcionar sus más características prestaciones, a saber: la eficiencia en la recuperación de información.

Toda búsqueda de información, ya sea en un libro, un catálogo de fichas, o una base de datos, comporta siempre una o más operaciones de comparación, operaciones que no se detienen hasta que se encuentra el elemento buscado: una página de libro, una ficha de un catálogo o un registro de una base de datos, respectivamente.

Cuanto más grande es el espacio de búsqueda, más operaciones de comparación debemos llevar a cabo, en promedio, para tener éxito. En este sentido, una de las prestaciones más apreciadas de una base de datos es su capacidad para proporcionar respuestas a una velocidad independiente del volumen de información que contiene la base de datos, y esa prestación la proporciona el índice de la base de datos.

Sugerencia de Actividad: Entre en las dos bases de datos usadas para la Actividad 1. Intente determinar para cada una de las bases de datos elegidas, algunas de las Entidades representadas (p.e. en IMDB una de las entidades son films, ¿hay alguna más?). Una vez identificadas una o más entidades para cada una de las bases de datos seleccionadas, escoja una de las Entidades (p.e. films en IMDB) y haga una lista de al menos 4 atributos o campos que ha podido identificar de esa Entidad (p.e. en films, tendremos el campo título y el campo año de producción o exhibición, ¿qué otros campos identifica?).

4. Modelo textual *versus* modelo relacional

Vamos a centrarnos ahora de los sistemas de gestión de bases de datos, es decir, de los programas que sirven para crear y explotar bases de datos.

Como ya hemos avanzado en la sección anterior, podemos encontrar en el mercado dos grandes clases de sistemas de gestión de bases de datos:

- Sistemas de **propósito general**, basados en el modelo relacional. Suelen denominarse sistemas de gestión de bases de datos, SGBD, en siglas
- Sistemas de **gestión documental**, basados en el modelo textual. Suelen denominarse sistemas de gestión documental, SGD, en siglas

En teoría, los SGBD pueden estar basados en diferentes modelos, como son el modelo relacional, el modelo eXtensible Markup Language (XML), el modelo JavaScript Object Notation (JSON)… Es verdad que para la representación e intercambio de datos estructurados en la web, estas tecnologías no relacionales están ganando mucho terreno, pero también es verdad que una gran parte de SGBD continúan basados en tecnología relacional. En todo caso, en el marco de este artículo, estableceremos la siguiente equivalencia:

sistemas de propósito general = sistemas relacionales

Por su parte, las principales prestaciones de los SGD están basadas en su capacidad para procesar información textual y para incorporar controles terminológicos. Esa capacidad está fundada en un modelo que toma al texto como base de representación y de recuperación de la información. De aquí la equivalencia que expresamos como:

sistemas documentales = sistemas textuales

4.1. Principales diferencias entre los SGBD y los SGD

Los SGBD (Sistemas de gestión de bases de datos), es decir, los sistemas basados en el modelo relacional, están orientados hacia la gestión de datos comerciales, administrativos, contables y, en general, de cualquier tipo, pero siempre muy estructurados. Sus características principales, por tanto, son:

- Están bien preparados para datos numéricos o para cadenas de caracteres cortas. En general, están bien preparados para informaciones que se gestionan bien en una tabla con valores siempre muy compactos (nombre, cifras, fechas, etc.).
- Presentan y manipulan la información en forma de tablas homogéneas con filas siempre formadas por el mismo número de columnas.

Por su parte, las principales características de los sistemas de gestión documental (SGD) son las siguientes:

- Están preparadas para utilizar y para explotar la información textual de tipo discursivo, como la que puede encontrarse en artículos de revista, libros, informes, tesis doctorales, ensayos, etc.
- Los campos son de extensión variable. No suelen requerir definición de extensión de los registros y ficheros.
- Disponen de módulos especiales de clasificación y listas de descriptores.
- Pueden utilizar controles terminológicos y lenguajes documentales como listas de términos autorizados o tesauros.

Los SGD constituyen, además, el núcleo de otros sistemas de información especializados, desde sistemas de gestión de referencias bibliográficas, a sistemas de gestión de unidades de información (bibliotecas, archivos…). Un ejemplo de este segundo tipo de aplicaciones es el propio DB/TextWorks, a partir del cual se ha diseñado DB/Text Library Suite http://www.inmagic.com/products/dbtext-library-suite.

Los SGBD están diseñados para procesar datos administrativos, y en su terreno son

insuperables. Por su parte, los SGD están diseñados para gestionar información textual o, más exactamente, conocimiento, y en ese aspecto son también insuperables. El alumno puede preguntarse si existen sistemas híbridos, es decir, si existen programas de gestión de bases de datos capaces de tratar al mismo tiempo con datos administrativos y con información textual o conocimiento puro. La respuesta es que sí, pero siempre bajo la forma de sistemas que integran, con mayor o menor fortuna, a dos subsistemas distintos: un subsistema relacional y un subsistema textual o documental. Lo que no existe es un único modelo ambivalente.

5. El diseño

Diseñar una base de datos con un sistema de gestión de bases de datos, ya sea relacional, documental o mixto, requiere básicamente lo siguiente:

1. Determinar las entidades que formarán parte de la base de datos
2. Determinar los atributos de las entidades
3. Determinar las relaciones, si las hay, entre las entidades
4. Desarrollar un documento, llamado, diccionario de datos donde se fijan por escrito los aspectos anteriores una vez se han determinado con seguridad (a veces harán falta varios ensayos y pruebas hasta llegar al documento final).

5.1. Modelo Entidad-Relación

El modelo entidad-relación (o modelo E-R) ayuda a detectar sin ambigüedad las entidades que formarán parte de la base de datos, es decir, los objetos que forman parte del sistema de conocimiento. Para ello, recuperaremos conceptos ya conocidos como los de entidad y atributo. Estas entidades son las que habrán de ser descritas en la base de datos e importa, por tanto, identificarlas con la mayor precisión posible. El modelo E-R añade un tercer concepto a los ya conocidos, con lo cual tenemos tres ítems:

- Entidad
- Atributo
- Relación

Como ya sabemos, si las bases de datos representan a cosas u objetos del mundo real, tales cosas deben ser identificables y deben tener algunas propiedades. La única restricción aplicable 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.

Los **atributos**, por su parte, sabemos también que 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. 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.

De este modo, si un tipo de entidad posee los atributos A, B, C, el modelo de registro debe poseer los campos A, B, C. En este punto, para cada atributo convertido en un campo necesitamos diferenciar entre los siguientes conceptos:

- **Etiqueta** del campo
- **Valor** del campo
- **Dominio** del campo

La **etiqueta** es el nombre del campo, es decir, una cadena de caracteres que identifica una zona del registro. El **valor** 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 el conjunto del cual puede tomar sus valores un campo. Por ejemplo, el dominio del campo Año de publicación, es el conjunto formado por los años de publicación de documentos. Un equivalente al término **Dominio** de un campo sería **Definición** o **Descripción** de un campo.

Figura 2: Un registro en representación de un libro

Título	Multimedia and hypertext: the Internet and beyond
Autor	Jakob Nielsen
Fuente	Boston: Academic Press, 1995
Año	1995
Páginas	480
ISBN	0-12-518408-5

Descriptores	Hipertextos, Multimedia, Sistemas de información, Publicaciones digitales, Documentación, Bases de datos, Internet, World Wide Web

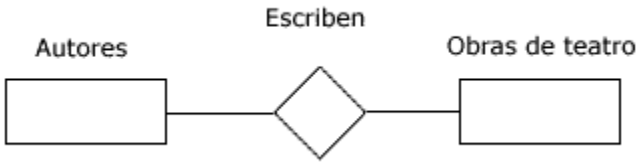
Veámoslo con el ejemplo anterior. De acuerdo con el registro de la figura 2, el segundo campo o zona de información se puede analizar o descomponer así:

- **Etiqueta** del campo: Autor
- **Valor** del campo: Jakob Nielsen
- **Dominio del campo**: Nombre y apellido de los autores o responsables intelectuales de los documentos

5.2. Generalizaciones y abstracciones

Al igual que distinguimos entre 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.

Figura 2: Diagrama de una relación entre dos entidades. 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 catedráticos de universidad y universidades es de **1:N**, (de uno a muchos) porque cada catedrático pertenece a una sola universidad, y una universidad tiene diversos catedráticos.

Finalmente, una relación de **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 que hemos puesto en el diagrama anterior.

5.2.1. Toma de decisiones

En conclusión, 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 también para determinar otros elementos del diseño.

Por ejemplo, y volviendo al caso anterior, donde se nos pide diseñar una base de datos sobre teatro. 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 dado por supuesto en el diagrama.

Si adoptamos 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:

1. En caso de duda, tratar las cosas como entidades distintas
2. Determinar la relación entre entidades
3. Determinar su grado
4. Si la relación es de grado 1:1, entonces se trata de una sola entidad, y un solo modelo de registro (o una sola tabla) 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
5. Si la relación es de grado N:1 se trata de dos entidades y, por lo tanto, necesitamos dos modelos de registro (o dos tablas), uno para cada entidad. Los dos modelos de registro deben contar con un campo compartido, lo cual proporciona dos campos con un dominio común. Esto último permitirá el cruce de datos.

6. Si la relación es del grado N:M necesitamos tres modelos de registro (o tres tablas): uno para cada entidad y otro para representar la relación. El modelo (o tabla) de la relación contendrá al menos un campo de identificación de cada una de las entidades. Por ejemplo, el modelo de registro o tabla para representar la relación en nuestro caso hipotético tendrá al menos dos campos: uno con el identificador (nombre y apellido o clave de identificación) de los autores de obras de teatro y otro con el identificador (p.e. un número único asignado al entrar los datos de la obra) de cada obra de teatro.

¿Qué sucedería si no procedemos como indica este modelo? 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 escrito 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 que si necesitamos llegar a la implantación para detectar los errores, tal vez después de meses de trabajo que, de golpe, se revelaron inútiles.

Sugerencia de Actividad: Imagine que tiene que diseñar una base de datos de cinematografía (parecida a la de IMDB que ya hemos tratado antes). Imagine que identifica dos Entidades que necesitará tratar: la entidad **Films** (o películas) y la entidad **Directores** (directores de las películas). Aplique el modelo Entidad-Relación y diga si se trata de una relación 1:1, N:M o N:M y, en consecuencia, cuántos modelos de registro o tablas necesitaría para su base de datos de cinematografía.

5.3. El diccionario de datos

El diccionario de datos 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, de tal manera que el diccionario de datos marcará decisivamente el rendimiento y la calidad global del sistema de información.

Consiste en la lista detallada de cada uno de los campos que forman los distintos modelos de registro de la base de datos. A cada campo de cada modelo de registro se le aplica una parrilla de análisis que contempla, como mínimo, los siguientes aspectos:

1. Etiqueta
2. Dominio
3. Tipo de datos
4. Indexación
5. Tratamiento documental
6. Lengua
7. Otros controles de validación u observaciones

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 podría tener entonces esta forma:

Campo Título Etiqueta: Título
Dominio: Título del documento.
Tipo: Alfanumérico.
Indexación: Sí
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.
Observaciones: El título se transcribe de la siguiente forma Título: antetítulo: subtítulo.

Campo Descriptores Etiqueta: Descriptores
Dominio: Palabras clave normalizadas que expresan los conceptos principales contenidos en el documento, según el siguiente principio general: si el artículo contiene n conceptos relevantes se asignan n descriptores.
Tipo: Alfanumérico
Indexación: Sí
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.

Campo Fecha de Publicación Etiqueta: FechaPub
Dominio: La fecha de publicación de la noticia, indicada con el siguiente formato, DD/MM/AAAA.
Tipo: Fecha
Indexación: Sí
Tratamiento documental: No procede
Lengua: No procede
Controles de validación: No admite valores fuera de rango.

Estudiando el ejemplo de diccionario de datos anterior, formado únicamente por tres campos, podemos observar, por vía de ejemplos, algunos aspectos importantes para el diseño de bases de datos que retomamos ahora:

(1) ***Dominio***. En el contexto del diccionario de datos, el dominio se refiere al conjunto del que un campo puede obtener sus valores. Dicho de otra forma, describir el dominio de un campo consiste en describir el contenido teórico de ese campo, es decir, la clase de contenidos que puede admitir el campo.

(2) ***Tipo (o data type)***. Es el tipo de dato que admite el campo. Los tipos de datos suelen ser: numérico, alfanumérico, fechas y lógico. El tipo dato (data type) establece cuáles son las operaciones válidas que puede hacerse y el rango de valores aceptable. Por ejemplo, el tipo de datos alfanumérico permite realizar operaciones de ordenación, en cambio, no admite operaciones aritméticas. Por el contrario, un tipo de dato numérico solamente admite números y operaciones aritméticas. Por su parte, un campo de fecha 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, etc.

(3) ***Etiquetas***. En los tres ejemplos anteriores hemos seleccionado etiquetas sujetas a algunas restricciones habituales: no más de ocho caracteres, sin espacios y sin acentos. Lo hemos hecho así para ilustrar la diferencia entre el nombre del campo en lenguaje natural (sin limitaciones especiales) y su etiqueta, que suele estar sometida a restricciones como las señaladas.

(4) ***Tratamiento documental***. Este parámetro 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 thesaurus o de una lista de autoridades.

(5) ***Lengua***. La lengua o idioma de un campo puede ser, o bien la lengua del documento, o bien

la del centro de documentación. Si observamos los ejemplos anteriores esto significa que, en el caso de un documento escrito en inglés, el título estaría en inglés, pero los descriptores en castellano, siempre de acuerdo con el diccionario de datos precedente.

La descripción funcional, por su parte, debe incluir los siguientes elementos:

- Qué clase de información se tratará y cómo entrará la información en el sistema
- Qué procesos documentales se llevarán a cabo
- Qué servicios y productos generará el sistema, y/o a qué aplicaciones podría dar soporte

El primer punto debe describir en qué consisten las entradas del sistema. El punto dos debe proporcionar una idea sobre qué procesos de tratamiento documental automatiza la base de datos, y el punto siguiente debe explicar en qué consisten las salidas del sistema.

Sugerencia de Actividad: Imagine que le piden el diccionario de datos de una base de datos. Tomemos el ejemplo de nuevo de una base de datos de cinematografía. Indique, para el campo Título del Film, de esta base de datos, qué indicaciones daría al diccionario de datos de ese campo. A continuación, presentamos la posible ficha para ese campo y dejamos algunas descripciones sin completar para que usted las complete:

1	Etiqueta	Título
2	Dominio	Título del film en la versión estrenada en España
3	Tipo de dato	
4	Indexación	
5	Tratamiento documental	No
6	Lengua	La lengua del film
7	Otros controles	

6. Conclusiones

Hemos visto una de las tecnologías más presentes en el mundo de la información actual (tan presentes que nos pasa desapercibida), como son las bases de datos, la clase de software que sirve para gestionarlas y, finalmente, hemos visto los componentes fundamentales de una metodología que permite encarar el diseño de una nueva base de datos.

En algunas investigaciones académicas (y de cualquier clase en general) necesitamos controlar

y describir un número de entidades con diversas propiedades y facetas. Un caso típico, en la investigación en Comunicación, por ejemplo, serían los análisis de contenidos donde podemos necesitar analizar y describir centenares o miles de piezas. En todas estas situaciones realizar un buen diseño es crítico para después poder explotar la información de forma adecuada, incluso de formas no inicialmente previstas. Lo contrario, proceder a una entrada de datos masiva sin las precauciones de análisis necesarias conducirá inevitablemente, o bien a perder posibilidades de explotación o a tener que repetir todo el trabajo.

7. Referencias

- ABADAL, E.; CODINA, L. (2005). *Bases de datos documentales: características, funciones y método*. Madrid : Síntesis, 2005, 220 p.

- AENOR. (1990). Norma UNE 50-106-90. *Documentación. Directrices para el establecimiento y desarrollo de tesauros monolingües*. Madrid: AENOR, 1990, 47 p.

- ANDERSON, James D.; PÉREZ-CARBALLO, José (2005). *Information retrieval design: principles and options for information description, organization, display, and access in information retrieval databases, digital libraries, catalogs, and indexes*. St. Petersburg, Fla.: Ometeca Institute, 2005. 617 p.

- AVISON, D. E.; FITZGERALD, Guy (2003). *Information systems development: methodologies, techniques and tools*. 3rd ed. London [etc.] : McGraw-Hill, cop. 2003.

- BAEZA-YATES, Ricardo; RIBEIRO, Berthier de Araújo Neto (2011). *Modern information retrieval: the concepts and technology behind search*. 2nd ed. Harlow [etc.]: Addison-Wesley / Pearson, 2011.

- *Bases de datos* (2010). Barcelona : UOC, 2010.

- CACHEDA SEIJO, Fidel; FERNÁNDEZ LUNA, Juan Manuel; HUETE GUADIX, Juan Francisco (2011). *Recuperación de la información : un enfoque práctico y multidisciplinar*. Madrid : Ra-Ma, cop. 2011. XVIII, 232 p.

- CHECKLAND, P. B. (1999). *Soft systems methodology: a 30-year retrospective.* Chichester [etc.]: John Wiley, cop. 1999. 330 p.

- CHECKLAND, P. B.; HOLWELL, Sue (1998). I*nformation, systems and information systems: making sense of the field*. Chichester [etc.]: John Wiley & Sons, cop. 1998. 262 p.

- CHECKLAND, P. B.; POULTER, J. (2006*). Learning for action: a short definitive account of soft systems methodology and its use for practitioner, teachers, and students*. Hoboken, NJ: Wiley, cop. 2006.

- CHEN, P.P-S. (2002). "Entity-relationship modeling: historical events, future trends, and lessons learned". BROY, M.; DENERT, E. (eds*.). Software pioneers: contributions to software engineering.* New York: Springer-Verlag, 2002, p. 296-310.

- CODINA, Lluís (1998). "Metodología de análisis de sistemas de información y diseño de bases de datos documentales: aspectos lógicos y funcionales". BARÓ, J.; CID, P. (eds.). *Anuario SOCADI de Documentación e Información 1998*. Barcelona: SOCADI, 1998, p. 195-210.

- GIL LEIVA, Isidoro (2008). *Manual de indización : teoría y práctica*. Gijón : Trea, 2008. 429 p.

- LAUDON, Kenneth C.; LAUDON, Jane P. *Management information systems: managing the digital firm.* 12th ed. Global ed. Harlow, Essex; Boston : Pearson Education, 2011. 630 p.

- TEOREY, Toby J. (2011). *Database modeling and design: logical design* [recurso electrónico]. 5th ed. Burlington, MA : Morgan Kaufmann Publishers/Elsevier, cop. 2011.

Forma recomendada para citar esta publicación:
Lluís Codina. *Sistemas de gestión de bases de datos documentales: características principales y metodología de diseño.* Barcelona: UPF, 2015
Colección ***Materiales***

- *Máster Universitario Online en Documentación Digital* http://documentaciondigital.org/
- *Máster Universitario en Comunicación Social* http://www.upf.edu/mastercomunicacio/es/

Otros materiales y recursos en el sitio web del autor: www.lluiscodina.com

---