

Análisis de sistemas y metodología de diseño de bases de datos documentales

Lluís Codina

Universidad Pompeu Fabra de Barcelona
Departamento de Ciencias Políticas y Sociales

0.1. Resumen

Se propone una metodología general para el diseño de bases de datos documentales, basada en el análisis de sistemas. El modelo aborda el problema desde tres puntos de vista complementarios: el aparato conceptual, el aparato instrumental y el aparato procedimental. Conforme al primero, los sistemas de información pueden caracterizarse en dos niveles: como un sistema de actividades humanas (SAH) y como un sistema de conocimiento (SCO). El segundo se apoya en tres procedimientos de representación: el modelo entidad-relación, el diccionario de datos y la norma ISBD. El tercero propone una secuencia de desarrollo de los proyectos en tres fases sucesivas y comunes a todo sistema de información: análisis, diseño e implantación.

Palabras clave: Análisis y diseño de sistemas de información. Bases de datos documentales. Metodologías.

0.2. Abstract

A general methodology for documentary database design is proposed. The methodology is based on systems analysis. The model considers the problem from three complementary points of view: the conceptual frame, the representational models and the methodological procedures. The first one consists of two levels: the human activities system and the knowledge system. The second one relies on three representational procedures: the entity-relation model, the data dictionary and the ISBD standard. The third one proposes a standard development cycle in three successive stages: analysis, design and implementation.

Keywords: Information systems analysis and design. Information retrieval systems. Metodologies.

1. El lugar de las metodologías en los sistemas de información documentales

El uso del término *metodologías*, en el contexto de los sistemas de información, produce equívocos a menudo, particularmente porque se espera de ellas cosas que no pueden dar, y otras veces porque se atribuye, un poco abusivamente, el carácter de metodología a meras recomendaciones o aún a reflexiones sobre el tema que, sin duda, pueden ser muy valiosas, pero que no constituyen en sí mismas una metodología.

Empezaremos por indicar lo que, a nuestro juicio, no debe esperarse de una metodología. Entendemos que no es un algoritmo. No es una lista finita de operaciones, expresadas sin ambigüedad, que conducen a la solución de un problema. Si así fuera, no se necesitarían seres humanos para diseñar y para crear sistemas de información. Podría codificarse la metodología y encargarse su ejecución a un robot.

Una metodología es una guía para el desarrollo (de sistemas de información), que sirve para que el resultado final se deba, en lo más posible, a la planificación consciente y, en lo menos posible, al azar y/o al método de ensayo y error.

Es evidente que, mediante la planificación consciente, cualquier profesional tiene derecho a esperar un mayor éxito que si se toman decisiones al azar y/o por el método del ensayo y error. Sin embargo, por muy correcta que sea una metodología, ningún lego tiene la más mínima posibilidad de sacar adelante un proyecto. Una metodología solo puede aprovecharla un profesional o una persona con formación en la materia. Sería ridículo que este autor, por ejemplo, en cuya formación no aparecen para nada las ciencias médicas, creyera que la simple lectura de una metodología le proporciona la competencia para realizar operaciones de cirugía, por ejemplo.

Por tanto, la diferencia entre utilizar una metodología o no, está, nada más (y nada menos), que en la parte final del producto, que debe atribuirse:

- al azar
- al ensayo y error
- a la planificación consciente.

De ello se desprende que siempre se desliza algo de azar en el diseño de sistemas de información, así como siempre hay la necesidad de recurrir al ensayo y error para refinar el resultado. La cuestión clave radica en que la parte de planificación consciente debe ser la mayor posible en el resultado final, tanto por razones de eficiencia como por razones de economía. Lo contrario, que el azar y el ensayo y error, tengan un gran peso, no sólo produce sistemas peores y a mayor

coste, sino que puede llegar a colapsar el propio proceso de diseño o incluso puede provocar la aniquilación del proyecto.

Hemos dicho que unas meras reflexiones, o unos consejos, no son una 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 primero, el *aparato conceptual* o grupo de elementos conceptuales, tiene la misión de proporcionar a los responsables del desarrollo de sistemas de información unas bases conceptuales mínimas que faciliten su entendimiento de todo el proyecto y que faciliten, así mismo, la comunicación entre los diferentes sectores 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 ayuda a reducir al mínimo posible los riesgos debidos a la improvisación.

Por otro lado, importa señalar que la metodología que se expone aquí se ha obtenido, básicamente, por la utilización de tres tradiciones científicas y académicas distintas, que este autor ha intentado fusionar en una metodología unificada y, hasta cierto punto, consistente. Se trata de las siguientes tradiciones académicas y/o tecnológicas:

- La tradición del *análisis de sistemas*, proveniente de las ciencias informáticas. Unos de los autores más representativos y cualificados sería Yourdon (1993).
- La tradición de la *metodología general de resolución de problemas*. Concretamente, se han utilizado las aportaciones de la SSM (*Soft System*

Methodology), una metodología elaborada principalmente (pero no únicamente) por Checkland (Checkland, 1981; Checkland y Scholes, 1990; Lewis, 1994). Un aspecto importantes a considerar es que, tanto esta tradición como la anterior, se basan formal y expresamente en la teoría general de sistemas. La SSM es particularmente útil al análisis y desarrollo de sistemas de información documentales ya que, por un lado, se inspira fundamentalmente en la teoría general de sistemas y, por otro, proporciona métodos que se pueden aplicar en las fases *anteriores* a la intervención de tecnologías concretas, es decir, que proporciona métodos *independientes* de las tecnologías concretas, de aquí el término *soft* (blando) de su nombre. Puede servir, por tanto, para el desarrollo de los aspectos *conceptuales* de los sistemas de información documentales y que caen, por naturaleza propia en el dominio de la Documentación como disciplina académica y práctica profesional.

- La tradición, naturalmente, de los *métodos y procedimientos de trabajo de las ciencias de la documentación*, particularmente en lo que se refiere al aparato conceptual y a la norma ISBD, concebida como una abstracción para generar descripciones de clases de documentos en términos de atributos o propiedades.

Una vez expuestas estas consideraciones de tipo meta-metodológicas, se exponen en las secciones siguientes los elementos de una metodología que, a su vez, tiene sus fundamentos teóricos en un modelo conceptual sobre sistemas de información documental expuesto con más detalle en otro lugar (Codina, 1994a y Codina 1994b).

2. Aparato conceptual

Un primer punto de partida muy útil en el diseño de todo sistema de información y, por tanto, también en el diseño de una base de datos documental, consiste en definir un sistema de información como un sistema, *SI*, denominado sistema de información, que mantiene registros sobre otro sistema del mundo real, *S2*, denominado sistema objeto.

De este modo, el proceso de análisis y diseño puede concebirse como el intento de obtener un modelo de aquella parte de la realidad, o sistema objeto (*S2*) que resulta de interés para el sistema de información (*SI*). Tenemos entonces el par conceptual <sistema de información, sistema objeto>, o <*SI*, *S2*>, y la relación que les une es que el primero (*SI*) es un modelo del segundo (*S2*), exactamente en el mismo sentido en que un mapa será un buen sistema de información justo en la medida en que sea un buen modelo del territorio sobre el que informa.

El segundo punto de partida consiste en considerar que, desde el punto de vista de los intereses de la Documentación, todo sistema objeto (*S2*) se compone de dos subsistemas, que denominamos:

- a) Sistema de actividades humanas (SAH) (1)
- b) Sistema de conocimiento (SCO).

El SAH es el sistema social —es decir un sistema formado por personas y cosas— que justifica la existencia del sistema de información, porque en él desarrollan sus actividades los futuros usuarios que necesitarán que exista un sistema de información.

Por ejemplo, si pensamos en una biblioteca universitaria como en un sistema de información, entonces el sistema objeto que modela (y por tanto, el SAH) es la universidad, la cual necesita la biblioteca (así como otros recursos documentales) para sus actividades de creación y difusión del conocimiento. ¿En qué sentido la biblioteca modela a la universidad? En el sentido en que los temas y disciplinas científicas que cubre la biblioteca, la clase de documentos que adquiere, los procedimientos de trabajo, los servicios que presta, etc., son un reflejo de las características de la universidad.

Si consideramos la base de datos que automatiza el centro de documentación de una empresa determinada, este centro de documentación es el SAH respecto a la base de datos, y la propia empresa es otro SAH que, en este caso, actúa de entorno del centro de documentación. Como el entorno de un sistema siempre influye en él de alguna forma, los diseñadores de la base de datos, aunque deberán concentrarse en las características del centro de documentación, también deberán conocer las características de su entorno, esto es, de la empresa. Los ejemplos podrían multiplicarse fácilmente. Por ejemplo, si se trata de diseñar la base de datos de un museo, el SAH será el museo en cuestión, etc.

Por su parte, el sistema de conocimiento (SCO) está formado por los documentos o las entidades sobre los cuales el sistema de información debe mantener algún tipo de registros.

Por ejemplo, muchos profesionales de distintas ramas de la actividad económica y social necesitan explotar el conocimiento contenido en los documentos científicos y técnicos que se publican a diario en todo el mundo para desempeñar su trabajo. Por eso, los centros de documentación realizan, entre otras labores, una descripción de sus fondos documentales y una labor de representación del conocimiento que contienen esos fondos que, en cada caso serán distintos.

En el caso de la base de datos de un museo, por seguir con otro de los ejemplos mencionados, el SCO consistirá, según decisión de los analistas, o bien en los objetos expuestos en el museo, o bien en los documentos de su centro de

documentación o en ambos. En todo caso, este ejemplo nos demuestra que, a veces determinar qué forma parte del SCO puede ser una decisión técnica e incluso intuitiva, fruto de un análisis superficial, pero otras veces será el resultado de una decisión política más o menos elaborada.

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 procedimental o al instrumental. Aquí se he hecho una elección concreta, pero probablemente son posibles otras interpretaciones.

3. 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, el diccionario de datos y la norma ISBD.

3.1. El modelo entidad-relación

El modelo entidad-relación (o modelo E-R) se utiliza para detectar sin ambigüedades las entidades que formarán parte de la base de datos, es decir, los objetos que forman parte del sistema de conocimiento. 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 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 las cosas sobre las cuales almacena información una base de datos se las denomina entidades, y pueden ser cosas materiales (libros, personas, etc.) o conceptuales (ideas, teorías científicas, etc.).

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, 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

ta que, en principio, los atributos de una entidad son virtualmente ilimitados, será labor del documentalista seleccionar en cada caso cuáles son los que se consideren 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.

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.

Por tanto, si los registros describen entidades del mundo real, los campos corresponden a los atributos de la entidad. 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.

<i>Título</i>	Internet: manual de referencia
<i>Autor</i>	Harley Hahn; Rick Stout
<i>Fuente</i>	Madrid: Osborne Mc Graw-Hill, 1994
<i>Año</i>	1994
<i>Páginas</i>	692
<i>ISBN</i>	84-481-1882-0
<i>Descriptor</i>	Internet, Redes telemáticas, Bases de datos, Correo electrónico, Telnet, Usenet, FTP, Wais, World Wide Web

Fig. 1: Un registro representando a un libro

En este punto, necesitamos diferenciar entre los siguientes conceptos:

1. Nombre de campo
2. Valor de campo
3. Dominio de campo

El nombre del campo es una constante que identifica una zona del registro. El valor es una variable, se refiere al contenido concreto de un campo concreto y, efectivamente puede ser distinto para cada campo. 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

Veámoslo con otro ejemplo. De acuerdo con el registro de la figura 1, el segundo campo o zona de información se puede analizar así:

- Nombre del campo: Autor
- Valor del campo: Harley Hahn; Rick Stout
- Dominio del campo: El conjunto de los nombres de responsables intelectuales de los documentos.

3.1.1. 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.

3.1.2. Relaciones

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 de teatro] y [autores teatrales], y veríamos que entre ambas entidades existe la relación <escriben>, que significa más explícitamente que [autores teatrales] <escriben> [obras de teatro].

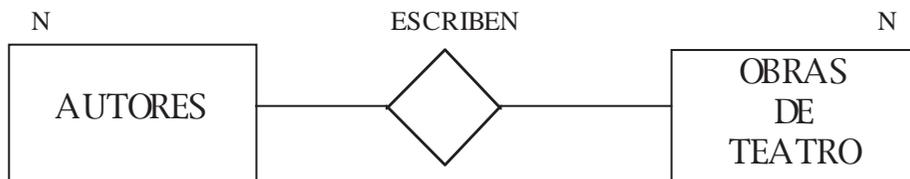


Fig.2

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 de teatro]. 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. A su vez, las entidades se identifican con sustantivos y las relaciones con verbos. En la figura 2 podemos ver un ejemplo de tales diagramas:

Así, por ejemplo, la relación que existe entre el número de ISBN y un libro es una relación de 1:1 (que se lee “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, porque cada profesor pertenece a una sola universidad, y una universidad tiene muchos profesores.

Finalmente, una relación de N:M (o de N:N) 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.

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 de gestión que suelen utilizar tecnología relacional, 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 tecnología relacional, ni necesitan modelar relaciones complejas entre entidades, como las que se dan en los sistemas de gestión administrativos.

En muchos sistemas documentales, las entidades, de hecho, no mantienen relaciones entre ellas que deban ser reflejadas en el modelo E-R. Por ejemplo, en una típica bases de datos documental sobre literatura científica y técnica 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 “sólo” aporta una cierta 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 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 dado por supuesto en el diagrama.

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 pun-

to 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 las cosas 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, o N:M, 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.

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. El campo con un dominio común podría ser el campo *Autor*, que debería figurar en ambos registros.

¿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 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 (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 una tabla más. Los interesados en esos aspectos de diseño pueden consultar Jackson (1990).

En general, la tecnología relacional debería ser necesaria cuando se trata sobre todo de modelar actividades (relaciones) y los datos relativos a cada entidad son relativamente simples o están muy estructurados. La mayoría de las actividades de gestión administrativa de una empresa son de esa clase y por eso utilizan sistemas relacionales. En cambio, deberíamos utilizar sistemas documentales en la situación simétricamente opuesta a la anterior, es decir, cuando se trata de modelar depósitos de conocimiento más que actividades, y los datos no son en realidad datos, sino información no estructurada o extremadamente compleja. La mayoría de las actividades de la Documentación responden a ese perfil y por eso utilizan sistemas documentales.

3.2. 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. Dominio
2. Tipo
3. Tratamiento indexación
4. Tratamiento documental
5. Lengua
6. Otros controles de validación

Por ejemplo, supongamos, a efectos de esta explicación, una base de datos documental sobre noticias de actualidad con sólo tres campos: <Título>, <Descriptor> y <Fecha de publicación>. El diccionario de datos tendría entonces esta forma

- 1) *Campo Título*
 - a) *Dominio: Título del documento.* El título se transcribe de la siguiente forma: *Título: antetítulo: subtítulo.*

- b) *Título del documento*. El título se transcribe de la siguiente forma:
Título: antetítulo: subtítulo.
 - c) *Tipo*: Alfanumérico
 - d) *Tratamiento indexación*: Indexado
 - e) *Tratamiento documental*: Lenguaje libre
 - f) *Lengua*: Lengua del documento
 - g) *Controles de validación*: No puede quedar vacío. Si por alguna razón, el documento careciera de título, el documentalista asignará un título descriptivo.
- 2) *Campo Descriptores*
- a) *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, procurando no asignar más de 20 descriptores por documento.
 - b) *Tipo*: Alfanumérico
 - c) *Tratamiento indexación*: Indexado
 - d) *Tratamiento documental*: Lenguaje controlado
 - e) *Lengua*: Del centro de documentación
 - f) *Controles de validación*: No puede quedar vacío y sólo admite valores extraídos de una lista de términos autorizados.
- 3) *Campo: Fecha de publicación*
- a) *Dominio*: La fecha de publicación de la noticia, indicada con el siguiente formato: DD/MM/AAAA.
 - b) *Tipo*: Fecha
 - c) *Tratamiento indexación*: Indexado
 - d) *Tratamiento documental*: No procede
 - e) *Lengua*: No procede
 - f) *Controles de validación*: No admite valores fuera de rango.

Estudiando el ejemplo de diccionario de datos anterior, formado únicamente por tres campos, 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 datos que admite el campo. Los tipos de datos suelen ser: numérico, alfanumérico, fechas y lógico. Un campo numérico permite efectuar cálculos aritméticos o búsquedas por rangos de valores. Un campo alfanumérico admite tanto cadenas de caracteres como números, pero trata a estos últimos como caracteres. Un campo de fechas sólo admite fechas en un formato establecido y permite búsquedas por rangos de fechas o por valores superiores o infe-

riores 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 *Descriptor*, 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 *descriptor*es en castellano, siempre de acuerdo con el diccionario de datos precedente.

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

- 1º) Qué clase de información y cómo entra la información en el sistema.
- 2º) Qué procesos documentales se llevan a cabo.
- 3º) Qué servicios y productos genera el sistema, y/o a qué aplicaciones puede 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.

Por ejemplo, siguiendo con la base de datos de un medio de comunicación (supongamos que se trata de un diario), la descripción funcional podría explicar algo del tenor siguiente:

- Diariamente, entran en un archivo temporal de la base de datos las noticias que se publicarán al día siguiente. Esta entrada se realiza mediante una exportación automática de los archivos informáticos generados por la redacción.
- Los documentalistas revisan las noticias entradas el día anterior, la analizan y asignan *descriptor*es a cada una de ellas, utilizando la lista de *descriptor*es autorizados. En el caso de las fotografías, redactan un título o copian el

pie de foto si resulta adecuado como título y les asignan también los descriptores correspondientes.

- Tanto el texto completo de las noticias como la imagen facsímil de las noticias y las fotos quedan archivados en la base de datos, asociados a los descriptores asignados por los documentalistas, que actúan de puntos de acceso a la noticia, junto con el texto completo del documento en el caso de los artículos.”
- Sobre el modelo de normativa de indexación, únicamente cabe advertir que debe dar indicaciones lo más completas posibles sobre la política de indexación que se recomienda para los campos controlados. En particular, debería recomendarse la aplicación de las normas ampliamente aceptadas por la comunidad internacional. En particular, puede recomendarse la aplicación de la norma UNE sobre construcción de tesauros (Norma UNE 50-106-90; Currás, 1991), así como ofrecer orientaciones y recomendaciones sobre su aplicación concreta a la base de datos que se ha diseñado.
- Por ejemplo, siguiendo el caso anterior, y teniendo en cuenta las características de la documentación periodística, el Modelo de Normativa de Indexación debería proponer una indexación postcoordinada, basada en la utilización de descriptores extraídos de un lenguaje documental controlado que, en su día, puede llegar a ser un tesauro, y para elegir la forma de los descriptores, su nivel de especificidad, etc., se podría recomendar la mencionada norma UNE.

Sobre los extremos relacionados con la descripción documental no nos extenderemos aquí, y tan sólo diremos que tales recomendaciones deben ser tan detalladas como sea posible.

3.3. La ISBD y los modelos canónicos

Por otro lado, no deberíamos olvidar que, en Documentación, la experiencia previa ha dejado bien sentados cuáles son los atributos de algunas entidades e incluso cual 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 testado.

Por ejemplo, los atributos estructurales de cualquier clase de documento pueden ser adecuadamente modelados siguiendo la norma internacional ISBD. Recordemos que esa norma internacional representa 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 audiovisual, 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 la lectura pública.

Entiendo 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 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 una raya”, etc.

4. 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 circular 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. Por ejemplo, adquiriendo primero el programa y el ordenador y diseñando después la base de datos, o pensando en la forma y el tamaño de las estanterías y diseñando después el sistema, etc.

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.

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), ignorando de momento los aspectos físicos, así como comenzar por analizar el sistema objeto y sólo después de conocerlo bien, podemos 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 que, por otro lado, es universal para todo sistema de información:

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:

- | |
|--|
| <ul style="list-style-type: none">- De lo conocido a lo desconocido- De los aspectos lógicos a los aspectos físicos- De lo general a lo concreto |
|--|

Fig. 3. La dirección del diseño en el ciclo de vida de un sistema de información

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 de la figura 4:

<i>1. Análisis</i>	
1a	Análisis del sistema de actividades humanas
1b	Análisis del sistema de actividades humanas
<i>2. Diseño</i>	
2a	Diseño del modelo conceptual
2b	Determinación de los procedimientos de tratamiento documental (descripción, análisis e indexación documental, etc.) si es el caso
<i>3. Implantación</i>	
3a	Elaboración del presupuesto y del calendario de implantación, en su caso.
3b	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.
3c	Instalación, pruebas de rendimiento y reelaboración, 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

Fig. 4. Cuadro sinóptico del ciclo de vida de una base de datos documental

Aunque expresado en fases y enumeradas secuencialmente el proceso parece estrictamente lineal, en realidad, el proceso de diseño también tiene mucho de circular, porque 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., idea que recoge la figura 5.

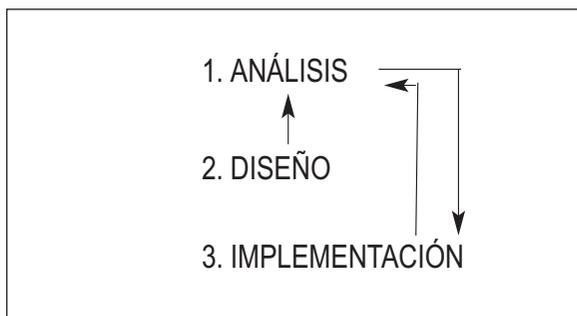


Fig. 5. El ciclo de vida de un sistema de información como un proceso circular

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 como un modo natural 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 haya que rehacerlo 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 de 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 las fases de diseño y de análisis las realice el departa

mento de documentación, mientras que todos los aspectos de implantación corran a cargo del departamento de informática.

Este reparto de tareas parece el más conveniente para todas las partes, ya que sólo el departamento de informática tiene una visión global de la arquitectura informática de la empresa, y sólo el departamento de informática posee la competencia técnica y las atribuciones que le permiten tomar decisiones que afectan a esa arquitectura corporativa.

Sin embargo, lo que nunca es conveniente sustraer es que el proceso de análisis y de diseño se sustraiga del departamento de documentación, que es el único que tiene competencia técnica en los aspectos lógicos de esa clase de sistemas, exactamente por la misma razón que el diseño de una biblioteca lo debe desarrollar un experto en biblioteconomía, aunque su automatización la desarrolle el departamento de informática; el sistema contable, un experto en contabilidad, aunque lo automatice el departamento de informática, etc.

Cada una de las fases precedentes (Análisis, Diseño, Implantación) tiene unos objetivos, debe producir unos resultados concretos y utilizar unas herramientas determinadas.

3.1. La fase de análisis

El objetivo de esta fase es conocer bien aquella parte del mundo real, llamada 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)
- Un sistema de conocimientos (SCO).

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 que debe producir esta fase de análisis es una descripción textual que puede incluir gráficos de ser necesario, sobre el SAH, que suele denominarse *modelo esencial*, y que debe incluir, como mínimo, los siguientes aspectos:

1. Propósito y objetivos del SAH
2. Actores principales del SAH
3. Actividades más relevantes del SAH
4. Entorno del SAH

La herramienta principal aquí es la realización de entrevistas con representantes del SAH y el análisis de cualquier documentación, del y sobre el SAH, que pueda aportar una comprensión global del sistema. Entre tales documentos podemos citar organigramas, documentos fundacionales, memorias, etc.

Aunque el *modelo esencial* consiste, básicamente en una descripción textual, puede incluir, si el analista lo considera necesario, diagramas o gráficos que faciliten su comprensión.

El *modelo esencial* no debe ser muy extenso, sino que, tal como indica su nombre, debe consistir únicamente en una descripción que recoja los aspectos esenciales de la naturaleza y de las actividades del SAH. Además, como una base de datos documental no persigue el modelado de esas actividades, probablemente cinco o seis párrafos deberían ser suficientes para aportar el conocimiento necesario para los objetivos perseguidos.

Este modelo podrá formar parte del producto final, pero no es necesario que sea así, ya que, principalmente su misión es asegurarse de que el responsable del proyecto y otros actores que intervengan en él tienen una adecuada concepción de la naturaleza del SAH.

Por su parte, el propósito de la fase del análisis del sistema de conocimiento consiste en conocer el componente clave en este caso del sistema objeto, a saber, los documentos o las cosas sobre las cuales la base de datos deberá recoger información.

El resultado de esta fase debe consistir en la identificación clara y sin ambigüedades de los documentos o las cosas (entidades) sobre las cuales la base de datos deberá mantener información, así como debe poner de manifiesto las propiedades más relevantes de esas entidades.

La herramienta más adecuada para esta fase, es el modelo entidad-relación (modelo E-R), un modelo bastante intuitivo que, sin embargo, resulta de gran utilidad para enfocar este tipo de análisis. Este modelo se explicará en el apartado dedicado a las herramientas.

3.2. La fase de diseño

El propósito de la fase de diseño es obtener un *modelo conceptual* de la base de datos y una *propuesta de tratamiento documental*. El primero contiene los elementos necesarios para orientar el proceso de implantación. El segundo establece criterios y orientaciones sobre el proceso de descripción y de representación del contenido semántico de los documentos o entidades de los que tratará la base de datos.

Los dos modelos mencionados son el resultado de la fase de diseño y deben ser aprobado 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. Una *definición raíz*, que es una definición, al estilo del modelo esencial, que describa los aspectos más importantes del SAH y del SC, aunque sin necesidad de entrar en detalles, mencionando el propósito de la base de datos y los usuarios del sistema.
2. Una *definición del dominio* de la base de datos.
3. Una *identificación de las entidades* representadas en la base de datos.
4. Un *diccionario de datos*
5. Una *descripción funcional* del sistema.

El *dominio* de la base de datos es el conjunto de los temas o entidades sobre los que mantiene información la base de datos. Como todo dominio, 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 dominio de la base de datos LISA Plus son las *Ciencias de la Documentación*. O puede consistir en una frase, por ejemplo, el dominio de la base de datos TESEO se enuncia diciendo que está formado por *las tesis doctorales publicadas por universidades españolas*.

Las herramientas para producir el documento anterior son, entre otras, las siguientes:

1. La definición raíz.
2. El modelo entidad-relación.
3. El diccionario de datos.

La *definición raíz* expresa qué es la base de datos o, si se quiere, expresa la clase de problemas que podrá solucionar y a qué categoría de usuarios dará servicio. Esta descripción debe mencionar a los usuarios de la base de datos. No debe ser más larga de tres o cuatro párrafos. La información necesaria para construir la definición raíz se obtuvo del *Modelo esencial*, que forma parte de la fase de análisis y que vimos en su momento.

Un ejemplo sencillo podría ser la definición raíz de la base de datos documental de un medio de comunicación podría adoptar la siguiente forma: “El propósito de esta base de datos es satisfacer las necesidades de información retrospectiva de los redactores del diario, permitiéndoles recuperar selectivamente cualquier información publicada anteriormente por el diario”.

Al igual que en la identificación del dominio de la base de datos, elaborar la definición raíz puede ser una tarea fácil e intuitiva, resultado de un mero análisis

técnico, o bien puede ser producto de una refinada decisión política. Lo que es importante es que, sea cual sea el proceso de decisión, ésta quede documentada y expresada y formalmente detallada por escrito.

3.3. La fase de implantación

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

1. Se selecciona el sistema informático (software + hardware) que pueda satisfacer los requerimientos del modelo conceptual y del modelo de normativa de indexación, se realiza la primera instalación y se nombra a un administrador de la base de datos que, a partir de ahora, será el máximo responsable de ella.
2. Se realizan pruebas con una colección-test de documentos.
3. Se realizan los cambios o ajustes necesarios.
4. Se procede a la formación del personal técnico y de los usuarios finales.
5. Se edita la versión 1 del *Libro de estilo de la base de datos*, que incluye:
 1. La versión definitiva del modelo conceptual.
 2. La normativa de tratamiento documental, en su caso.
6. Acciones de promoción, formación de usuarios finales.
7. Definición de una política de mantenimiento y explotación, para lo cual será también esencial disponer del libro de estilo mencionado en el punto 5. En esta política de mantenimiento se detallan los procesos y las rutinas de carga de datos, consulta y explotación.

5. 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.

6. Notas

- (1) El término Sistema de actividades humanas es uno de los conceptos característicos de la SSM de Checkland.

7. Bibliografía

- AENOR. (1990). Norma UNE 50-106-90. Documentación.: Directrices para el establecimiento y desarrollo de tesauros monolingües. Madrid : AENOR, 1990.
- Checkland, P. B. (1981). Systems thinking, systems practice. Chichester : Wiley, 1981.
- Checkland, P. B.; Scholes, J. (1990). Soft systems methodology in action. Chichester : Wiley , 1990.
- Chen, P. P-S. (1976). The entity-relationship model : towards a unified view of data. // ACM transactions on databases systems. 1 : 1 (1976) 9-36.
- Codina, L. (1994a). Sistemes d'informació documental: concepció, anàlisi i disseny de sistemes de gestió documental amb microordinadors. Barcelona : Pòrtic, 1994.
- Codina, Lluís (1994b). Modelo conceptual de un sistema de información documental. // Revista Española de Documentación científica. 17 : 4 (Octubre-diciembre 1994) 440-449.
- Codina, L. (1995). Metodología de creación de bases de datos documentales : Parte I. // Information world en español. 33 (abril 1995) 10-11.
- Codina, L. (1995). Metodología de creación de bases de datos documentales : Parte II. // Information World en Español. 34 (mayo 1995) 9-12.
- Currás, E. (1988). La información en sus nuevos aspectos. Madrid : Paraninfo, 1988.
- Currás, E. (1991). Thesaurus: lenguajes terminológicos. Madrid : Paraninfo, 1991.
- Jackson, G. A. (1990). Introducción al diseño de bases de datos relacionales. Madrid : Anaya, 1990.
- Lewis, P. (1994). Information systems development. London : Pitman, 1994.
- Van Slype, Georges. (1991). Los lenguajes de indización: concepción, construcción y utilización en los sistemas documentales. Madrid : Fundación Germán Sánchez Ruipérez, 1991.
- Walker, D.W. (1991). Sistemas de información basados en ordenador. Barcelona : Marcombo, 1991.
- Yourdon, E. (1993). Análisis estructurado moderno. México : Prentice-Hall Hispanoamericana, 1993.