# DISEÑO DE UNA BASE DE DATOS DE PRENSA CONTROLADA POR UN LENGUAJE FACETADO DE ESTRUCTURA COMBINATORIA («THESAURUS»)

**José María Izquierdo Arroyo***
**Luis M. Moreno Fernández***

**Resumen:** Este proyecto se inscribe en el marco del Plan de Información y Documentación de las Comunidades Autónomas, uno de cuyos objetivos es favorecer la creación y coordinación de Bases de Datos. Describe la construcción y el mantenimiento de una base de datos y un lenguaje documental de estructura combinatoria («Thesaurus»), para el tratamiento de la documentación de actualidad referida a la Comunidad Autónoma de la Región de Murcia. Incluye los diversos tipos de registros que integrarán las Bases de Datos, el equipo informático propuesto para el almacenamiento y recuperación de la información, las características del Thesaurus que ha de controlarla y el plan de trabajo a seguir en la elaboración de ambos productos. Tareas que se realizan de modo conjunto.

**Palabras clave:** Base de datos, recuperación de la información, prensa, lenguajes de indización, thesaurus.

**Abstract:** This project has been developed within the framework of the Plan de Información y Documentación de las Comunidades Autónomas (Documentation and Information Plan of the Spanish Autonomous Communities) having as a specific aim the creation and coordination of informational data bases. It describes the development and housekeeping of a data base and a combinatory-structured documentary language for the management of current general information related to the Murcian Comunidad Autónoma. The various record structures managed by the data base are dealt with, as well as the specific type of hardware suggested for data storage and retrieval, constituent features of the controlling thesaurus, and work guidelines in developing and implementing these products in an integrated fashion.

**Key words:** Database, information retrieval, press, indexing language, thesaurus.

## 1. Antecedentes y estado actual del tema

Uno de los rasgos distintivos de las sociedades en esta segunda mitad del siglo XX es la consolidación de una industria de la información que trata de organizar y servir a los distintos usuarios la información que se produce. Consolidación que corre pareja a la diversificación de la oferta de información en un mercado cada vez más amplio. De modo que a las bases de datos de ciencia y tecnología servidas por un gran distribuidor en la década de los 70 vienen a sumarse otras relacionadas con las actividades económicas (bolsa, empresas, legislación, patentes, prensa...), que experimentan un crecimiento importante porque están relacionadas con el desarrollo económico de los países. A ellas hay que añadir las relacionadas con la vida cotidiana, servidas por videotex, y las distribuidas en CD-ROM y

* Universidad de Murcia
Recibido 15-9-91

tecnologías ópticas en general, que ofrecen alta capacidad de memoria, seguridad en el manejo de los datos y flexibilidad (1, 2, 3).

En nuestro país esta industria no ha alcanzado aún pleno desarrollo, pero continúa expandiéndose desde que surgiera en la década de los 70, siguiendo la tónica general.

Las bases de datos de prensa no han sido una excepción. La débil presencia de ellas en el seno de la industria de la información es paliada parcialmente por BARATZ, DINESA y EFEDATA. Aparte de éstas cabría mencionar la de Información Española de Actualidad, constituida en 1982 en el seno del Departamento de Documentación de la Facultad de Ciencias de la Información de la Universidad Complutense de Madrid. A estos esfuerzos se suman los de aquellos diarios españoles que están poniendo en marcha sus propias bases de datos.

En el plano internacional hay bases de datos de prensa con unas funciones similares a las que debe tener la que hemos proyectado, pues son de texto íntegro y circunscriben su radio de acción a escala regional y local. Nos referimos a las bases de datos «Documentation Le Meridional», producida por Le Meridional, en funcionamiento desde 1981; y a «Documentation La Provençal», producida por Le Provençal, y que presta servicio desde 1984. Ambas distribuidas por SEMITEL.

También de prensa y texto íntegro, aunque de ámbito nacional, son «The Guardian» —Reino Unido—, de una temática variada, semejante a la nuestra, producida por The Guardian y funcionando desde 1984; «Today», igualmente del Reino Unido, de temática sociodeportiva, producida por Today. Las distribuye PROFILE. Por no abundar en una descripción farragosa de bases de datos análogas, mencionaremos tan sólo los nombres de aquellas consideradas más relevantes: INFOBANK (1972-); BUSINESS WIRE (1986); NEWSEARCH (1985-); NEXIS (1970-); The Washington Post (1983).

El común denominador de la gran mayoría de ellas es que tienen como finalidad primordial el tratamiento de la información para la elaboración de la noticia, y nuestro propósito es relativamente distinto: Proyectamos organizar la base de datos —y el «Thesaurus» correspondiente— para el tratamiento de la noticia ya elaborada, desechando, por ejemplo, fuentes o datos que sí aprovechan aquéllas, tales como la información que suministran los teletipos.

En virtud de esta diferencia funcional creemos que no hay bases de datos de esta clase a escala regional. Las bases de datos de la Comunidad Autónoma de la Región de Murcia tienen por finalidad, por ejemplo, informar al ciudadano acerca de aspectos de su vida cotidiana diferentes a los aquí tratados: callejero, guía turística, etc., amén de alimentarse de fuentes ajenas a las que queremos explotar. Es por eso por lo que se conocen popularmente como «Servicios de Información al Ciudadano». Esto no obstante, en cierta medida, dependiendo del desarrollo que adquiera esta base de datos de prensa, es posible contemplarla como un servicio complementario de los anteriores.

En el plano nacional sabemos que se han construido bases de datos similares en otras Comunidades Autónomas. Así, en el País Vasco ha sido creado recientemente un archivo iconográfico adscrito a la Secretaría General de la Presidencia, precediéndole una base de datos de información política (4).

## 2. Objetivos del proyecto

Entre los objetivos del PLAN DE INFORMACION Y DOCUMENTACION DE LAS COMUNIDADES AUTONOMAS se ha destacado la necesidad de «favorecer y coordinar la producción de Bases de Datos que permitan, entre otros aspectos, controlar la documentación producida en cada Comunidad Autónoma, sistematizar y recuperar la información estadística sobre su ámbito de actuación y producir los directorios de información individualizada necesarios para identificar las actividades de los organismos públicos y empresas privadas».

Partiendo de estas directrices, el presente trabajo quiere ser una respuesta, no una panacea, a la necesidad que tiene el mundo de hoy —la España de hoy— de organizar la información producida con vistas a su mejor aprovechamiento. De ahí que nuestro proyecto se encamine a la construcción de una base de datos y a arbitrar los instrumentos oportunos para recuperar la información almacenada en ella. Sin embargo, conscientes de la imposibilidad de asumir en su totalidad los objetivos del Plan de Información y Documentación de las Comunidades Autónomas, por lo limitado de los recursos materiales y humanos disponibles, así como por el período de tiempo establecido para llevar a cabo el trabajo, la base de datos ha de nutrirse de fuentes muy concretas —prensa fundamentalmente—, pero que contienen una carga informativa de sumo interés para la ciudadanía en general y los gobernantes y la Administración en particular. En consecuencia, éstos son los objetivos propuestos:

2.1. Construcción de una base de datos para la Comunidad Autónoma de la Región de Murcia con documentación de actualidad —diarios, publicaciones periódicas— en sus vertientes textual e icónica, al objeto de que usuarios de distintos perfiles —gabinete de prensa de la C.A., gobernantes, políticos, funcionarios, público en general e investigadores— puedan ver satisfechas sus demandas de información en tres niveles:

a) Respuestas del sistema a preguntas concretas.
b) Seguimiento de determinados temas durante un período cronológico (B.R.)
c) Elaboración de perfiles selectivos en virtud de las diversas demandas de información del usuario (D.S.I.).

El almacenamiento estructurado de la noticia (base de datos) implica someterla, previa selección, a las habituales operaciones de tratamiento documental: descripción externa —identificación—, descripción substancial (resumen) y descripción característica —indización—. Así, la base de datos ha de permitir efectuar la búsqueda de referencias —asientos, palabras clave, etc.— y la recuperación del texto o los documentos icónicos almacenados en forma digitalizada en la memoria de masas —disco óptico—.

2.2. La recuperación de la información de modo pertinente y exhaustivo requiere el empleo de un «Thesaurus». Elaborarlo es el segundo objetivo de este trabajo, y, en nuestra opinión, se encuentra tan imbricado con el anterior porque no puede obtenerse información de una base de datos de prensa a plena satisfacción si no se cuenta con un «Thesaurus», intermediario preciso entre la información introducida en aquélla y el usuario. De manera que empresas con servicios de

documentación similares al que nos estamos refiriendo —caso de El Correo Español del Pueblo Vasco – han creído rentable acometer la tarea de elaborar uno. Camino que mucho antes abrieran el New York Times y Le Monde. En realidad, las bases de datos relativas al vasto campo de las ciencias sociales y humanas ofrecen un mayor rendimiento cuando utilizan un lenguaje controlado en la indización y recuperación de documentos.

Considerando el carácter general de la información que acerca de nuestra Región traen los diarios y publicaciones periódicas, así como el mencionado perfil de los usuarios, el «Thesaurus» proyectado, además de permitir el control terminológico y ser flexible —revisión y actualización—, debe estar sectorialmente alejado de la especialización —vid. apartado correspondiente—.

2.3. El objetivo final de cualquier centro de documentación es facilitar al usuario la información que pide, ya de modo directo, ya arbitrando los procedimientos adecuados para que aquél sea capaz de obtenerla por sí mismo.

Ahora bien, este objetivo resulta inalcanzable si antes no se ha planificado correctamente la difusión de la información sobre la base de tres variables: el presupuesto del que se dispone; la situación institucional del centro, es decir, su dependencia con respecto a otro; y la tipología de los usuarios.

Habida cuenta de que la base de datos se incardina en la Administración regional, que prevé dar servicio en una primera fase a su gabinete de prensa, y posteriormente al resto de la Administración y público en general, y que para ese menester dispone de un equipo informático que se pretende conectar en red, estimamos que a raíz de la entrada en funcionamiento del sistema —primera etapa— el servicio de difusión estará en condiciones de ofrecer las siguientes prestaciones:

a) Búsquedas retrospectivas (B.R.).
b) Prestación de documentos primarios de la base de datos en soporte papel, previa solicitud, tras reproducirlos por impresora, o directa y libremente mediante red a miembros del gabinete de prensa en principio, y a consejeros, funcionarios y organismos integrantes de la estructura políticoadministrativa de la Comunidad Autónoma más adelante.
c) Confección de «revistas» y «dossiers» de prensa.
d) En el marco de la difusión de documentos secundarios, se cree de gran interés la preparación de boletines de resúmenes.
e) Difusión Selectiva de la Información (D.S.I.), según los «perfiles de interés» de los distintos focos de usuarios, previo control del flujo de la información.

A tal efecto se diseñará y mantendrá un archivo de usuarios —fig. 1—, con especificación de los distintos tipos y categorías de intereses, que estará organizado conforme a un organigrama jerárquico, agrupando campos y subcampos y especificidad dentro de los organismos, competencias y departamentos. La incorporación al registro de los usos concretos del sistema de documentación facilitará el mantenimiento y actualización de la tipología del usuario. El modelo de ficha de lo que será el registro contempla asimismo las prestaciones solicitadas (B.R., D.S.I.), incluyendo la especificación de las consultas formuladas y la información suministrada —por referencia al archivo textual—.

Trazados, previo estudio de campo, los «perfiles» de los usuarios —habituales y ocasionales—, en virtud de las necesidades internas y de proyección al exterior del Servicio de Documentación, se hará llegar con regularidad a los demandantes del servicio de D.S.I. el tipo de información requerida desde el momento en el que se efectuara la B.R. El período tipo de difusión será inicialmente mensual.

Un tratamiento computadorizado permitirá la elaboración de n diarios «a la medida» para los diversos organismos que configuran la estructura orgánica de la C.A. La modalidad de edición de las fichas de novedades que integran ese «diario» —a la medida— podrá ser, según convenga, periódico-libraria o informática: mediante computador personal conectado.

**Figura 1**

Fecha ..... Demanda N.º .....

Nombre.....

Apellidos .....

Sexo V ..... H .....

Fecha y lugar de nacimiento.....

Ocupación.....

C/..... Núm. ..... Piso ..... Puerta ..... Tf. .....

C.P. ..... Localidad.....

A RELLENAR SOLO POR FUNCIONARIOS Y PERSONAL ADSCRITO A LA A.R.

1. ADSCRIPCION ORGANICA:

1.1. Consejería.....
1.2. Centro directivo.....
1.3. Servicio .....
1.4. Sección.....
1.5. Negociado.....

2. LOCALIZACION:

2.1. Calle ..... Núm ..... Piso ..... Puerta ..... Tf ..... Ext.....
2.2. C.P ..... Localidad.....

3. RESPONSABLE:

3.1. Cargo .....
3.2. Primer apellido ..... Segundo apellido .....
3.3. Nombre..... Tf. ..... Ext. .....Horario M ..... Horario T .....

4. FINES/COMPETENCIAS/ACTUACIONES:

.....
.....
.....

5. TIPIFICACION FINES/COMPETENCIAS/ACTUACIONES:

A. ... Servicios administrativos sin repercusión externa a la Administración.
B. ... Materia de empleo.
C. ... Tramitación y/o concesión de ayudas.
D. ... Tramitación y/o concesión de becas.
E. ... Tramitación y/o concesión de subvenciones.
F. ... Promoción, organización de actos culturales.
G. ... Promoción, organización de actos deportivos.
H. ... Publicaciones.
I. ... Recaudación.
J. ... Pagos.
K. ... Otros. Especificar.

BUSQUEDA RETROSPECTIVA

Tema objeto de la búsqueda .....
Limitaciones en la búsqueda: 19..... a 19.....
N. de referencias obtenidas .....
N. satisfactorio de referencias obtenidas .....

DIFUSION SELECTIVA DE LA INFORMACION

Tema .....
Descriptores: Cite 4 ó 5 términos relacionados con el tema solicitado .....
Tipo de documento: Artículo de diario ..... Artículo de revista .....
Periodicidad: Mensual ..... Bimestral .....Trimestral ..... Semestral .....
Usuario: Habitual ..... Ocasional.....

Fin del servicio:

Este S.D.S.I. se proporcionará durante un año, a partir de la fecha de solicitud. Una vez concluido, cesará automáticamente si antes no se ha solicitado prórroga del mismo.

2.4. La investigación desarrollada genera, pues, una serie de productos documentales:

a) Libros:
«Thesaurus».
Manual de manejo y acceso a la base de datos.
Manual de indización y resumen para analistas.
Glosario-diccionario —coordinado en el «Thesaurus»—.
Repertorio biográfico —coordinado informáticamente con el archivo onomástico.
b) Publicaciones periódicas.
Boletín de noticias.
Informes colectivos para distintos organismos y departamentos.

## 3. Metodología y plan de trabajo

En cuanto se refiere al sistema del tratamiento del contenido en las bases de datos de prensa, baste con citar el utilizado por INFOBANK —New York Times—, descrito por M. Caridad entre otros (5).

Los Sistemas de Gestión Documental que sorportan las bases de datos han proliferado en la última década. Pero todos los autores coinciden en señalar que un Sistema de Gestión Documental debe satisfacer los requisitos que en forma sumaria exponemos:

— Estructuración de la base de datos diseñada por el documentalista.
— Gestión de multibases.
— Activación de ficheros inversos.
— Almacenamiento de ficheros en formato variable sobre disco.
— Búsquedas booleanas y otras.
— Factor de expansión razonable.
— Gestión de léxicos y/o «Thesauri».
— Texto íntegro.
— Capacidad aritmética.
— Edición de documentos secundarios, etc. (6, 7, 8, 9, 10, 11).

En virtud de estos requerimientos hemos seleccionado el siguiente sistema informático:

— Ordenador «Hewlett-Packard Vectra RS/25C, microprocesador Intel 80386 a 25 Mhz., disco duro de 330 Mb., disketera 3"-1/2 de 1,44, RAM de 4 Mb., monitor color súper VGA.
— Sistema operativo UNIX
— Software de Gestión Documental BRS para trabajar en el entorno hardware arriba expuesto y con el sistema UNIX (multiusuario):
    a) BRS/SEARCH: Módulo principal.
    b) BRS/DEMON: Editor de pantallas.
    c) BRS/THESAURUS.
— Software de captura, almacenamiento y recuperación de imágenes de ScripNet.
— Servidor de disco óptico con controladora SCSI: Juke-Box Sony WDA-610 (con capacidad hasta 50 discos de 328 Gbytes).
— Escáner: TDC DS-4270.
— Impresora LaserJet III.

El plan de trabajo y su metodología se detallan en tres fases, a partir de la creación de la base de datos, la construcción, mantenimiento y evaluación del «Thesaurus», y el desglose de tareas con diagrama de tiempos.

### 3.1. Creación de la base de datos

En el tratamiento documental de la noticia, tanto textual como icónica, seguiremos las conocidas operaciones de la cadena documental: Selección, descripción física, indización y resumen... en teoría; la práctica, como veremos, no se

ajusta exactamente a este orden. A esas operaciones, podría añadirse la de «informe», es decir, el tratamiento relacional de contenidos (12).

### 3.1.1. *La selección de la noticia*

Para evitar que la acumulación de documentos lastre el trabajo del equipo consumiendo tiempo, energías y recursos materiales caros y limitados, es preciso proceder a seleccionar las noticias generadas por las publicaciones regionales y nacionales.

Obviamente el primer criterio de selección lo impone la delimitación territorial, pues sólo nos conciernen aquellas noticias acaecidas en el marco geográfico de la C.A. de la Región de Murcia, y las de ámbito nacional que incidan de modo directo en la Región o se refieran expresamente a ella.

El segundo nivel de selección lo determina la temática que interesa al usuario del sistema. Y su exponente es el índice analítico del «Thesaurus», desglosado en un apartado posterior.

En el tercer nivel de selección juegan varios elementos. Así, antes de analizar las noticias seriadas, serán sometidas a un seguimiento puntual para establecer su grado de veracidad —entendiendo por veracidad aquí el que no se trate de un error o de un rumor— y relevancia, desestimando las noticias reiterativas o falsas. Con este fin se confeccionarán «dossiers» sobre personas, acontecimientos, etc., ya que el examen del conjunto de documentos relativos a un tema permite seleccionar con mayor rigurosidad la documentación a analizar en una fase posterior. Siempre se pondrá especial cuidado en recoger los diversos enfoques que acerca de las noticias proporciona la prensa de diferente signo, o sea, n periódicos, siendo $n \geqslant 2$.

La selección se llevará a cabo en equipo, para normalizar procedimientos y evitar errores o apreciaciones subjetivas en el curso de esta tarea.

### 3.1.2. *Descripción física del documento: Registro, fichero, base de datos*

La base de datos está constituida por dos subsistemas de archivos. El primero comprende a su vez dos tipos de ficheros: uno de carácter textual y otro icónico. El segundo lo integran tres ficheros: el de usuarios, el onomástico —biográfico— y el geográfico.

Para ambos tipos de archivos se utilizará el sistema de acceso directo, por diseñarse con el criterio de relación biunívoca entre unidades documentales y campos de descripción (13). La correlación entre los registros está asegurada asimismo, debido a la homogeneidad de los criterios de descripción, y al establecimiento de campos formales —características físicas del documento— y semánticos —referidos al contenido—. Se pretende con ello establecer enlaces cruzados con la informción de personas o materias relacionadas entre sí.

Cada uno de los registros que componen los ficheros textual e icónico se corresponden, «a priori», con una unidad de contenido informativo, si no surgen datos colaterales o complementarios que aconsejen la apertura de nuevos registros.

Ambos modelos de registros llevan un campo de referencia que permite el acceso de uno a otro indistintamente, en el supuesto de que sean complementarios.

Por ejemplo; un reportaje acompañado de fotos contará con dos campos en su correspondiente registro donde conste si va acompañado de fotos, y qué número de registro se adjudicó a éstas en el archivo icónico: y viceversa, el documento icónico ha de remitir a su contexto si lo hubiere.

El registro del fichero textual y el del geográfico están basados en los modelos que expusieran G. Gutiérrez y Lucas Fernández (1987), aunque hemos introducido modificaciones en ellos; por ejemplo, nosotros no contemplamos la posibilidad de incluir asientos bibliográficos.

Los registros que integran el primer subsistema de archivos —textual e icónico— están formados por tres tipos de campos. Los del fichero de carácter textual son:

a) Campos que describen las características externas del documento: tipo, autor, fuente, edición, etc.
b) Entradas —llamémoslas así, aunque no se haya previsto acceder al registro a través del resumen— que describen el contenido del documento. En el campo de descriptores procuramos acoger «la mayor cantidad de Laswellianos —qué, quién, cómo, por qué, dónde, cuándo y para qué o con qué finalidad—».
c) Campos que aportan datos de interés para el documentalista, tales como fecha, análisis, analista, pero que no son recuperables —fig. 2.

**Figura 2**

N. de referencia ..... Fecha ..... de entrada en la B.D.

| | |
|---|---|
| A. ASIENTO HEMEROGRAFICO | A1 TIPO DOC.<br>A2 FORMALIT.<br>A3 AUTOR<br>A4 TITULO<br>A5 FUENTE<br>A6 EDICION<br>A7 SECCION<br>A8 AGENCIA<br>A9 PAG.<br>A10 COL.<br>A11 ART.<br>A12 FECHA PUB.<br>A13 FECHA NOTICIA<br>A14 ICONOS<br>A15 N. REGISTRO ICONO(S) |
| B. CONTENIDO | B1 RESUMEN<br>B2 MATERIA<br>B3 DESCRIPTORES: Qué, cómo, por qué, para qué o con qué finalidad.<br>B4 ONOMASTICOS: Quién<br>B5 TOPONIMOS: Dónde |
| C. CONTROL | C1 ANALISTA<br>C2 F. ANALISIS |

Los campos recuperables forman el índice inverso de la base de datos, en tanto que los no recuperables pasan a integrar el índice secuencial de aquélla.

Los documentos del archivo icónico reciben un tratamiento similar. Una vez seleccionados y analizados engrosan los correspondientes registros diseñados «ad hoc»; y al igual que en el caso anterior existen tres tipos de campos: Los de utilidad para el documentalista —ref., analista, fecha análisis— que le sirven para llevar un control de las fichas que se introducen a diario y conocer su autoría; los destinados a describir los rasgos externos de la imagen —formato, soporte, fecha, lugar, número, fotos asociadas, autor y procedencia—; y aquellos que describen el contenido del documento —título, materias, descriptores, lugar, personas, resumen. Fig. 3.

**Figura 3**

N. de referencia .....  Fecha ..... de entrada en la B.D.

| | |
|---|---|
| A. ASIENTO HEMEROGRAFICO | A1 NUMERO: Asignado al doc. colocado en el archivo fotográfico, y que es importante para su recuperación.<br>A2 CARPETA<br>A3 N. FOTOS ASOCIADAS<br>A4 COPIAS<br>A5 SOPORTE<br>A6 CALIDAD<br>A7 FORMATO<br>A8 REPORTAJE<br>A9 FECHA FOTO<br>A10 LUGAR FOTO<br>A11 AUTOR<br>A12 TITULO<br>A13 PROCEDENCIA<br>A14 CON TEXTO<br>A15 N. REGISTRO TEXTO |
| B. CONTENIDO | B1 RESUMEN<br>B2 MATERIA<br>B3 DESCRIPTORES<br>B4 ONOMASTICOS<br>B5 TOPONIMOS |
| C. CONTROL | C1 ANALISTA<br>C2 F. ANALISIS<br>C3 UTIL ..... FECHA ..... LUGAR .....<br>RAZON ..... PREST ..... DATOS.....<br>C4 EXPURGO ..... FECHA ..... RAZON..... |

Aunque el orden de los distintos campos es irrelevante a efectos de acceso informático, se ha procurado buscar la ordenación lógica más apropiada para las secuencias de lectura del usuario.

El registro del segundo subsistema de archivos, dejando a un lado el modelo para usuarios, expuesto ya en la primera parte de este proyecto —fig. 1— ofrece estas características:

a) La ficha biográfica —relacionada con el campo de descriptores onomásticos del fichero textual—, está concebida de manera simple; refleja los datos personales, tales como nombre y apellidos, padres, estudios y ocupación. Otro campo, con una extensión de 60 caracteres, recoge la «circunstancia sociopolítica» de la persona cuya microbiografía nos interesa. Datos de gran valor para la investigación histórica y sociológica, porque permiten seguir la trayectoria vital de un ser humano, estudiar élites, etc., sin cortapisas de ninguna clase. De suerte que una persona puede figurar registrada en el archivo onomástico tantas veces como se considere oportuno. Por ejemplo: Fulano de Tal, unas veces aparecerá —«circunstancia sociopolítica», estudios, ocupación— como masón y diputado por el grupo X, ingeniero y director gerente, y otras como Alcalde independiente, etc.

Hemos renunciado a introducir una biografía en una ficha no sólo por no desperdiciar memoria, sino porque ni la más compleja de las fichas llegaría a ser lo bastante completa como para albergar todos los datos de interés que aparezcan en el decurso del tiempo acerca de un individuo —fig. 4.

**Figura 4**

N. de referencia ..... Fecha ..... de entrada en la B.D.

Apellidos.....
Nombre.....
Apodo/Título .....
Sexo V ..... H .....
Fecha y lugar de nacimiento .....
Estudios ..... Fecha inicio ..... Fecha terminación .....
Ocupación ..... Fecha inicio ..... Fecha terminación .....
Residencia ..... Fecha inicio ..... Fecha terminación .....
Circunstancia sociopolítica.....

Con texto
N. registro de texto .....
Materia/Personas con las que se relaciona.....

DATOS DEL PADRE

Apellidos.....
Nombre.....
Estudios ..... Fecha inicio ..... Fecha terminación .....
Ocupación ..... Fecha inicio ..... Fecha terminación .....
Residencia ..... Fecha inicio ..... Fecha terminación .....
Circunstancia sociopolítica.....

DATOS DE LA MADRE

Apellidos.....
Nombre.....
Estudios ..... Fecha inicio ..... Fecha terminación .....
Ocupación ..... Fecha inicio ..... Fecha terminación .....
Residencia ..... Fecha inicio ..... Fecha terminación .....
Circunstancia sociopolítica.....

b) El archivo geográfico está relacionado con el campo de descriptores topónimos del archivo textual y organizado por entidades de población. Detrás de estos encabezamientos geográficos, estructurados por cada mes, aparecen los registros —con los niveles analíticos que señalamos antes, al tratar los archivos del primer subsistema— en orden cronológico de días o semanas. Fig. 5.

**Figura 5**

MURCIA-1990

ENERO
El Palmar
1 Registro
2 Registro
3 .....
Monteagudo
4 Registro
.....

Los registros llevan un código topográfico que tiene dos funciones: conocer el número de noticias que se han generado en un territorio durante un año —datos útiles para efectuar análisis infométricos— y localizar la referencia de que se trate desde el índice de descriptores. En el índice general de la base de datos, cada registro va acompañado de una referencia alfanumérica en la que se aprecia el nombre de la entidad geográfica, seguida del año y el número de la noticia. Fig. 6.

**Figura 6**

Cartagena - 90 - 15 (Noticia 15 correspondiente a 1990 en Cartagena).

3.1.3. *Indización y resumen*

La indización automática se revela hoy por hoy más eficaz en el tratamiento de la literatura científica, portadora de una terminología más homogénea que la que

traen las fuentes manejadas en este trabajo. Es por eso por lo que hemos optado por la indización humana, más lenta —se emplean de 5 a 15 minutos por texto, dependiendo de su extensión, complejidad, profundidad de la indización y familiaridad del indizador con el tema que maneje— que la automática, pero de mayor eficacia en el terreno documental en el que habremos de movernos.

Aunque algunos autores sostengan la conveniencia de operar sobre el texto original, la indización se verificará sobre resúmenes, porque en la modalidad adoptada, el resumen informativo, quedan reflejadas todas las partes pertinentes de la estructura y esencia de la noticia, incluido el vocabulario controlado, en aras de mejorar la recuperación de la información. Esto no obstante, también contemplamos la posibilidad de extraer los términos no sólo del campo documental expresado, sino de diccionarios, glosarios y otros «Thesauri» existentes cuya temática nos concierna, siempre que consideremos que pueden resultarnos de utilidad.

A fin de comprobar la calidad de la indización, cuya profundidad media situamos entre 8 y 12 términos, someteremos a análisis el listado de «descriptores candidatos», midiendo sus tasas de coherencia y especificidad o pertinencia (14).

## 3.2. El Thesaurus

El lenguaje documental propuesto es un lenguaje combinatorio de descriptores estructurados en un «Thesaurus», con sistema facetado. Las facetas se hacen corresponder con un desarrollo del paradigma de Lasswell, adecuado a la información de actualidad. Se incorporará al mismo un sublenguaje sintáctico —por nexos y/u orden— ligado a la estructura de los resúmenes informativos, y, eventualmente, a la fijación de referencias a documentos primarios para la redacción de informes o reseñas de actualidad.

### 3.2.1. *Construcción del «Thesaurus»*

Una vez seleccionadas las fuentes —prensa, diccionarios, glosarios, «Thesauri», etc.—, realizada la indización y sometidos a análisis los términos candidatos a descriptores —predescriptores—, se ordenarán alfabéticamente, procediéndose luego a elaborar un listado de microthesauri —microdisciplinas— y a distribuir los términos entre ellos según su afinidad. La posibilidad de consultar en el fichero inverso de la base de datos todos los términos por orden alfabético, y de conocer el número de registros en el que aparecen (número de ocurrencias), permite disponer de un glosario mediante el cual conocer el contenido temático de aquélla, depurar errores y manejar una información de base con la que elaborar el Thesaurus.

Agrupados los términos por categorías temáticas, toca la elaboración del lenguaje documental propiamente dicho. Para ello es menester: a) depurar ese vocabulario, eliminando sinónimos, casi sinónimos y términos demasiado específicos; b) establecer las relaciones semánticas de sustitución, que permitan el paso del lenguaje natural al lenguaje documental y precisar el sentido de los términos ambiguos a base de notas de alcance (SN); y c) estructurar el lenguaje mediante

una red de relaciones semánticas —jerárquicas y asociativas— entre las palabras controladas.

Es en este punto cuando la informática puede ayudarnos, enriqueciendo la indización humana por el procedimiento de «autorreenvío». En efecto, algunos programas de gestión documental disponen de una opción de «autorreenvío genérico», en virtud de la cual todo descriptor asignado por la persona indizadora se verá automáticamente completado por la máquina gracias a otros descriptores vinculados con ése en el «Thesaurus» con una relación jerárquica ascendente —remite al B.T. desde el N.T. Esto significa que es posible insertar niveles de especificidad en una determinada faceta del indizado. Finalmente, la validación de la indización la realiza el programa informático sobre los indizados de los registros, comparando los descriptores asignados a los documentos con los del «Thesaurus» (15, 16, 17, 18, 19, 20, 21, 22, 23).

### 3.2.2. *Características cualitativas y cuantitativas del «Thesaurus»*

Dados el perfil del usuario, la naturaleza de las fuentes y la función de la base de datos, el «Thesaurus» no debe ser monotemático, sino que en principio englobará diversas grandes categorías temáticas, escogidas a título de hipótesis de trabajo de la división de toda actividad humana en sectores productivos desde Colin Clark, y de las secciones que trae la prensa. Por lo tanto, estas categorías temáticas, en principio —repetimos— son:

— Agricultura, ganadería, silvicultura y pesca.
— Asuntos sociales.
— Ciencia y tecnología.
— Comunicaciones.
— Comunidades europeas.
— Cultura y Humanidades.
— Educación y Universidad.
— Empresa.
— Fuerzas Armadas.
— Industria y Energía.
— Leyes.
— Medio ambiente.
— Negocio y finanzas.
— Ocio, diversión.
— Onomásticos.
— Política y Gobierno.
— Relaciones exteriores.
— Topónimos.
— Trabajo y Seguridad Social.
— Transporte.
— Turismo.

La norma ISO 2788 (1986) distingue tres modelos básicos de «Thesauri»: Alfabéticos, sistemáticos y gráficos. Pues bien, el «Thesaurus» proyectado consta

de dos partes esenciales: una sistemática y otra alfabética. Las describimos de modo somero a continuación.

La primera presenta los descriptores ordenados por categorías temáticas y facetas, puesto que lo más frecuente es que en su elaboración se utilicen simultáneamente ambas técnicas. La complementa un índice permutado, que facilita la búsqueda inicial de los distintos términos —ya sean sintagmáticos o no— y remite al usuario mediante un código al apartado anterior.

En la segunda parte, descriptores y no descriptores aparecen dispuestos siguiendo una secuencia alfabética única. Los no descriptores van acompañados de reenvíos (USE, USE FOR), hacia el término preferente. Los descriptores pueden llevar anejas notas explicativas (SN). Y también se establecen los reenvíos hacia los términos más genéricos (BT, TT), los términos más específicos (NT) y los términos asociados (RT) y el sistema de códigos que remite al usuario a la parte sistemática.

Muy en síntesis, éstos son los tipos de relaciones: Jerárquicas; específico-genérico, todo-parte; asociativas; monoequivalencia semántica —EM, EP—; notas explicativas: definición, nota de aplicación, nota histórica; sistema de facetas agente, instrumento, modo, acción, materia, etc.

Además, está previsto incluir en el «Thesaurus» índices auxiliares que coadyuven a controlar terminológicamente los conceptos básicos de las materias susceptibles de merecer la atención de los usuarios del sistema. Nos referimos a los índices geográfico y onomástico.

Las características cualitativas de la obra son:

a) Representación de «conceptos» mediante descriptores estándar; descriptores auxiliares para «infraconceptos» —tipo súper—; explicitación de reglas de composición a partir de componentes semánticos.
b) Como ratio de precoordinación cada descriptor contendrá una media de 1,5 a 2 «palabras significativas».
c) La ratio entre el número de no descriptores y descriptores —«tasa de equivalencia»— se situará entre 0,5 y 2. Esta tasa relativamente elevada —gran número de descriptores— propicia la coherencia y precisión de la indización, al aumentar el número de entradas alfabéticas al «Thesaurus».
d) La ratio entre el número de relaciones jerárquicas y asociativas y el de descriptores —«la tasa de enriquecimiento»— habrá de situarse entre 1 y 3; cada descriptor está ligado al menos con otro descriptor.
e) Flexibilidad: La ratio de «palabras significativas» simples utilizadas para la composición de descriptores se situará en torno a 0,6.

El «Thesaurus» tendrá un tamaño estimado en unos 1.000 descriptores y será monolingüe. Y en cuanto a sus características formales, ofrecerá una presentación nominal, con términos cuya longitud no excederá de 50 caracteres, preservando la menor tasa de precoordinación, y con riqueza tipográfica: mayúsculas, redonda, negrita, cursiva.

El S.G.D. proporciona la representación icónica —terminogramas— de la estructura de los descriptores.

### 3.2.3. *Mantenimiento del «Thesaurus» de descriptores*

Los léxicos documentales han de permanecer abiertos a la actualización si se desea que cumplan la función para la cual fueron concebidos. Así, dado que la comunicación social se enriquece constantemente con nueva terminología, reflejada por la documentación seleccionada y analizada para conformar la base de datos, el mantenimiento o puesta al día del «Thesaurus» es una actividad ineludible para que su eficacia sea óptima y su manejo —y la consiguiente recuperación de la información— no periclite con el paso del tiempo.

La actualización del «Thesaurus» requiere efectuar un seguimiento de los errores y lagunas que hayan podido cometerse en su construcción, y controlar la evolución de las materias que abarca.

El mantenimiento se realizará en dos etapas:

Primera: Seguimiento de uso, que se desglosa a su vez en tres dimensiones:

a) Estimar la frecuencia de uso de los descriptores utilizados en la indización: Número de documentos de la base de datos indizados por medio de ellos.
b) Comprobar la posible ausencia de conceptos dentro del «Thesaurus» por lagunas en la construcción o ampliación de servicios.
c) Subsanar las dificultades surgidas en la utilización del «Thesaurus», y que han podido ocasionar la inclusión de descriptores difusos, duplicados o de acepción poco clara; relaciones semánticas ausentes o innecesarias; y divergencias entre las diversas presentaciones del «Thesaurus».

Los indizadores registrarán estas dificultades a medida que van surgiendo, mediante el campo reservado en el registro para la descripción característica de los documentos.

Segunda: Puesta al día. Periódicamente —transcurridos seis meses— se centralizarán y editarán todos los datos registrados en el seguimiento de uso del «Thesaurus», tomándose entonces decisiones sobre el mantenimiento, disgregación, supresión, fusión de descriptores y no descriptores; incorporación o cancelamiento de relaciones jerárquicas y/o asociativas; adición, supresión o revisión de notas explicativas, etc.

### 3.2.4. *Evaluación del «Thesaurus»*

Desde la óptica de su utilización, todos los Thesauri deben valorarse en función del tipo de sistema documental del que forman parte, de su capacidad para indizar los documentos previstos y de satisfacer al usuario

La validación del sistema documental desde la perspectiva del lenguaje documental empleado se desdobla en dos planos:

Primero: Macroevaluación. Que nos la proporciona la tasa de «llamada» y la tasa de «precisión». La primera mide la calidad de la indización de

documentos y consultas, y se sitúa entre 0,6 y 0,8. Y la segunda advierte la calidad de la relación sistema-documentalista —entre 0,2 y 0,8— y sistema-usuario, un poco más reducida que la anterior.

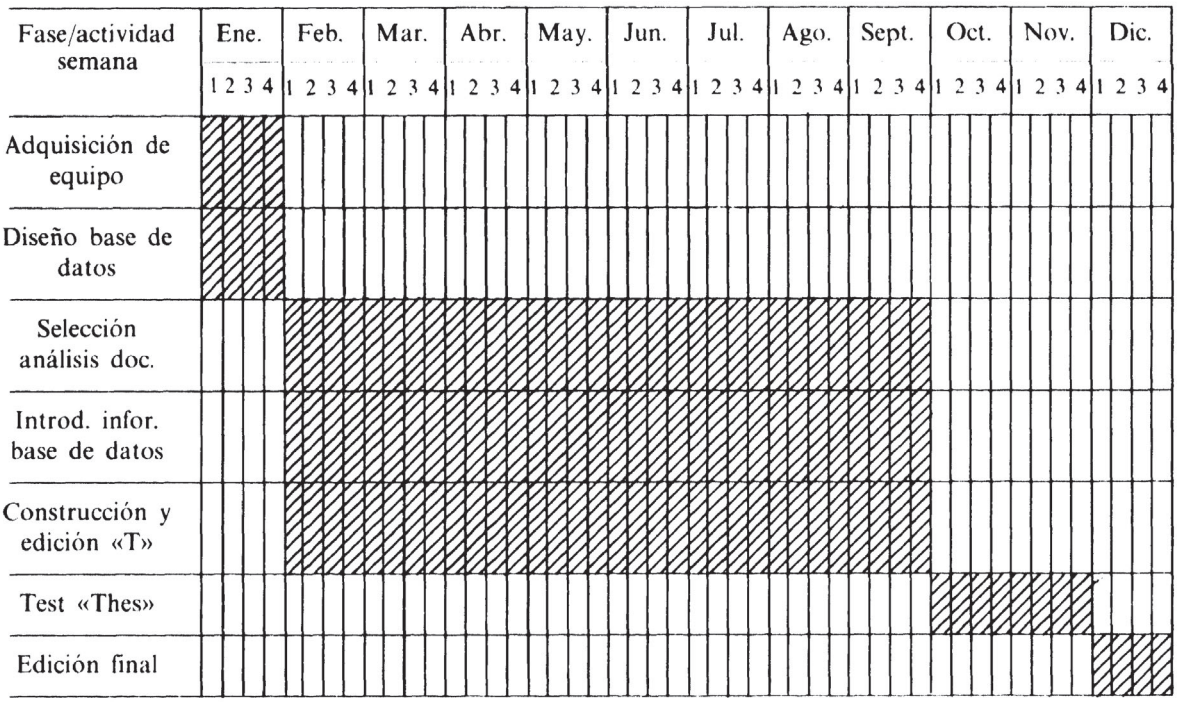
Segundo: Microevaluación. Su objetivo es identificar las causas de la disfuncionalidad, mediante el análisis de una muestra de varios cientos de búsquedas documentales. Permite efectuar un diagnóstico sobre los motivos de las respuestas inadecuadas: indización insatisfactoria, formulación deficiente de las consultas. Asimismo hace posible el mantenimiento de la coherencia de la indización, por medio del ejercicio de la indización en grupo, procedimiento que obviará necesidades de actualización en la formación de los analistas. Finalmente, el análisis de las disfuncionalidades asegurará el adecuado mantenimiento del «Thesaurus».

### 3.3. Desglose de tareas

La parte final del plan de trabajo necesario para la construcción de este sistema documental ha sido ya realizada y expuesta en páginas anteriores —estructura organizativa y plan del proyecto; requerimientos; sistema a desarrollar; selección del equipo informático. Es decir, hemos efectuado un estudio global de los objetivos a alcanzar y de los métodos a seguir, así como de los medios humanos y materiales que estimamos precisos para llevar adelante el proyecto. Por lo tanto, sólo nos queda exponer el desglose de tareas y el tiempo que dedicaremos a ellas.

A grandes rasgos, la actividad a desarrollar en el plazo de un año, independientemente de que se comience en un mes u otro, aparece estructurada en cuatro fases interrelacionadas, como podemos apreciar en el diagrama 1.

**Diagrama 1**



La primera comprende la adquisición del equipo y el diseño de la base de datos, elementos imprescindibles para introducir la información conforme se va elaborando.

La segunda dura ocho meses, e incluye los trabajos de selección, análisis de la documentación, introducción de la información en la base de datos y la construcción y edición 0 del «Thesaurus».

La tercera —dos meses— está centrada en el test o validación del «Thesaurus»; y la cuarta y última en su edición.

Pero si éste es el programa de trabajo a largo plazo, a corto plazo, el diseño del sistema comprende una actividad cotidiana que se resume del siguiente modo:

**Diagrama 2**

Investigadores

| Horas diarias | Lunes | Martes | Miércoles | Jueves | Viernes | Horas semanales |
|---|---|---|---|---|---|---|
| 1 h. | Selección noticias y elaboración «dossiers» sobre temas | | | | | |
| 2,5 h. | | | | Selección descriptores y construcción «Thesaurus» | | 10 h. |

Los investigadores —diagrama 2— seleccionarán diariamente durante una hora las noticias y elaborarán «dossiers» con los temas que interesa tratar. Aparte de ello, jueves y viernes procederán a seleccionar los predescriptores y a construir el «Thesaurus», dedicando a esta labor cinco horas semanales —repartidas entre el jueves y el viernes.

Por su parte, los documentalistas centrarán su tarea en el análisis documental, al cual asignamos dos horas diarias. La introducción de la información en la base de datos ocupará una hora diaria para evitar acumulación de documentación por automatizar. En cualquier caso, esta actividad será mecánica, porque los analistas, previamente, en fichas iguales a los registros que aparecen en la pantalla del monitor, han rellenado los distintos campos con la información pertinente.

**Diagrama 3**

Documentalistas

| Horas diarias | Lunes | Martes | Miércoles | Jueves | Viernes | Horas semanales |
|---|---|---|---|---|---|---|
| 2 h. | Análisis documental | | | | | |
| 1 h. | Introducción información base de datos | | | | | 15 h. |

Ultimados los trabajos de construcción de la base de datos y su correspondiente «Thesaurus» se procederá a la implantación del sistema, todo lo cual entrañará la formación del personal encargado de manejarlo, y de abordar las operaciones de análisis documental, de tal modo que se hallen en disposición de colaborar con el equipo de investigación en el mantenimiento del «Thesaurus».

**Agradecimientos**

Los autores agradecen a los Dres. Dña. Mercedes Caridad Sebastián y D. Félix Sagredo Fernández, las sugerencias que hicieron a la primera versión de este trabajo.

**Bibliografía**

1. BERMEJO, C. A. et al. Desarrollo de lenguajes documentales formalizados en lengua española. 2. Evaluación de los tesauros en lengua española, *Revista Española de Documentación Científica*, 12-3 (1989) 283-305.
2. CARIDAD, M. Curso «El acceso automático de la información». En *El acceso automático a la información*, Granada, 1990.
3. ESPINOSA, B.; IZQUIERDO, J. M.; SAGREDO, F. Automatización y tecnologías ópticas en información y documentación, *Cuadernos EUBD Complutense*, 1-1 (1991) 5-100.
4. OLAIZAOLA, J.; OLAIZAOLA, J. Archivo iconográfico ARGAZKI. En *Terceras Jornadas Españolas de Documentación Automatizada*, 1 ed., vol. 2, Secretariat de Publicacions i Intercanvi Cientific de la UIB, Palma de Mallorca, 1990, 1004-1018.
5. CARIDAD, M. Estructura del Banco de Datos del New York Times, *Documentación de las CC de la Información*, 4 (1980) 139-155.
6. BERTRAND, R. Les logiciels documentaires pour microordinateurs, *Documentaliste*, 26-6 (1988) 248-254.
7. GARCIA GUTIERREZ, A. L.; LUCAS, R. *Documentación Automatizada en los medios informativos*, Paraninfo, Madrid, 1987.
8. SLYPE, G. van. Seminario «Indización automática». En *El acceso automático a la información*, Granada, 1990.
9. CODINA, L. Bases de datos documentales para microordenadores. En *Terceras Jornadas Españolas...*, op. cit., 618-627.
10. FROCHOT, D. Les logiciels documentaires, *Documentaliste*, 25-6 (1988), 255A-257C.
11. RUIZ, A. A.; DEL ALAMO, I. Los archivos y las bases de datos documentales con almacenamiento de imágenes. En *Terceras Jornadas Españolas...*, op. cit., 1202-1213.
12. IZQUIERDO ARROYO, J. M. Cuatro trabajos en curso, *Documentación de las CC de la Información* (1992. En prensa).
13. COLLE, R. *Tecnologías de la información*, Pontificia Universidad Católica de Chile, Santiago de Chile, 1988.
14. CAFFO, R.; PROSSOMARITI, M. *Indicizzacione, 1975-1987: Bibliografía*. Associazione Italiana Biblioteche, Roma, 1989.
15. AITCHISON, J.; ALLEN, G. G. *Bibliografía de vocabularios. «Thesauri», encabezamientos de materias y esquemas de clasificación de ciencias sociales (mono y plurilingües)*. UNESCO, París, 1983.
16. AITCHISON, J.; GILCHRIST, A. *Thesaurus construction: A practical manual*, 2.ª ed., Aslib., Londres, 1987.
17. BARA, L. Problems of modelling and terminological control in documentary information retrieval systems: a re-examination of the «Thesaurus» concept, *Probleme de Informare si Documentare*, 23-3 (july-sept. 1989), 98-105.

18. BATTY, D. Thesaurus Construction and Maintenance: A Survival Kit, *Database*, 12-1 (feb. 1989), 13-20.
19. CURRAS, E. *Tesauros. Lenguajes terminológicos*, Paraninfo, Madrid, 1991.
20. IRAZAZABAL, A. de, et al. Aplicación de la cadena elaboración-desarrollo de tesauros en un microordenador a la confección del tesauro de la programación del CSIC (1987-1988). En *Terceras Jornadas Españolas...*, op. cit., 876-889.
21. JANIK, S.; BRUNET, L. La mise a jour d'un thesaurus, *Documentaliste*, 24-6 (nov.-dec. 1987), 215-229.
22. LAGUNA SERRANO, E., et al. Confección automática de tesauros, *Revista Española de Documentación Científica*, 12-2 (abril-junio 1989), 129-140.
23. ROHOU, C. La gestion automatisée des thesaurus: étude comparative de logiciels, *Documentaliste*, 24-3 (may-june 1987), 103-108.