GENERACION DE FICHEROS DE INPUT PARA LA CONSTRUCCION Y EL MANTENIMIENTO DE TESAUROS EN BRS MEDIANTE UN SISTEMA DE GESTION DE BASES DE DATOS EN MICROORDENADOR

C. B. Amat* y A. Baquedano Alcocer**

Resumen: Se presenta un procedimiento que ha permitido la creación en lote de un tesauro en el entorno BRS. Dicho tesauros comprende 7.500 términos correspondientes a la geografía mundial y controla bases de datos bibliográficas y audiovisuales de información de actualidad. La estructuración del fichero de carga y la delimitación de las relaciones terminológicas se han realizado empleando un sistema de gestión de base de datos en microordenador. El alcance semántico de los términos se ha representado mediante códigos clasificatorios alfanuméricos que permiten identificar las relaciones de equivalencia (igualdad de códigos) y jerárquicas (inclusión de los códigos más genéricos en los de los términos específicos). El procedimiento es igualmente aplicable al establecimiento de relaciones asociativas. Los programas, detallados a lo largo del trabajo, se basan en la relación de un fichero general en que los términos se ordenan alfabéticamente y otro en que los términos preferentes se disponen en orden jerárquico según sus códigos. Dos ficheros adicionales contienen los términos no preferentes y los asociados. Como resultado se obtienen tanto las ediciones alfabética y jerárquica simples como una combinación de ambas, apta para la carga posterior del fichero en el módulo tesauro de BRS u otros grandes sistemas.

Palabras clave: BRS, creación de tesauros, mantenimiento de tesauros.

Abstract: A method is presented for batch creation and maintenance of BRS thesaurus files. The thesaurus elaborated contains 7.500 terms corresponding to world geography and is used to control input and retrieval from several bibliographic and audiovisual databases. The load file structure, as well as the delimitations of terminological relationships, have been performed using a microcomputer database management system. Semantic scope of every term has been represented by alphanumeric classificatory codes, equivalence relationship being represented by equality of notation and jerarquic relationships by inclusion of generic codes in the more specific ones. The procedure is also applicable to associative relationships. Processing is based on the connection between two main term files, one containing all terms in alphabetic arrangement and the other sorting preferred term by their codes. Two additional files contain non preferred and associated terms. As a result, not only the alphabetic and hierarchical displays, but a combination of both are obtained, suitable for a later loading of the file in BRS thesaurus module. Source programs are provided.

Keywords: BRS, thesaurus maintenance, thesaurus construction.

* Unidad de Documentación.
** Sección de Informática. Radiotelevisión Valenciana. Burjassot. Valencia.
Recibido 20-12-92.

1. Introducción

La mayoría de los actuales sistemas de recuperación de información sobre grandes instalaciones emplean un módulo de gestión de tesauro para el control del vocabulario. Al igual que en la creación y el mantenimiento de los ficheros principales, es posible la adición de elementos online o por lotes en la creación o mantenimiento de tesauros (1, 2, 3).

El sistema de gestión documental BRS dispone de un completo módulo guiado por menús para la creación y mantenimiento online de tesauros, que ayuda a seleccionar los términos introductorios *(lead terms)* y las relaciones adjudicadas a cada uno. Tras la validación de las relaciones, las recíprocas se generan automáticamente. Desgraciadamente, la tarea de transferir un tesauro ya estructurado término a término es en exceso farragosa si se realiza online. Por otra parte, raro es el caso en que el mantenimiento de un tesauro exige la adición de una o pocas relaciones. Es más habitual la trasferencia al fichero de tesauros de un gran número de relaciones en cada actualización, sobre todo en entornos de trabajo con un alto número de indizadores. Todo ello apoya la alternativa de crear o mantener los ficheros de tesauros mediante el procesamiento de lotes de términos y relaciones.

BRS, al igual que otros grandes sistemas de gestión documental, posibilita esta segunda alternativa con la única exigencia de que los ficheros de input estén convenientemente estructurados. Ello supone que se debe especificar cada relación anteponiendo a los términos de la misma los adecuados prefijos. Así, la relación «los cuerpos y fuerzas de seguridad incluyen al de la guardia civil, que es lo mismo que decir a los guardias civiles» se expresaría como

LT CUERPOS Y FUERZAS DE SEGURIDAD
NT GUARDIA CIVIL

y

LT GUARDIA CIVIL
UF GUARDIAS CIVILES

quedando el sistema encargado de la generación automática de las relaciones recíprocas

LT GUARDIAS CIVILES
USE GUARDIA CIVIL

y

LT GUARDIA CIVIL
BT CUERPOS Y FUERZAS DE SEGURIDAD

El problema, entonces, es hallar el medio de dotar al fichero de términos que se pretende incluir en lote, de la estructura conveniente para que el sistema pueda interpretar adecuadamente las relaciones.

Habitualmente, se distinguen dos tipos de fuentes en la selección de los términos y relaciones a incluir en un tesauro: las fuentes estructuradas y las fuentes abiertas (8). Las primeras son aquellas que ofrecen no sólo una colección de términos, sino que indican también las relaciones semánticas que mantienen o, al menos, los agrupan por campos semánticos. La utilización de un tesauro ajeno, para su adaptación al sistema en cuestión, es un ejemplo habitual de recurso a las

fuentes del primer tipo. Las fuentes abiertas ofrecen, en el mejor de los casos, simples indicaciones de la equivalencia semántica entre términos y quizás de alguna relación de tipo asociativo. El empleo del índice de materias de una monografía especializada, como fuente para la selección de términos en un campo determinado, es el ejemplo más claro de este segundo procedimiento.

En el caso de la utilización de un tesauro ajeno, la información sobre términos y relaciones se puede transferir a un sistema de procesamiento de textos grabando términos e indicadores de las relaciones (prefijos) de forma sencilla. La fuente se puede hallar ya en soporte magnético: puede existir en edición electrónica o bien se puede descargar a partir de un CD-ROM o (menos probable, en relación con la extensión) a partir de una base de datos accesible online. En este caso el trabajo se facilita aún más, salvando las posibles complicaciones idiomáticas.

En el segundo caso, tanto si la obra se halla o puede obtenerse en soporte magnético como si es impresa, es necesario disponer de un instrumento para establecer las relaciones terminológicas y conceptuales adecuadas.

Las unidades de documentación e informática de Radiotelevisión Valenciana, responsables de las tareas de creación y mantenimiento del sistema de gestión documental de esta empresa informativa, han puesto a punto un procedimiento para la estructuración de ficheros de input con información terminológica, para la creación y el mantenimiento de tesauros en el entorno BRS, utilizando como soporte inicial de los ficheros el sistema relacional de bases de datos para microordenador dBASE III. Fruto de este procedimiento es la creación de un tesauro de topónimos de algo más de 7.500 términos, instrumento auxiliar de singular importancia en el entorno de la información de actualidad. Ejemplos de este tesauro se mostrarán a lo largo del presente trabajo.

2. Método

En esencia, la elaboración de todo vocabulario controlado supone la organización de los términos y conceptos en tres estructuras: la estructura de sinónimos-homónimos, la estructura de equivalencia y la estructura clasificatoria (4). La primera de ellas se basa en el establecimiento de la relación de sinonimia entre términos diferentes que denotan un mismo concepto y en la distinción entre homónimos: términos de igual grafía que corresponden a diferentes conceptos. La estructura de equivalencia agrupa términos no estrictamente sinónimos, aunque considerados como tales a efectos de indización y recuperación en el contexto de cada sistema. La estructura clasificatoria, por su parte, supone la ordenación de los términos en función del mayor o menor alcance (la mayor o menor generalidad) del concepto que cada uno representa.

La norma española de elaboración de tesauros monolingües distingue, en efecto, una edición alfabética y una edición sistemática de los tesauros. La primera pone en evidencia las relaciones de sinonimia, polisemia y cuasisinonimia, mientras que la segunda presenta los términos relacionados jerárquicamente. Ambas ediciones se pueden presentar combinadas (5).

Teniendo en cuenta todo lo anterior, es preciso hallar un medio de representar: 1) los términos; 2) los conceptos que denotan, y 3) las decisiones sobre los términos.

No existe dificultad alguna en la representación de los términos mediante una estructura de base de datos. Bastará con un archivo en que cada registro contenga un campo de caracteres con la extensión adecuada. Tampoco existe problema en representar las decisiones sobre los términos: básicamente, existen términos preferentes, términos no preferentes e indicadores clasificatorios. El que un término sea preferente supone su inclusión tanto en la edición alfabética como en la sistemática del tesauro. Los términos no preferentes sólo se incluyen en la presentación alfabética y los indicadores clasificatorios pueden, en caso de que se empleen, figurar en la edición sistemática. El tipo de término (y la decisión que representa) se puede incluir también en la estructura de la base de datos en cuestión. Para representar los conceptos, lo más simple es recurrir a una solución casi centenaria: la notación, esto es, la adjudicación a los términos de códigos clasificatorios. La estructura de los registros del fichero de base de datos contendrá, así, los siguientes elementos:

Nombre	Tipo	Extensión
TERMINO	CARACTER	40
TIPO	CARACTER	1
CODIGO	CARACTER	20

La relación de sinonimia o equivalencia originará la inclusión en el fichero de dos registros, como en el caso siguiente:

PAISES BAJOS		HOLANDA
1	frente a	0
EUO030		EUO030

La relación de homonimia se resuelve añadiendo calificadores a los términos. Por ejemplo:

MISSISSIPPI (RIO)		MISSISSIPPI (ESTADO)
1	frente a	1
ANR025		ANE035

Por último, la relación jerárquica se expresa haciendo corresponder, a cada término preferente, una extensión del código clasificatorio de su término genérico:

EUROPA OCCIDENTAL		ESPAÑA
1	y	1
EUO		EUO05

La polijerarquía, es decir, la posible existencia de más de un genérico para cada término, se resuelve igualmente, generando tantas entradas como relaciones en el tesauro. Por ejemplo:

ESPAÑA	
1	
EUO05	(donde EUO representa EUROPA OCCIDENTAL)

ESPAÑA
1
PME015 (donde PME son los PAISES MEDITERRANEOS)

ESPAÑA
1
LAM035 (donde LAM corresponde a LATINOAMERICA), etc.

Las relaciones básicas en el tesauro quedarían, pues, definidas en el fichero de base de datos por las siguientes reglas:

1) Dos términos son sinónimos si su código clasificatorio es el mismo.
2) Un término es genérico de otro si el código clasificatorio del primero está incluido en el código clasificatorio del segundo.
3) Un término es preferente cuando es del tipo 1. Este término se incluirá en las ediciones alfabética y sistemática del tesauro. En la edición alfabética, además, deberá acompañarse de sus posibles sinónimos.
4) Un término no preferente es del tipo 0 y se deberá incluir en la edición alfabética del tesauro con un reenvío a su preferente (esto es, deberá haber siempre un término del tipo 1 cuyo código clasificatorio coincida con el término tipo 0).

La selección de los términos del tesauro geográfico de Radiotelevisió Valenciana se basó en la consulta de obras de tipo enciclopédico y documentación oficial de carácter normativo. Aunque las obras no eran de naturaleza totalmente estructurada, ofrecían una agrupación razonable de los términos. El diseño del fichero de base de datos incluyó los tres campos mencionados anteriormente: término, código y tipo. Los términos fueron grabados en soporte magnético aprovechando dicha estructura y distribuidos en 5 grupos, cada uno de ellos correspondiente a los países de cada continente. Tras cada término de un país, se consignaron los correspondientes a la distribución por sus estados, territorios, regiones, departamentos y ciudades. A continuación, se determinó la sintaxis de la notación, que quedó configurada como un código alfanumérico con tres caracteres alfabéticos iniciales y grupos de tres cifras separadas por un punto para facilitar su lectura. De forma similar, se procedió al registro de los términos (comunidades, provincias y ciudades) de la geografía española.

Los ficheros se dividieron en grupos de registros que representaban topónimos de cada continente y se utilizó la función de sustitución (REPLACE en la sintaxis de dBASE III) para adjudicar a cada continente y a los países, estados, regiones, territorios y ciudades, los tres caracteres alfabéticos iniciales del código clasificatorio. Posteriormente, los códigos se fueron extendiendo mediante grupos de caracteres numéricos hasta alcanzar la adecuada sintaxis. Así, del grupo inicial

AMERICA DEL NORTE
ESTADOS UNIDOS DE AMERICA
NUEVA YORK (ESTADO)
ALBANY
NUEVA YORK (POBLACION)

se pasó a

AMERICA DEL NORTE	AMN
ESTADOS UNIDOS DE AMERICA	AMN002
NUEVA YORK	AMN002.075
ALBANY	AMN002.075.010
NUEVA YORK (POB)	AMN002.075.020

Una vez adjudicado el código correspondiente a España, el fichero que contenía los topónimos correspondientes a las comunidades autónomas, las provincias y las ciudades se trató del mismo modo. Se generó un tercer fichero para los topónimos correspondientes a la Comunidad Valenciana y se procedió del mismo modo con los términos agrupados en las tres provincias.

Se combinaron los tres ficheros para obtener una primera versión del conjunto de términos y, una vez constituido el fichero unificado, que se denominó FTP.DBF (Fichero de Términos Preferentes), se procedió a su ordenación alfabética, generando el fichero índice auxiliar basado en el campo TERMINO.

3. Eliminación de la homonimia

El fichero ordenado alfabéticamente presentaba juntos los términos homónimos. Para resolver la homonimia se procedió a añadir a los términos calificadores basados en su ubicación geográfica o en su significado o ambas cosas. Así, se distinguieron, por ejemplo:

VALENCIA (VENEZUELA)
de VALENCIA

NUEVA YORK
de
NUEVA YORK (POB) y otros casos

4. Generación del fichero de términos no preferentes

Aunque, como se ha apuntado anteriormente, tanto los términos preferentes como los no preferentes, se incluirán al final del proceso en un mismo fichero, los términos no preferentes requieren un tratamiento inicial diferenciado en un segundo fichero, denominado FNP.DBF (Fichero de términos No Preferentes). La estructura de este segundo fichero es idéntica a la del FTP, y su manejo es el siguiente: una vez localizado un término preferente que dispone de sinónimo, se incluye en el FTP. A continuación, se incluye en el FNP el o los sinónimos localizados con el mismo código clasificatorio y tipo 0. Por ejemplo:

LA VALL DE TAVERNES	1	EuO060.240.030.275.095
es sinónimo de		
TAVERNES DE VALLDIGNA	0	EuO060.240.030.275.095

y

ESTADOS UNIDOS DE AMERICA	1	AmN020

es sinónimo de
ESTADOS UNIDOS 0 AmN020
y de
EUA 0 AmN020

El proceso puede parecer farragoso, pero se facilita si los dos ficheros de base de datos se encuentran abiertos en diferentes áreas de trabajo. Por otra parte, es preciso reconocer que un campo semántico como el tratado no presenta un excesivo número de sinónimos.

5. Definición final de los ficheros

A partir del fichero principal, FTP, se ha de generar un segundo fichero que es copia exacta del mismo: FTP2. Por otra parte, los términos no preferentes incluidos en FNP se añaden a FTP. Serán distinguibles en razón de su tipo 0. El fichero imagen FTP2 de términos preferentes se indiza según el campo CODIGO.

6. Elaboración de la edición alfabética del tesauro

La figura 1 contiene el programa que elabora la edición alfabética del tesauro. Aunque se pueden apreciar las remisiones a los programas adicionales de elaboración de la edición combinada (ver más adelante «Combinación de las ediciones alfabética y jerárquica»), el esquema básico se describe en esta sección.

Figura 1

Programa principal de procesamiento de los términos

```
CLEA ALL
SET TALK OFF
PUBLIC MCODG,MCOD2,MNIV1,MNIV2,MPOSI
* ASIGNACION DE LOS FICHEROS A LAS DIFERENTES AREAS DE
* TRABAJO
SELE 3
USE FTP2 INDE XFTP2
SELE 2
USE FNP INDE XFNP
SELE 1
USE FTP INDE XFTP
GO TOP
* ALMACENAMIENTO DE LAS VARIABLES CORRESPONDIENTES AL
* PRIMER TERMINO DEL FICHERO PPRINCIPAL EN ORDEN ALFABETICO
* LA ORTOGRAFIA DEL TERMINO:
MTERM = TERM
* LA CADENA DE CARACTERES DE LA NOTACION JERARQUICA:
MCODG = CODG
* Y EL NUMERO DE CARACTERES DE SU NOTACION (GRADO DE
* ESPECIFICIDAD)
MNIV1 = LEN(TRIM(CODG))
```

```
* Y SU POSICION (NUMERO DE REGISTRO) EN EL FICHERO
* PRINCIPAL
MPOSI = RECNO()
*
? "LT "+TERM
*
* SI EL TERMINO ES PREFERENTE SE INVESTIGAN SUS SINONIMOS A
* TRAVES DE LA IGUALDAD DE CODIGOS EN EL FICHERO DE
* TERMINOS NO PREFERENTES
*
IF TIPU = "1"
SET RELA TO CODG INTO FNP
SELE 2
DO WHILE .NOT. EOF() .AND. CODG = MCODG
? SPAC(3)+"UF "+TERM
SKIP
ENDDO
SELE 1
GO MPOSI
*
* Y TAMBIEN SUS ESPECIFICOS Y GENERICOS A TRAVES DE LA
* ACTIVACION DE LOS PROGRAMAS AUXILIARES
DO REDC22
DO REDC20
*
* EN CASO CONTRARIO, SE LOCALIZA SU PREFERENTE A TRAVES DE
* LA IGUALDAD DE CODIGOS EN EL FICHERO DE PREFERENTES
* ORDENADOS JERARQUICAMENTE
*
ELSE
SET RELA TO CODG INTO FTP2
SELE 3
?SPAC(3)+"USE "+TERM
ENDIF
SELE 1
GO MPOSI
SKIP
*
* SE INICIA AHORA EL PROCESAMIENTO DE TODOS LOS TERMINOS
* DEL FICHERO PRINCIPAL COMENZANDO POR EL SEGUNDO
*
DO WHILE .NOT. EOF()
*
* SE PRODUCE LA ACTUALIZACION DE LAS VARIABLES, EXCEPTO LA
* CORRESPONDIENTE A LA ORTOGRAFIA DEL TERMINO
*
MCODG = CODG
MNIV1 = LEN(TRIM(CODG))
MPOSI = RECNO()
*
* SE PLANTEA AHORA LA CONDICION DE QUE EL TERMINO APAREZCA
* POR SEGUNDA VEZ EN EL FICHERO PRINCIPAL
*
IF TERM = MTERM
```

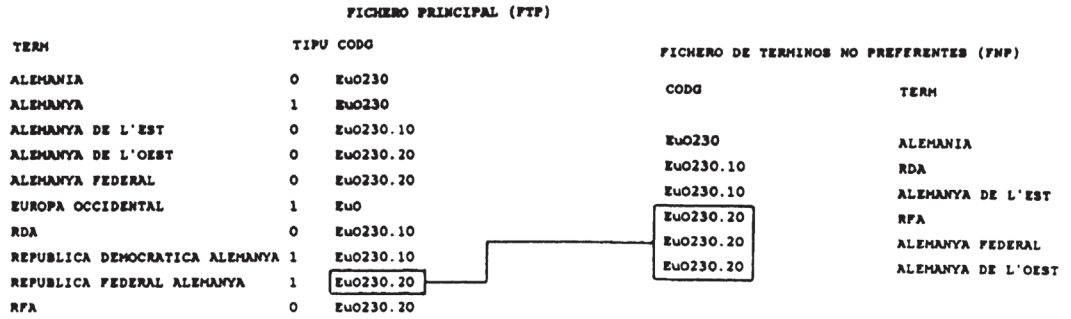
```
*
* EN CASO DE QUE HAYA COINCIDENCIA, SE ACTIVA EL PROGRAMA
* DE IDENTIFICACION DE GENERICOS
*
DO REDC20
SELE 1
GO MPOSI
*
* UNA VEZ IDENTIFICADOS, EL PROCESAMIENTO SIGUE EN EL
* TERMINO SIGUIENTE
*
SKIP
LOOP
ENDIF
? "LT "+TERM
*
* EN CASO DE QUE EL TERMINO SEA DIFERENTE, EL PROCESAMIENTO
* CUBRE TANTO LA IDENTIFICACION DE SINONIMOS COMO LA
* ADJUDICACION DE GENERICOS Y ESPECIFICOS SI ES PREFERENTE
*
IF TIPU = "1"
SET RELA TO CODG INTO FNP
SELE 2
DO WHILE .NOT. EOF() .AND. CODG = MCODG
? SPAC(3)+"UF "+TERM
SKIP
ENDDO
SELE 1
GO MPOSI
DO REDC22
SELE 1
GO MPOSI
DO REDC20
SELE 1
GO MPOSI
ELSE
*
* O LA IDENTIFICACION DEL PREFERENTE SI NO LO ES
*
SET RELA TO CODG INTO FTP2
SELE 3
? SPAC(3)+"USE "+TERM
ENDIF
SELE 1
* LA VARIABLE DE LA MORFOLOGIA DEL TERMINO SE ACTUALIZA
*
MTERM= TERM
SKIP
*
* Y EL PROCESAMIENTO CONTINUA CON EL SIGUIENTE TERMINO
*
ENDDO
RETU→
```

El FTP2, de términos preferentes indizados según el código, se sitúa en la tercera área de trabajo. El FNP, de términos no preferentes, ordenados también por su código, se sitúa en la segunda. En la primera área de trabajo, que será el área activa al inicio del procesamiento, se abre el FTP, fichero de términos preferentes y sinónimos ordenados alfabéticamente.

El primer término en el FTP es también el primero de la ordenación alfabética. Con éste, como con los restantes, la primera operación consiste en determinar si es preferente o no, es decir, si su tipo es 1 ó 0. En el caso de que sea un término preferente, se establece una relación a través de su código con el FNP. Si no hay ningún término en este segundo fichero cuyo código coincida con el del término preferente, el término no tiene equivalentes y el procesamiento continúa con el siguiente término en el orden alfabético. Si, por el contrario, existe un término no preferente cuyo código coincide con el de FTP, se anota como sinónimo de éste y se sigue examinando el fichero hasta que no haya ningún término en el FNP cuyo código resulte coincidente con el del preferente. La figura 2 esquematiza la relación entre ficheros para la identificación de términos equivalentes de un preferente.

Figura 2

Relación entre ficheros para la identificación de los sinónimos de un término preferente



En el caso de que el término sea de tipo 0, es decir, sea un término equivalente de otro preferente, la relación se establece entre el FTP y el FTP2. Es evidente que debe haber en este segundo fichero un término cuyo código coincida con el del término no preferente del FTP. En este caso, el término localizado a través de la coincidencia de códigos será el preferente del primero. En la figura 3 se esquematiza la relación para la identificación del preferente de un término equivalente. El hecho de que se realice el procesamiento del primer término de la ordenación alfabética y de que se almacene su ortografía (MTERM = TERM), antes de iniciar el procesamiento del resto de los términos del fichero principal con el segundo término, obedece a la posible existencia en el fichero de más de una aparición de un término, lo que sucede siempre que existe polijerarquía. La activación de los programas de adjudicación de genéricos y específicos (REDC20 y REDC22) se detalla más adelante (elaboración de la edición jerárquica del tesauro).

Una vez realizado todo el procesamiento de los términos ordenados alfabéticamente en FTP, el resultado es similar al que se muestra en la figura 4: una edición alfabética con entradas correspondientes a cada término y sus relaciones de sinonimia y equivalencia.

Figura 3

Relación entre ficheros para la identificación del término preferente correspondiente a un sinónimo

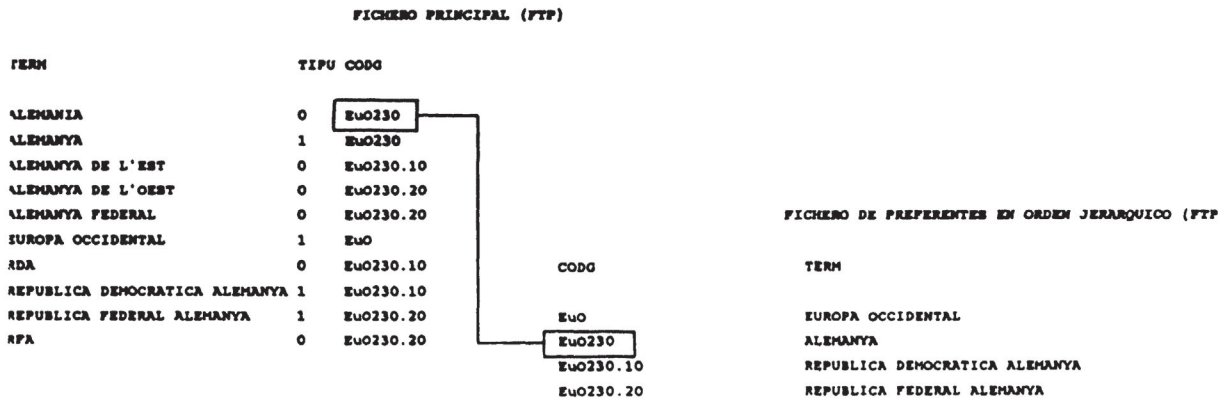


Figura 4

Edición alfabética con especificación de las relaciones de sinonimia (USE y UF)

```
ALEMANIA
   USE ALEMANYA

ALEMANYA
   UF ALEMANIA

ALEMANYA DE L'EST
   USE REPUBLICA DEMOCRATICA ALEMANYA

ALEMANYA DE L'OEST
   USE REPUBLICA FEDERAL ALEMANYA

ALEMANYA FEDERAL
   USE REPUBLICA FEDERAL ALEMANYA

EUROPA OCCIDENTAL

RDA
   USE REPUBLICA DEMOCRATICA ALEMANYA

REPUBLICA DEMOCRATICA ALEMANYA
   UF RDA
   UF ALEMANYA DE L'EST

REPUBLICA FEDERAL ALEMANYA
   UF RFA
   UF ALEMANYA FEDERAL
   UF ALEMANYA DE L'OEST

RFA
   USE REPUBLICA FEDERAL ALEMANYA
```

7. Elaboración de la edición jerárquica del tesauro

Los términos se ordenan jerárquicamente tomando como valor el alcance del concepto o conceptos que representan. Puesto que el código clasificatorio empleado en el campo CODIGO expresa dicho alcance y puesto que el FTP2 contiene los términos indizados según dicho valor, la edición jerárquica se obtiene con gran sencillez listando los términos preferentes en el orden marcado por sus códigos. La figura 5 muestra un fragmento de la edición jerárquica así obtenida. La inclusión de los códigos clasificatorios es opcional y puede obtenerse una lista únicamente de los términos con un sangrado similar haciendo preceder cada término de la cantidad de espacios en blanco equivalente a la extensión ajustada de su código clasificatorio.

Figura 5

Un fragmento de la edición jerárquica con inclusión de los códigos clasificatorios (opcional)

```
EuO       EUROPA OCCIDENTAL
EuO230      ALEMANYA
EuO230.10       REPUBLICA DEMOCRATICA ALEMANYA
EuO230.20       REPUBLICA FEDERAL ALEMANYA
```

8. Combinación de las ediciones alfabética y jerárquica

El fichero de input para la carga de un tesauro BRS en lote presenta ciertas exigencias de formato: todos los términos preferentes deben ir precedidos de una etiqueta (LT, por *lead term*) y, a continuación, se deben listar todos los términos relacionados con el LT con las correspondientes abreviaturas de relación. Ello plantea la exigencia de combinar, para cada término preferente, las relaciones de sinonimia, equivalencia y jerarquía en conjunto. Por otra parte, se ha comprobado que si se ofrecen en el fichero de input las relaciones directas (USE, BT) junto con las recíprocas (UF, NT), la cantidad de tiempo necesario de procesamiento en el proceso de verificación de la carga se reduce extraordinariamente. Así, en la experiencia de Radiotelevisió Valenciana, una primera carga sin expresión de las relaciones recíprocas consumió alrededor de 100 horas de proceso. Cuando la carga del mismo conjunto de términos se realizó aportando las relaciones recíprocas, el tiempo de procedimiento se redujo ¡a 10 minutos! A continuación se esquematiza el proceso de estructuración del fichero terminológico de acuerdo con estas exigencias.

El punto de partida es, nuevamente, el FTP. Por supuesto que no hay cambios en el caso de que el término sea no preferente (TIPO = 0), puesto que su única relación posible (y única) es la de USE + el preferente correspondiente. En el caso de los términos preferentes, es necesario aportar, además de las eventuales relaciones UF, las relaciones jerárquicas tanto ascendentes (BTs) como descendentes (NTs). Para ello, se establece la relación entre FTP y FTP2 basada en los códigos clasificatorios. Recuérdese que, según la segunda de las reglas enunciadas al principio del apartado metodológico, un término es genérico de otro sólo si su código clasificatorio está incluido en el código clasificatorio del segundo. Por tanto, dado

un código correspondiente al término del FTP que se está procesando, y establecida la relación con el FTP2, se recorre el sentido ascendente este segundo fichero hasta hallar un código contenido en el de referencia. En el momento en que se halla, el término de FTP2 correspondiente es anotado como genérico del término de referencia. Como en una imagen especular, el procesamiento se devuelve al punto de FTP2 donde se situaba el código originario y, desde allí, desciende hacia el final del fichero en busca de términos cuyos códigos contengan el código de referencia. Todos aquellos que cumplan la condición corresponderán a términos específicos del preferente en cuestión. La figura 6 muestra los programas que, activados desde el programa principal de la figura 1, identifican las relaciones jerárquicas entre términos preferentes. Por su parte, las relaciones entre los ficheros se ilustran en la figura 7.

Figura 6

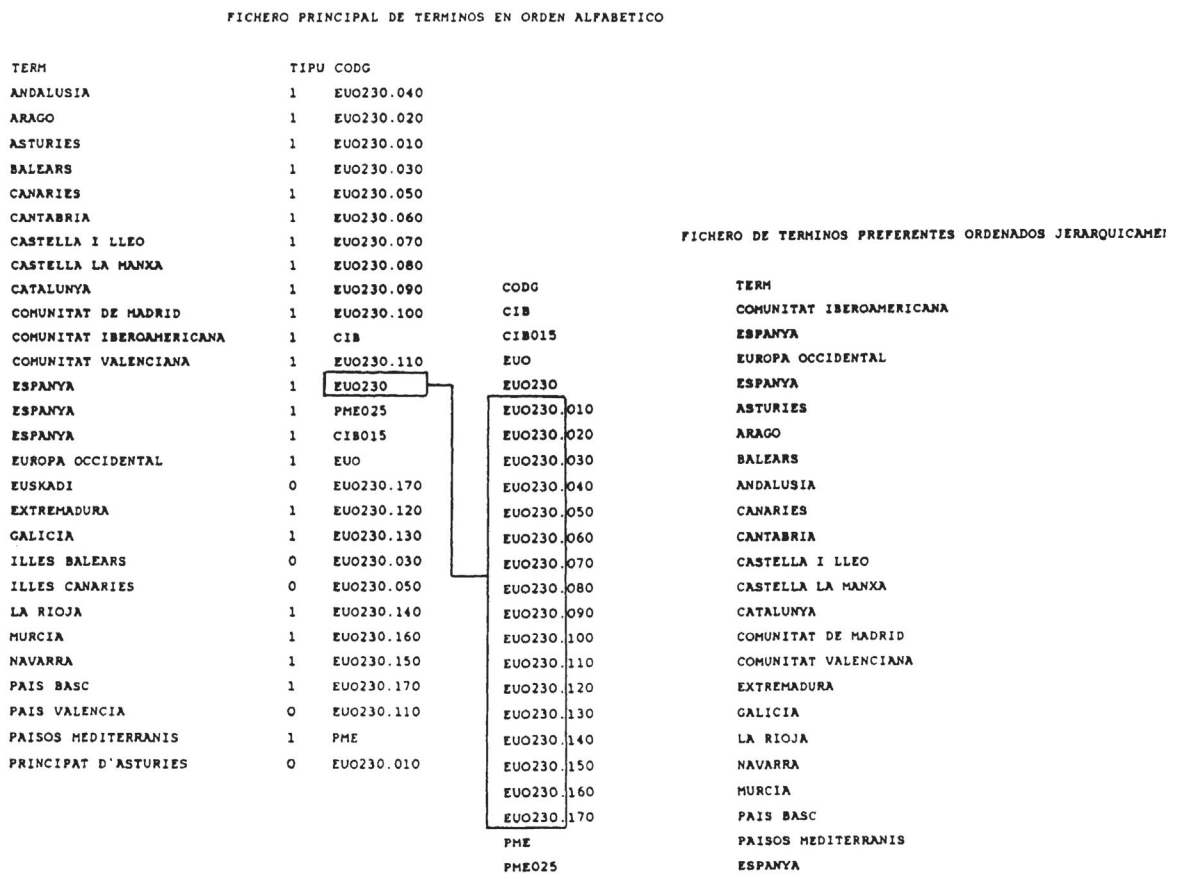
Programas activados desde el principal para la determinación de las relaciones genérico-específicas

```
* IDENTIFICACION DE GENERICOS DE UN TERMINO PREFERENTE

SET RELA TO CODG INTO FTP2
SELE 3
SKIP -1
MCOD2 = TRIM(CODG)
*
* EL FICHERO DE TERMINOS PREFERENTES SE RECORRE EN ESTE
* CASO EN ORDEN ASCENDENTE A PARTIR DEL TERMINO PRINCIPAL
*
DO WHILE .NOT. BOF()
IF AT(MCOD2,MCODG) = 1
? SPAC (3)+"BT "+TRIM(FTP2->TERM)
EXIT
ELSE
SKIP -1
MCOD2 = TRIM(CODG)
ENDI
ENDDO
RETU→
*IDENTIFICACION DE ESPECIFICOS DE UN TERMINO PREFERENTE

SET RELA TO CODG INTO FTP2
SELE 3
MCODG = TRIM(CODG)
MNIV1 = LEN(TRIM(CODG))
*
* EL CODIGO CLASIFICATORIO Y EL NIVEL DE ESPECIFICIDAD DEL
* TERMINO JERARQUICAMENTE POSTERIOR AL QUE SE PROCESA SE
* ALMACENAN EN SENDAS VARIABLES
*
SKIP 1
MCOD2 = TRIM(CODG)
MNIV2 = LEN(TRIM(CODG))
*
* EL FICHERO DE PREFERENTES SE RECORRE EN ORDEN DESCENDENTE
*
DO WHILE .NOT. EOF()
*
```

```
* SI EL CODIGO DEL TERMINO PRINCIPAL ESTA CONTENIDO EN EL
* DEL TERMINO SIGUIENTE
*
IF AT(MCODG,MCOD2) = 1
*
* Y SI SU NIVEL JERARQUICO ES INMEDIATAMENTE INFERIOR
*
IF MNIV2 = MNIV1+3 .OR. MNIV2 = MNIV1+4
*
* SE SELECCIONA COMO ESPECIFICO
*
? SPAC(3)+"NT "+TRIM(FTP2->TERM)
ENDIF
SKIP
MCOD2 = TRIM(CODG)
MNIV2 = LEN(TRIM(CODG))
ELSE
EXIT
ENDI
ENDDO
SELE 1
RETU→
```

Figura 7

Relación entre ficheros para la adjudicación de genéricos y específicos de un término preferente



Puesto que las diversas apariciones de un término que presenta polijerarquía (es decir, que cuenta con más de un genérico) aparecen agrupadas en el fichero FTP, ordenado alfabéticamente, bastará con comparar, al inicio de cada bucle, si el término cuyo procesamiento se inicia coincide con el anterior. En este caso, la búsqueda se reduce al genérico correspondiente a cada caso.

El fichero resultante (un fichero ASCII generado mediante la orden SET ALTERNATE TO, admisible por cualquier programa de procesamiento de textos u otros) se muestra fragmentariamente en la figura 8.

Figura 8

Un fragmento del fichero de carga resultante del procesamiento. Por simplificar, se han omitido los específicos de los términos correspondientes a las autonomías y otros

```
LT ANDALUSIA
   BT ESPANYA
LT ARAGO
   BT ESPANYA
LT ASTURIES
   UF PRINCIPAT D'ASTURIES
   BT ESPANYA
LT BALEARS
   UF ILLES BALEARS
   BT ESPANYA
LT CANARIES
   UF ILLES CANARIES
   BT ESPANYA
LT CANTABRIA
   BT ESPANYA
LT CASTELLA I LLEO
   BT ESPANYA
LT CASTELLA LA MANXA
   BT ESPANYA
LT CATALUNYA
   BT ESPANYA
LT COMUNITAT DE MADRID
   BT ESPANYA
LT COMUNITAT IBEROAMERICANA
   NT ESPANYA
LT COMUNITAT VALENCIANA
   UF PAIS VALENCIA
   BT ESPANYA
LT ESPANYA
   NT ASTURIES
   NT ARAGO
   NT BALEARS
   NT ANDALUSIA
   NT CANARIES
   NT CANTABRIA
   NT CASTELLA I LLEO
   NT CASTELLA LA MANXA
   NT CATALUNYA
   NT COMUNITAT DE MADRID
   NT COMUNITAT VALENCIANA
   NT EXTREMADURA
   NT GALICIA
   NT LA RIOJA
   NT NAVARRA
   NT MURCIA
   NT PAIS BASC
   BT EUROPA OCCIDENTAL
   BT PAISOS MEDITERRANIS
   BT COMUNITAT IBEROAMERICANA
LT EUROPA OCCIDENTAL
   NT ESPANYA
LT EUSKADI
   USE PAIS BASC
LT EXTREMADURA
   BT ESPANYA
LT GALICIA
   BT ESPANYA
LT ILLES BALEARS
   USE BALEARS
LT ILLES CANARIES
   USE CANARIES
LT LA RIOJA
   BT ESPANYA
LT MURCIA
   BT ESPANYA
LT NAVARRA
   BT ESPANYA
LT PAIS BASC
   UF EUSKADI
   BT ESPANYA
LT PAIS VALENCIA
   USE COMUNITAT VALENCIANA
LT PAISOS MEDITERRANIS
   NT ESPANYA
LT PRINCIPAT D'ASTURIES
   USE ASTURIES
```

9. Discusión

Frente a los programas destinados específicamente a la elaboración y mantenimiento de tesauros en microordenador, revisados por Jessica Milstead (3), la utilización de software de propósito general presenta ventajas. Entre ellas destaca la popularidad de programas como las diferentes versiones de DataBase o FoxBase, que se han convertido en «standards» en la práctica totalidad de entornos de trabajo. Su flexibilidad es una ventaja añadida. Así, por ejemplo, se han descrito procedimientos de descarga de datos bibliográficos e importación a dBASE III+ que permitirían la rápida recopilación de un vocabulario técnico (6).

Por otra parte, la utilización de este tipo de programas en la construcción de vocabularios controlados no es nueva. David Batty ha propuesto la utilización de dBASE II para el diseño y mantenimiento de pequeños (hasta 500 términos) tesauros (7). Más recientemente, el propio Batty menciona el sistema de gestión de tesauros del Instituto Centroamericano de Investigación y Tecnología Industrial, basado en el mismo programa (8). En ninguno de los dos casos, sin embargo, se ofrecen los programas ni la estructura de ficheros empleados. Además, debe tenerse en cuenta que la versión II de dBASE no es capaz de establecer relaciones entre tantos ficheros como los requeridos en el procedimiento aquí descrito.

El presente trabajo, por ende, no propone la construcción de tesauros y su utilización en el entorno de los sistemas de gestión de bases de datos para microordenador, sino la mera utilización de estos sistemas para la estructuración de ficheros de carga en lote de los tesauros asociados a la operación de grandes sistemas, como BRS o BASIS. Desde este punto de vista, el procedimiento propuesto presenta dos puntos débiles: en primer lugar, la adjudicación de los códigos clasificatorios para representar el contenido semántico de cada término y la determinación de relaciones asociativas.

En relación con el empleo de notación, el manual de Aitchison y Gilchrist, por ejemplo, no ofrece ningún procedimiento para aliviar la tarea de asignación de códigos, a pesar de exponer las ventajas de su utilización (7, págs. 57-60). Por su parte, Dagobert Soergel, en el capítulo de su extensa obra dedicado al empleo de ordenadores en la construcción de tesauros, admite la automatización parcial de la asignación de códigos, sólo una vez establecido el índice jerárquico del tesauros (8, págs. 420-448). Es difícil, en efecto, creer que la asignación de códigos se pueda realizar automáticamente de forma aleatoria (7). Más bien se trata de una tarea a realizar en paralelo con la de análisis conceptual de los términos, aunque la utilización de fuentes estructuradas y, ciñéndose al ejemplo del presente trabajo, la relativa simplicidad del establecimiento de relaciones, puede aligerarla. En todo caso, es cierto que, para conjuntos de términos muy amplios del mismo nivel jerárquico, sí sería posible la extensión de las cadenas que representan la notación mediante la representación de los caracteres alfabéticos iniciales de cada uno mediante los códigos ASCII correspondientes. Así, por ejemplo, los específicos de AFRICA (AFR), CAMERUN y CONGO se podrían sustituir por AFR132 y AFR139 («AFR»+«(67+65)», etc.), lo que, naturalmente, seguiría permitiendo la interpolación de nuevos términos.

Entre términos que no son equivalentes ni están conectados jerárquicamente puede establecerse una relación de asociación. Esta relación advierte al indizador

o al recuperador de la existencia de puntos adicionales de acceso a la información. El tesauro presenta relaciones apriorísticas entre conceptos y la relación de asociación está influida en buena medida por la práctica de los indizadores y recuperadores de un sistema determinado. A pesar de lo anterior, y de la no existencia de relaciones asociativas entre los conceptos incluidos en el tesauro geográfico aquí tratado, la relación de asociación podría expresarse mediante la creación de un nuevo fichero que contuviera, para cada término relacionado, el código de su término asociado. Al iniciarse el proceso de revisión de los términos por orden alfabético, cada término preferente sería investigado en este fichero adicional y el código asociado conduciría, a través del fichero ordenado por códigos, al término relacionado en cada caso. Esta solución puede ser más efectiva que la sugerida por Batty (7). La generación de relaciones recíprocas podría, en este caso, apoyarse en el funcionamiento del sistema global.

No cabe la menor duda de que, en manos de un programador experto, el procedimiento descrito ha de resultar mucho más ágil. Otro tanto puede decirse del empleo de microordenadores con capacidad de procesamiento elevada. A pesar de su aparente farragosidad, sin embargo, el conjunto de reglas, registros, ficheros y relaciones descritos puede rendir un alto servicio a las tareas de control de vocabulario incluso en grandes sistemas.

10. Addenda

Durante la realización del presente trabajo, se ha comunicado la existencia del programa BEAT, creado por Josep Sau en el contexto del sistema documental Sinera del Programa d'Informática Educativa de la Generalitat de Catalunya, para la elaboración, mantenimiento y consulta de tesauros documentales sobre microordenadores PC compatibles (*Information World en Español*, núm. 4, mayo 1992, pág. 6). No cabe la menor duda de que tal programa se podrá utilizar para alcanzar los mismos objetivos que los aquí descritos.

Referencias bibliográficas

1. MILSTEAD, J. L. Thesaurus software packages, *Proc. ASIS*, 3-15, 1990.
2. MILSTEAD, J. L. Specifications for thesaurus software, *Inf. Process. Manag.*, 27 (2-3), 165-175.
3. MILSTEAD, J. L. Thesaurus software packages for personal computers. Database 1990 december, 61-65.
4. SORGELD, D. Indexing languages and thesauri: construction and maintenance. Los Angeles, Melville, 1974.
5. UNE 50-106. Directrices para el establecimiento y desarrollo de tesauros monolingües. Madrid, AENOR, 1990.
6. AMAT, C. B. Generación de archivos bibliográficos personales mediante descarga e importación de Indice Médico Español en CD-ROM a dBASE III+, *Med. Clin.* (Barc.), 97 (20), 579-588, 1991.
7. BATTY, D. Microcomputers in Index Language Design and Development, *Microcomput. Inf. Manag.*, 1 (4), 303-312, 1984.

8. BATTY, D. Thesaurus construction and maintenance: a survival kit. Database 1989 february, 13-20.
9. AITCHISON, J., y GILCHRIST, A. Thesaurus construction: a practical manual, 2nd ed., London, ASLIB, 1987.