TESAUROS EN HTML. UN MODELO DE DISEÑO Y ESTRUCTURA PARA SU CONSULTA EN LA MALLA MUNDIAL (WWW)

Antonio Valle Bracero, Alfredo del Rey Guerrero, Jorge Páez Mañá y Reyes Valle Bracero*

Resumen: Se describe un modelo de diseño y estructura de tesauros para su consulta y utilización en tratamientos de indización de bases de datos a través de la malla mundial (WWW). Como fuentes de estudio de partida se analizaron los modelos empleados en diferentes gestores de documentación de ayuda: AcroRead, HTML Help, NetHelp, etc., utilizados para facilitar, a los usuarios de las diferentes aplicaciones, las normas de uso de las mismas. Para la elaboración del trabajo, se han adoptado las directrices expuestas en «Microsoft's HTML Help».

El formato origen de los tesauros piloto, empleados en las pruebas, ha sido el establecido para la aplicación CAT (Confección Automática de Tesauros) desarrollada en el CINDOC. Desde dicho formato, con la metodología descrita en el presente trabajo, se ha obtenido un árbol de ficheros HTML que incorpora índices de búsqueda, localización y enlace, que permiten al usuario, mediante un programa navegador, un posicionamiento en el lugar del texto que hace referencia al término objeto de su búsqueda. En dicho posicionamiento pueden observarse indicaciones sobre las relaciones del término seleccionado con otros términos. Estos últimos pueden servir, a su vez, como elementos idóneos para realizar nuevos posicionamientos.

Las pruebas realizadas sobre diferentes tesauros, tanto en red como en CD-ROM, han sido satisfactorias.

Palabras clave: informática documental, vocabularios científicos controlados, lingüística computacional, gestión de la información.

Abstract: A new model is described for outlining and structuring thesauri for their use as an aid in the database indexing through the World Wide Web (WWW). Several models used in the different on-line help managers were studied and used as main sources on the start up: AcroRead, HTML Help, NetHelp, etc. These on line help systems are used to provide easy access for the users to the operation guidelines for several commercial software packages. For this work the guidelines included in the «Microsoft's HTML Help» were used.

The format used for the prototype thesauri in the trials is the native format of the CAT (Confección Automática de Tesauros) application developed in CINDOC. From this format and using the methodology outlined in this paper, an HTML file system including search indexes, location indexes and links was developed. This file system allows the user to go to the precise point of the text that refers to the search term. In that point the user will find a number of indications about the relationship between the search term and other related items. These later ones can be also used as starting points for new searches.

* Centro de Información y Documentación Científica (CINDOC), (CSIC). Correo-e: tonio@cindoc.csic.es
Recibido: 1.ª versión: 12-11-99; 2.ª versión: 2-4-00.

The preliminary trials made with different thesauri, both on-line and on CD-ROM have provided very good results.
Keywords: automated documentation, controlled scientific vocabularies, computational linguistics, information management.

1 Introducción y antecedentes

En el marco de los trabajos sobre normalización de la lengua científica española, que viene desarrollando el CINDOC desde 1981, encuadrados dentro de la programación científica nacional, la elaboración del producto informativo que se presenta ha sido el resultado de los estudios encaminados a la introducción de las nuevas tecnologías informáticas (1-3) en la difusión de corpora terminológicos (4-10), tanto vía red telemática como en edición electrónica para consulta local.

El CINDOC, con la experiencia adquirida en la consecución de trabajos sobre traducción automática (11), servicios de traducción, detección de neologismos y acrónimos aparecidos en publicaciones científicas (12), elaboración de aplicaciones informáticas adecuadas al tratamiento de la información científica (13) y desarrollo de productos terminológicos, ha conformado diferentes equipos de investigación para la realización de proyectos encuadrados dentro de las directrices de la programación científica nacional. Entre éstos se encuentra el que ha dado lugar al presente trabajo.

Dicho centro dispone asimismo de un amplio conjunto de vocabularios controlados (glosarios multilingües, tesauros, etc.) que se utilizan tanto para la indización de textos científicos, como para la preparación de perfiles de búsqueda en sistemas automatizados interactivos.

En este trabajo se expone el desarrollo de un tratamiento informático que resuelve la puesta en formato HTML (14-18) de los citados vocabularios, permitiendo obtener una estructuración idónea de este tipo de información terminológica para su consulta vía red telemática o en forma local, facilitando las labores de indización.

En el mismo se indican las aplicaciones informáticas desarrolladas, tanto para llegar al diseño y estructura del árbol de directorios como para la creación de ficheros relativos a los vocabularios, puestos en forma que permita una consulta rápida utilizando un programa navegador, y se reseñan aquellos productos con los que se han hecho las pruebas piloto.

2 Metodología

2.1 Estudios iniciales

Para cubrir el objetivo de disponer de un árbol HTML que se asemeje en lo posible a las publicaciones impresas de tesauros, se realizó un estudio de las aplicaciones informáticas utilizadas para la preparación de ediciones electrónicas, analizando las diferentes opciones, principalmente AcroRead, HTML Help y Net Help, y observando que la mayoría de ellas, en su aspecto externo, disponen de parecidas prestaciones.

En el entorno de trabajo del equipo de investigación, se había adquirido experiencia en la preparación de tesauros y otros productos informativos en forma consultable informáticamente mediante la aplicación Microsoft Windows Help, es decir, en la confección y uso de ficheros con extensión HLP. A fin de rentabilizar esta experiencia se

decidió, en un primer momento, ver la viabilidad de conversión directa de los ficheros HLP a ficheros HTML, o en su caso, el paso desde los ficheros RTF (formato con texto enriquecido que se usa como intermedio normalizado dentro de las aplicaciones para edición electrónica e impresa), fuente de los ficheros HLP, a los ficheros HTML.

2.2 Versión HLP

Para la preparación de una edición electrónica de los tesauros (19-21) y glosarios, que venían editándose en el CINDOC en forma impresa, se utilizó la aplicación Help-Compiler de Microsoft como vía para preparar la edición informatizada, en CD-ROM, de dichos productos. El resultado final obtenido, para cada producto, quedó constituido por un fichero con formato HLP, transportable y entendible, bajo la plataforma Windows, por el gestor Windows Help de Microsoft.

Las especificaciones de los ficheros RTF de información y HPJ de organización de datos, conformantes de la estructura del proyecto, están explicitadas en las normas de uso del asistente de Microsoft® HC (HelpCompiler), que tiene como salida los ficheros HLP.

La aplicación desarrollada para la consulta en CD-ROM de los productos así obtenidos, consistió en un programa Borland® Delphi (22-23), que aúna, bajo Microsoft Windows, los cuatro índices de consulta previstos en su diseño.

Las características de los índices de los tesauros (24), son los siguientes:

a) ***Índice jerárquico***. Permite visualizar globalmente, en forma arborescente, la estructura completa del tesauro. Se presenta en dos columnas: la 1.ª contiene el código de clasificación asignado al descriptor, y la 2.ª el descriptor mismo al que corresponde dicha clasificación.

El código, que permite determinar la ubicación del descriptor en la organización temática se identifica con una secuencia de dígitos en la que cada posición, de izquierda a derecha, determina tanto la secuencia del nivel de profundidad del descriptor, como la relación jerárquica con los términos genéricos y específicos asociados a dicho nivel.

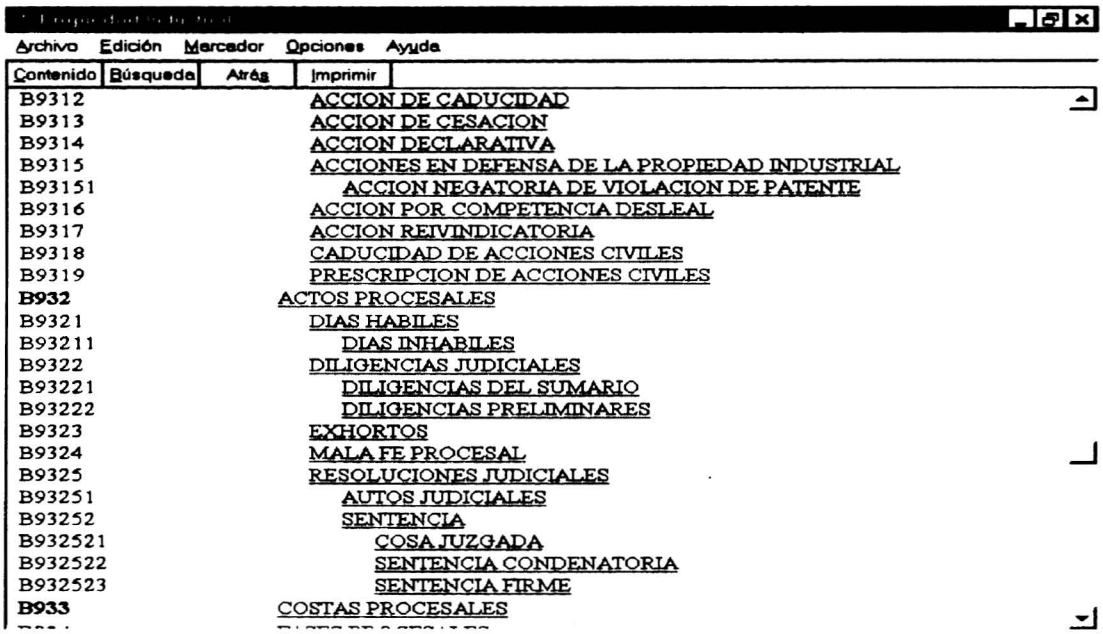
En la figura 1 puede observarse un ejemplo de este índice.

Cuando el tesauro se encuentra inserto en una estructura clasificadora temática (25), la aplicación almacena el contenido de la cabecera temática en un archivo aparte, a efectos de eludir que en los tratamientos de los descriptores (relaciones jerárquicas y asociativas, reenvíos, índice kwoc...) se introduzcan, como términos del tesauro, los encabezamientos de materias que conforman dicha cabecera.

La cabecera temática, que admite hasta un segundo nivel de profundidad, puede consultarse en el fichero de familias que sirve de supraestructura a las diferentes subfamilias del tesauro.

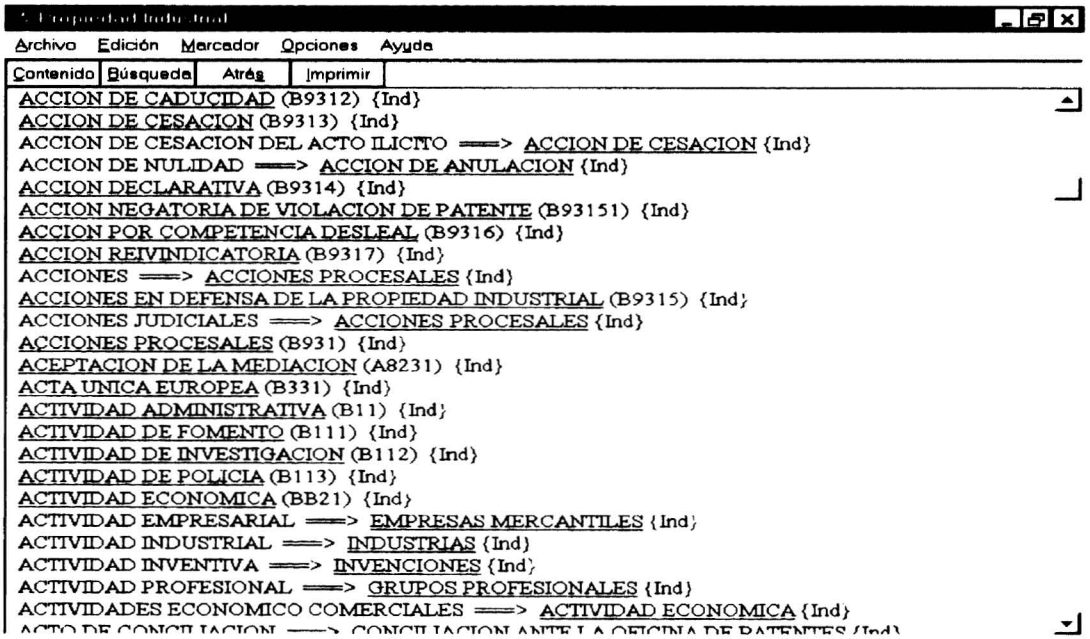
b) ***Índice alfabético***. Presenta, en orden alfabético, la totalidad de los descriptores (términos preferentes que representan conceptos determinados) y no descriptores o reenvíos (términos no preferentes o absorbidos por descriptores). Cada descriptor se ve complementado con la secuencia de dígitos indicativos de su código de clasificación. Cada no descriptor se complementa con el término descriptor al que ha sido reenviado.

Figura 1
Índice jerárquico



En la figura 2 puede verse un ejemplo de presentación de este glosario que puede ampliarse con un encabezado que muestre las letras del abecedario para facilitar el posicionamiento en pantalla de un término deseado.

Figura 2
Índice alfabético

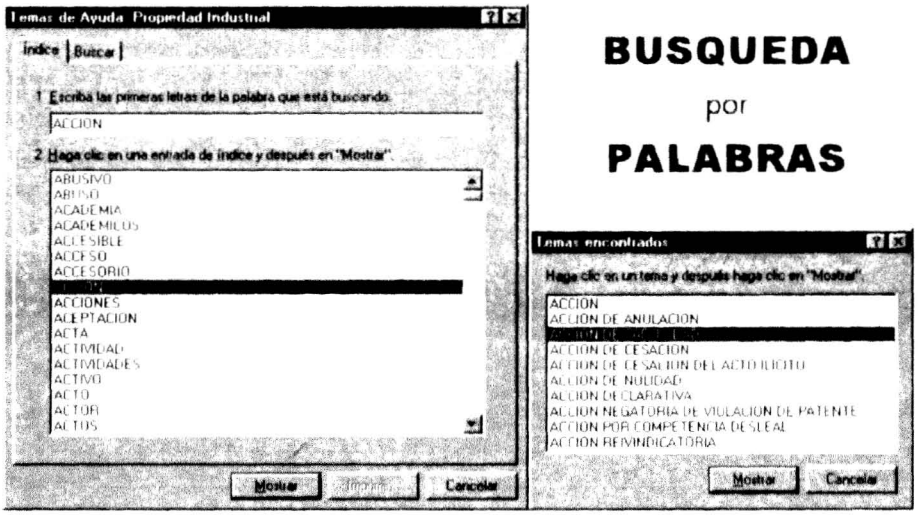


c) ***Índice permutado KWOC***. Permite conocer el conjunto de descriptores y no descriptores que contiene una determinada palabra en su representación formal con independencia del lugar que ésta ocupe en el contexto del término.

Dentro de la aplicación se accede a este índice pulsando un botón situado en la parte superior de la ventana con el texto «búsqueda». Esta acción visualiza un cuadro de doble ventana mediante el que, indicando la raíz de una deseada palabra, se localizan, en la de la izquierda, aquéllas que en el índice arrancan con dicha raíz, pasando luego, una vez seleccionada la deseada, a mostrar, en la de la derecha, los términos completos que la contienen. Seleccionando el término elegido y pulsándolo en forma doble, o en el botón «mostrar», se presenta en pantalla una nueva ventana con la página del tesauro jerárquico donde se encuentra dicho término.

En la figura 3 puede observarse un ejemplo de este índice.

Figura 3
Índice Kwoc



d) ***Índice conceptual***. Recoge, en forma alfabética, la totalidad de los términos integrados en el tesauro. Cada término lleva incorporado su entorno conceptual y temático. En el entorno conceptual se contemplan los siguientes aspectos:

1. Respecto a los términos preferentes o descriptores
   - — Nota de alcance
   - — Equivalencia idiomática
   - — Clasificación temática
   - — Sinonimia
     - • Usado en lugar de (*no descriptor*)
   - — Relación jerárquica
     - • Términos más genéricos
     - • Términos más específicos
   - — Relación de afinidad (término relacionado)

2. Respecto a los términos no preferentes o no descriptores
   - Use (*descriptor*)

Para la visualización de este índice se utiliza el gestor de recuperación SERIOMIC (figuras 4 y 5). Este gestor, tras la realización de la estrategia de búsqueda adecuada, sitúa en pantalla el término junto con sus entornos conceptual y temático, permitiendo enlazar con el índice KWOC de cada una de las palabras significativas que constituyen el término cabecera obtenido de la búsqueda previa.

En las figuras 4 y 5 pueden observarse dos ejemplos del índice conceptual.

Con la aplicación ya citada, Borland Delphi para Windows, se coordina la consulta de los diferentes índices.

2.3 Versión HTML

La permanente preocupación de disponer de metodologías que facilitasen la elaboración de productos informativos gestionados por las más modernas tecnologías de difusión de datos (publicación sin papel) generó la preparación de ediciones electrónicas de los productos del CINDOC tanto en soporte CD-ROM como en WEB incorporados a la red telemática.

Dado que las formas más actualizadas de amplia difusión de la información vía telemática utilizan, en forma generalizada, el formato HTML (hipertexto) y sus variaciones de adecuación al uso de multimedios, se decidió utilizar este formato para las ediciones electrónicas del CINDOC.

La experiencia adquirida en la preparación de la versión HLP, explicitada anteriormente, y los estudios de conversión de ficheros en formato RTF a formato HTML, dieron lugar a la realización de pruebas con diversas aplicaciones RTFaHTML. De estas primeras pruebas de conversión, para las que se utilizaron varias de las aplicaciones preparadas en formato HLP, se obtuvieron las siguientes conclusiones:

— Ninguno de los conversores utilizados respondió a las expectativas previstas en el objetivo propuesto.

Figura 4
Índice conceptual (término no preferente)

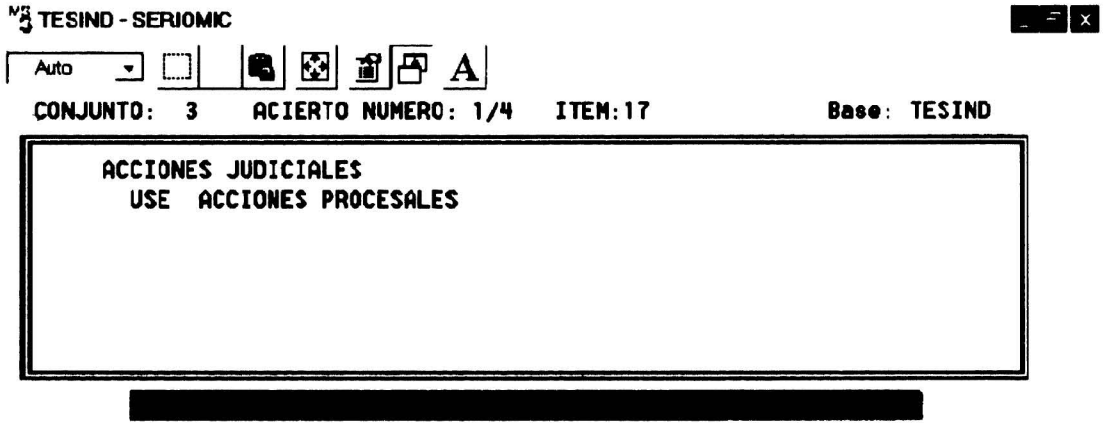


Figura 5
Índice conceptual (término preferente)

DELITOS CONTRA LA LIBERTAD SEXUAL
Cl Ikl
en Crime against sexual freedom
FR Délit contre le liberté sexuelle
na Se entienden por tales a los reguladosen el Título del Libro II del Código Penal
= Delito sexual
= Delitos contra la honestidad
< Delitos
< Derecho Penal
> Abusos sexuales
> Estrupo
> Incesto
> Acoso sexual
> Agresiones sexuales
> Violación
> Escándalo público
> Exhibicionismo
> Prostitución
> Proxenetismo
> Trata de blancas
> Provocación sexual
> Corrupción de menores
> Pornografía
> Difusión de material pornográfico
– Lugar de la acción delictiva
– Perdón del ofendido

— El nivel de RTF aceptado por las aplicaciones RTFaHTML evaluadas era inferior al empleado en el diseño de las aplicaciones HLP elaboradas en el CINDOC.
— La revisión que el resultado requería para adecuarlo al diseño final del objetivo previsto, incluida la consulta y visualización del resultado, resultaba excesivamente complicada.
— Por el camino iniciado los resultados obtenidos no eran suficientemente satisfactorios.

Teniendo en cuenta estas conclusiones se decidió, como mejor solución, desarrollar una aplicación que permitiera efectuar la conversión de formatos, formando parte, como un eslabón más, de la cadena de aplicaciones que, dentro de la preparación de tesauros, se venía desarrollando en el CINDOC.

Como diseño de presentación del tesauro se consideró adecuado el utilizado en la metodología Microsoft®, como equivalente HTML al formato HLP, ya que los ficheros obtenidos, referenciados con extensión CHM (HTML compilado), son de organización gemela a la establecida en los ficheros HLP, elegida en las versiones anteriores de edición electrónica elaboradas en el CINDOC.

Para la preparación de los ficheros CHM se partió de un conjunto de ficheros HTML, que se corresponden uno a uno con los RTF preparados para la versión HLP. Un fichero proyecto, HHP, especifica al programa constructor las directrices de compilación.

El asistente «HTML Help Workshop» detallado en «Official Microsoft HTML Help», permitió efectuar un conjunto de pruebas, en la línea del objetivo previsto, al ofrecer entre sus prestaciones cubrir las siguientes pautas:

— Convertir nuestros ficheros RTF a HTML
— Crear el fichero CHM correspondiente indicando mediante un fichero proyecto HHP las directrices de compilación. Este fichero es equivalente al fichero proyecto HPJ en la versión HLP.

Así pues, se encontró un paralelismo entre la versión HLP y la versión HTML en la forma:

RTF - HTML
HPJ - HHP
HLP - CHM

A la vista de los resultados se consideró oportuno estudiar con detalle esta solución, para ubicar los tesauros del CINDOC en la red telemática en formato HTML, teniendo presente que, anteriormente, se había preparado una versión de los mismos, en formato HLP, en forma satisfactoria.

Las conclusiones de este estudio llevaron al diseño de una aplicación propia, para la que se tuvieron en cuenta los siguientes apartados:

— Interfaz de consulta.
— Presentación de tablas-índices e información.
— Enlaces hipertexto entre los distintos índices de presentación del tesauro.
— Compatibilidad entre la ubicación en un servidor de red telemática y una edición electrónica en CD-ROM.
— Generalización del modo de presentación de los tesauros en un modelo normalizado, procurando conservar al máximo los criterios de presentación de las ediciones impresas.

El primer estadio a considerar fue cuál podría ser la información de partida para llegar al árbol de ficheros objetivo del proyecto. Conforme a las conclusiones del estudio previo, la presentación final podría estar formada bien por un único fichero CHM o bien por el conjunto de ficheros HTML, cuya compilación daría origen al CHM anteriormente indicado.

Las ventajas de esta segunda opción eran fundamentalmente dos:

— Posibilidad de interpretación desde diferentes plataformas.
— Posibilidad de transmisión parcial con el consiguiente ahorro de tiempo de respuesta a la consulta y en el espacio requerido por la misma en el soporte del cliente. En esta forma, el usuario recibiría únicamente el tramo que contiene la respuesta a lo solicitado.

En cuanto a la fuente de información, desde la que efectuar la conversión a HTML, fueron considerados dos formatos:

• EPI (Texto con epígrafes)
• RTF (Obtenido con la secuencia indicada para HLP)

Tras sopesar los caminos a seguir derivados de la elección del formato escogido, se concluyó que la solución que mejor se adecuaba a la experiencia del equipo investigador y a la idea del objetivo a cubrir, era la de utilizar el formato EPI (figura 6).

Figura 6
Formato EPI

```
TT: ACCION
US: ACCIONES PROCESALES

TT: ACCION DE ANULACION
CL: B9311
UP: ACCION DE NULIDAD
TC: ACCIONES PROCESALES

TT: ACCION DE CADUCIDAD
CL: B9312
TG: ACCIONES PROCESALES

TT: ACCION DE CESACION
CL: B9313
UP: ACCION DE CESACION DEL ACTO ILICITO
TG: ACCIONES PROCESALES

TT: ACCION DE CESACION DEL ACTO ILICITO
US: ACCION DE CESACION
```

Las razones fundamentales en esta decisión fueron:

— Es un formato texto simple sin adenda de controles como ocurre con RTF.
— Es ampliamente utilizado para la exportación desde diferentes editores de texto y bases de datos.
— Se dispone de útiles de tratamiento así como de aplicaciones para su conversión a otros formatos.
— Enlaza con el formato (figura 7) de la aplicación de confección automática de tesauros (CAT), elaborada por el CINDOC, permitiendo preparar ediciones impresas del tesauro.
— Al ser un formato utilizado en diversas aplicaciones documentales del CINDOC, resulta relativamente fácil su adecuación a la tabla de caracteres entendible por la aplicación en uso, evitando los problemas que, en la ordenación alfabética, plantean las vocales con signo diacrítico y la ñ, es decir, aquellos caracteres que, según la tabla elegida por la aplicación, tienen un código interno diferente.

La tabla de epígrafes utilizada, basada en la norma UNE 50-106-90, fue la siguiente:

TT: Término tratado o de cabecera
LT: Líder temático. Cabeza de serie jerárquica. Familia

CL: Código de noción. Clasificación temática
NA: Nota de alcance
UP: Término sinónimo
US: Término no preferente (reenvío)
TG: Término genérico
TE: Término específico
TR: Término relacionado

En el caso multilingüe, el epígrafe se corresponde con la codificación del idioma conformándose mediante dos caracteres (EN-Inglés, FR-Francés, DE-Alemán, etc.).

La secuencia de tareas a realizar para obtener el árbol HTML, a partir del tesauro alfabético conceptual, grabado en el formato de epígrafes anteriormente citado (formato EPI), se realizó a tenor de las siguientes fases (ver organigrama de diseño):

Figura 7
Formato CAT

```
ACCESIBLE AL PUBLICO
  USE DIVULGACION DE LA PROPIEDAD INDUSTRIAL
ACCESO A LOS ARCHIVOS ADMINISTRATIVOS
  CL  B181
  TG1 PROCEDIMIENTO ADMINISTRATIVO
  TR  PUBLICIDAD REGISTRAL
ACCION DE ANULACION
  CL  B9311
  UP  ACCION DE NULIDAD
  TG1 ACCIONES PROCESALES
ACCION DE CADUCIDAD
  CL  B9312
  TG1 ACCIONES PROCESALES
ACCION DE CESACION
  CL  B9313
  UP  ACCION DE CESACION DEL ACTO ILICITO
  TG1 ACCIONES PROCESALES
ACCION DE CESACION DEL ACTO ILICITO
  USE ACCION DE CESACION
```

a. Fase de validación

En esta fase, la aplicación VALITES, en cada término tratado (reseñado con el epígrafe TT):

— Revisa la estructura de las notas de alcance (anotadas como tales mediante el epígrafe NA).
— Analiza las relaciones de jerarquía (TG, TE), afinidad (TR) y sinonimia (UP, US).

Cuando la aplicación detecta algún error en las anteriores validaciones lo pone de manifiesto a fin de que éste pueda ser corregido en forma previa al tratamiento y transformación del tesauro.

La aplicación no valida ni la clasificación (CL) ni la procedencia de los términos líderes (LT), ya que tanto la primera como los segundos se establecen en forma automática por la misma, siendo introducidos a posteriori en el entorno conceptual de los términos de referencia.

b. Fase de preparación y puesta a punto de los ficheros base para la conversión a HTML

Denominamos ficheros base para la conversión a HTML a aquéllos que, en el diseño general de la aplicación, son origen de cada una de las partes interrelacionadas que conforman el árbol de páginas WEB. Éstos responden a cuatro apartados:

A) Familias (*tesa*.FAM —donde *tesa* debe indicar el nombre del tesauro—), y Jerárquico (*tesa*.JCD). Contienen respectivamente 1) la relación de términos líderes temáticos y 2) la estructura jerárquica completa del tesauro precedida del código clasificador de familia, obtenido automáticamente según un algoritmo diseñado al respecto.
B) Alfabético Conceptual (*tesa*.TXT). Almacena, en orden alfabético, todos los términos del tesauro y su entorno conceptual de relaciones con otros términos.
C) Indice KWOC del glosario de términos (xxxKWO.DLM —donde xxx debe identificar el tesauro del que se obtiene el índice Kwoc—). Fichero que contiene las palabras del glosario, ordenadas alfabéticamente, seguidas de los términos completos en los que se encuentran.
D) Indice KWOC de los glosarios de equivalencia idiomática (xxxmmKWO.DLM —donde xxx debe identificar el tesauro del que se obtiene el índice Kwoc y mm el idioma—). Ficheros gemelos del anterior pero obtenidos de los respectivos glosarios de idioma. Estos KWOC, en su posterior tratamiento, necesitan cada uno del fichero relativo al respectivo diccionario, que asimismo ha de prepararse.

Las aplicaciones complementarias MICTESA, KWOC-DLM, DICCIONARIOS, SEIK, CAT y CRINDO, se utilizan para obtener los ficheros indicados anteriormente.

c. Fase de conversión y organización final de los ficheros HTML

Todos los ficheros anteriormente especificados se trataron utilizando como patrón la tabla de caracteres pc8 (437 u 850). Para su tratamiento con la aplicación de conversión (CONVHTML) es necesaria su conversión a la tabla Latin1 (Windows-1252). Ello se realiza con la aplicación CRINDO

La organización final (árbol HTML) presenta el siguiente diseño:

- Directorio raíz del tesauro específico como subdirectorio del que contiene la página principal de llamada.
- Subdirectorio HTML del directorio raíz.

En el directorio raíz se incluyen:

- Fichero .HTM de inicio y archivos complementarios necesarios para el tratamiento.
- Botones de enlace e imágenes.
- Ficheros índice KWOC y Diccionarios.

En el subdirectorio HTML:

- Fichero Alfabético-conceptual.
- Fichero Jerárquico codificado
- Fichero Familias.
- Botones de Página anterior y Página siguiente.

Conformado el conjunto, se procede al paso de la aplicación CONVHTML, en la secuencia siguiente:

a) En subdirectorio HTML

— Etapa PagHTM.
- Entrada: Fichero Alfabético-conceptual
- Salida: Tabla de páginas (.TAB)

— Etapa TesHTML.
- Entrada: Fichero Alfabético-conceptual
Tabla de páginas
Fichero familias
- Salida: Ficheros HTML

— Etapa JerHTML.
- Entrada: Fichero Jerárquico codificado
- Salida: Ficheros HTML

b) En directorio raíz, copiada en el mismo la tabla de páginas

— Etapa KwocHHK.
- Entrada: Fichero KWOC
- Salida: Ficheros índice .HHK y HTM de enlace.

Esta última etapa KWOC ha de repetirse para cada uno de los idiomas.

c) Preparar e incluir en directorio raíz fichero HHC con la tabla de contenido del tesauro y enlaces con la introducción e índices de búsqueda por KWOC. Puede añadirse una entrada directa a las páginas del fichero jerárquico.

d) Preparar e incluir en directorio raíz una página de introducción y normas de consulta del árbol HTML.

3 Pruebas piloto

Para la realización de las primeras pruebas de la aplicación desarrollada CONVHTML se eligieron los tesauros de extensión media:

- Biología Animal
- Propiedad Industrial

Las pruebas iniciales incluían únicamente la preparación del índice KWOC y el tesauro alfabético conceptual, así como el pertinente enlace, observándose que:

a) La consulta resultaba lenta por utilizar los navegadores una metodología de copia a su disco duro de la página en visualización. El tiempo necesario para ello es dependiente de la extensión de la página, es decir, páginas cortas son rápidas y páginas extensas llevan su tiempo.
b) Internet Explorer no permitía la llamada desde el KWOC posicionándose en el término objeto de la búsqueda; siempre quedaba posicionado en el comienzo de página. En esta versión en el primer término del tesauro.

Estas dos incidencias se resolvieron:

a) Estableciendo un algoritmo de división del tesauro en un número de ficheros de longitud adecuada a un tiempo de transmisión razonable.
b) Insertando en cada página un JavaScript que subsanara la inhibición del salto al término buscado.

Comprobada la bondad en el uso de esta metodología en un servidor de red bajo Windows 9x y NT, se efectuaron pruebas bajo UNIX, encontrando:

a) Unix es un servidor muy estricto en el uso de mayúsculas y minúsculas para nombrar las páginas HTML.
b) Para el uso de los JavaScripts insertados en la aplicación no encontramos la solución adecuada.

A la vista de estas pruebas se decidió que en lugar de adecuar la aplicación para su uso bajo Unix, debía desarrollarse y mejorar la metodología de las pruebas anteriormente realizadas. Entre las mejoras se incluyeron las siguientes:

a) Incorporar la visualización del índice jerárquico, creando una página para cada líder temático. Esta página sería accesible mediante el índice de familias correspondiente.
b) Establecer enlaces bidireccionales entre los índices jerárquico y alfabético conceptual.
c) Crear una estructura lo más afín posible a las normas utilizadas en ediciones impresas, facilitando así su uso al interesado.

Desarrolladas y puestas a punto estas mejoras, la aplicación indicada se utilizó sobre un conjunto de tesauros de diferentes temas y procedencias, siendo satisfactorios los resultados obtenidos.

Queda pendiente para el futuro el estudio de subdivisión o agrupamiento de las páginas del índice jerárquico, pues generalmente las extensiones de las diferentes familias no guardan un tamaño homogéneo. Estas extensiones pueden estar en un abanico muy amplio, es decir, unas resultan pequeñas mientras otras quedan excesivamente grandes, con lo que, debido a su diversa extensión actual y la metodología de trabajo de los navegadores, los tiempos de consulta pueden variar de unos instantes a segundos.

La aplicación, si bien está operativa, requiere continuar con los estudios citados para homogeneizar los tiempos de consulta.

4 Relación de tesauros en los que se ha utilizado la aplicación

Además de los dos tesauros citados en las pruebas piloto, es decir, «Biología Animal» y «Propiedad Industrial», se han empleado los siguientes:

Tesauros:

- Ciencias Ambientales
- Derecho
- Economía
- Geología
- Ingeniería Civil
- Política Científica
- Psicología
- Topónimos
- Urbanismo

Ejemplos de Tesauros en versión HTML (figuras 8, 9 y 10):

Figura 8
Portada del Tesauro de Propiedad industrial

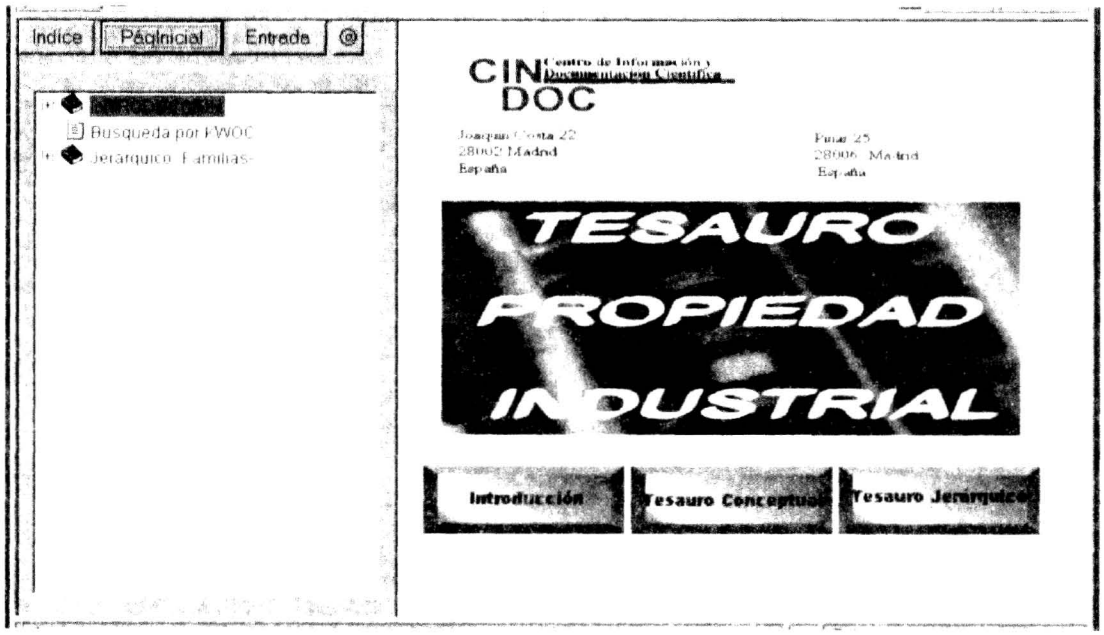


Figura 9
Familia de Identificadores del Tesauro de Propiedad industrial

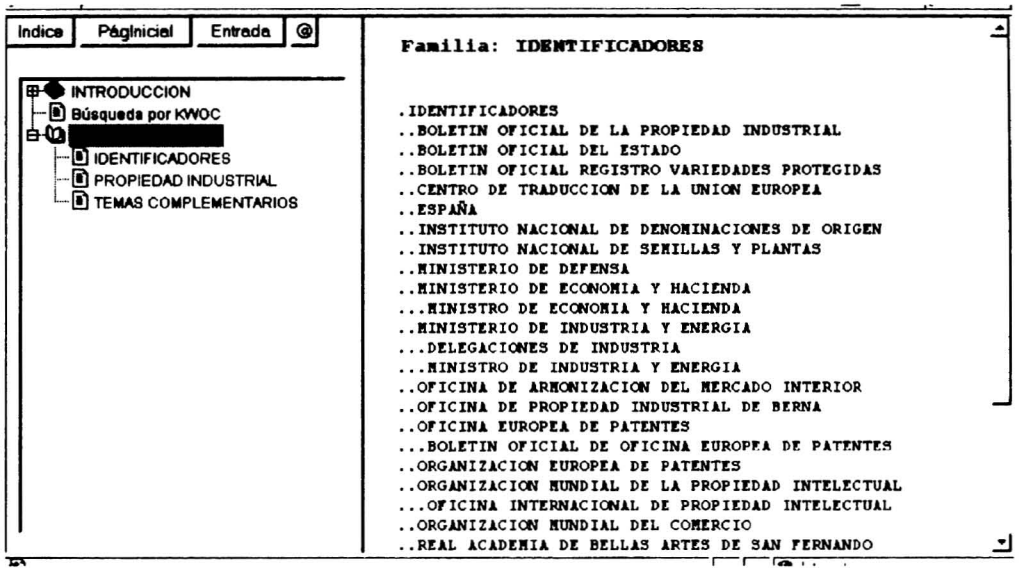
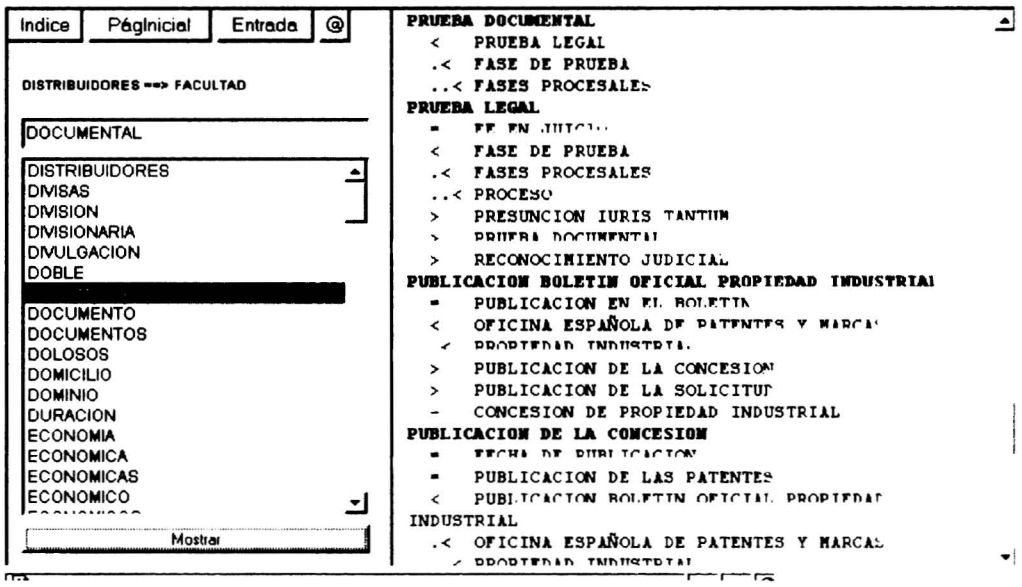


Figura 10
Índice Alfabético-conceptual del Tesauro de Propiedad intelectual



así como los glosarios multilingües:

Glosarios:
- Alimentos
- Drogas
- Máquinas Herramienta

Ejemplos de Glosarios en versión HTML (figuras 11, 12 y 13):

Figura 11
Portada del Glosario de Máquinas Herramienta

Figura 12
Índice Kwoc del Glosario de Máquinas Herramienta

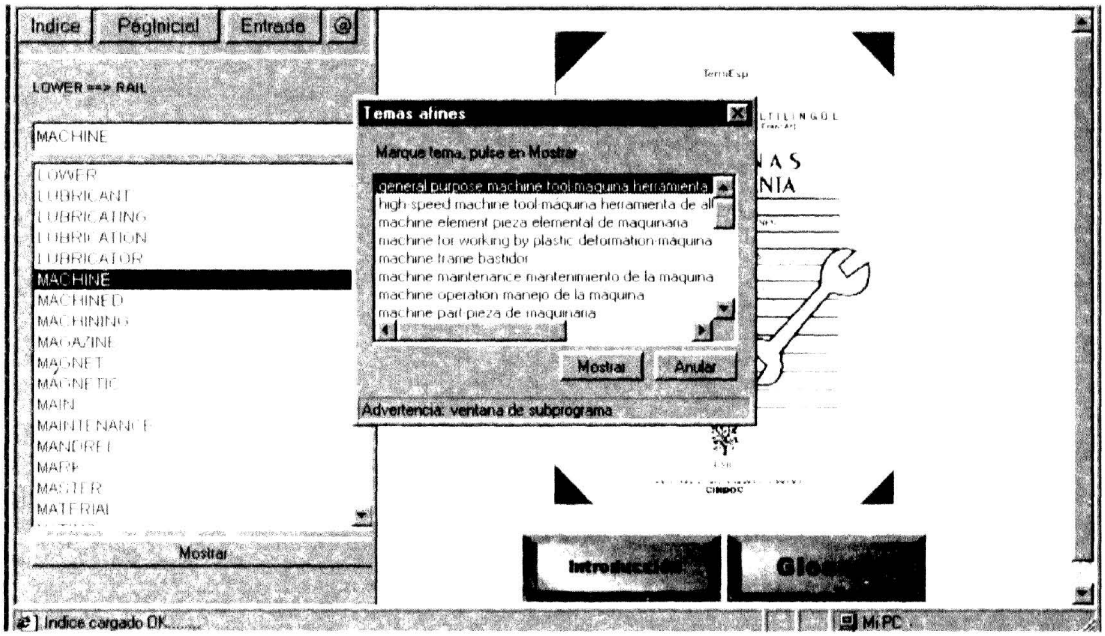
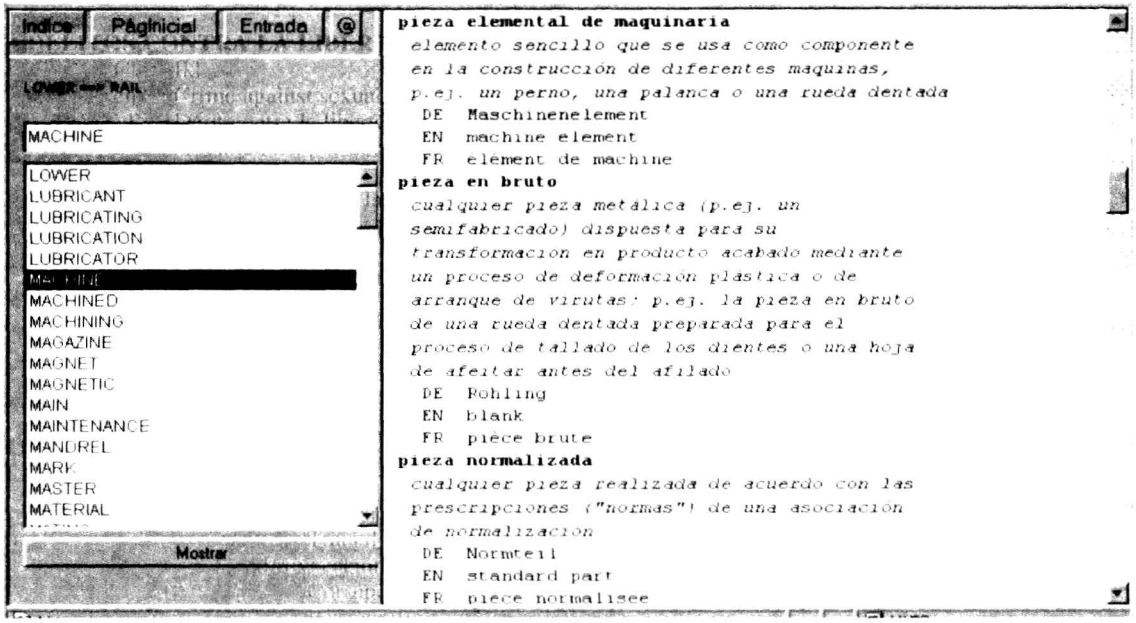


Figura 13
Índice Alfabético-conceptual del Glosario de Máquinas Herramienta



La puesta en el servidor de Red del CINDOC está pendiente, en la mayoría de los mismos, bien del productor o bien del propietario de los derechos de autor. Algunos asimismo están pendientes de revisión, modificación o ampliación en su desarrollo.

En la tabla I se reseñan aspectos documentales sobre los tesauros y glosarios utilizados.

Tabla I

Notas documentales sobre los tesauros y glosarios utilizados

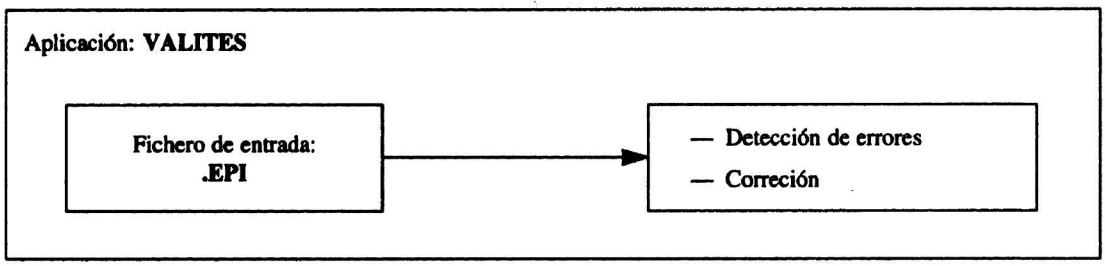
Nombre	*Términos*	*Familias*	*(Lideres tematicos)*	*Idiomas*
Alimentos	3.814			Esp., Ing., Fra., Aleman
Biología Animal	3.781	15	(66)	Español
Ciencias Ambientales	1.754	18		Español
Derecho	19.454	16		Español
Drogas	856			Esp., Ing., Fra., Aleman
Economía	6.888	13		Español, Ingles, Frances
Geología	2.108	23		Español
Ingeniería Civil	10.734	57	(1613)	Español, Ingles
Máquinas Herramientas	990			Esp., Ing., Fra., Aleman
Política Científica (Spines)	10.832	34	(486)	Español, Ingles, Frances
Propiedad Industrial	1.701	3		Español
Psicología	4.363	17	(334)	Español
Topónimos	27.890	200		Español
Urbanismo	4.357	15		Español

5 Beneficios que pueden aportar los resultados

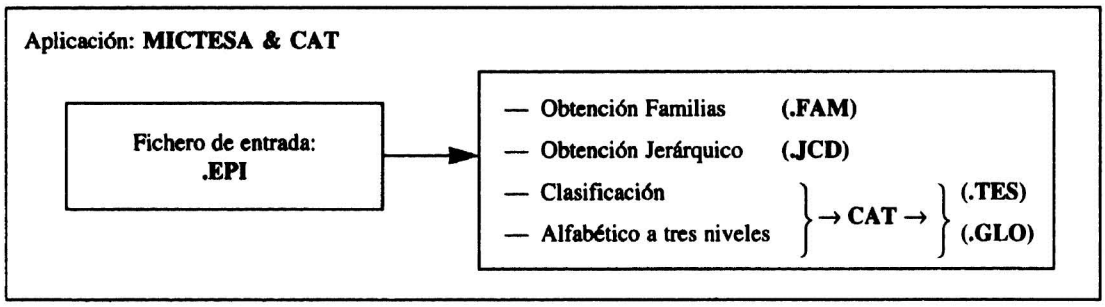
1. Aplicación de este tratamiento automatizado a cualquier tesauro informatizado con el formato descrito en la norma ISO-5964.
2. Automatización del uso de diccionarios terminológicos para la traducción, expresión de sinonimias y diferentes acepciones.
3. Facilitación de la localización de términos de indización y en la indización semiautomática (26).
4. Mejora de la difusión telemática de contenidos y grabación en CD para uso local.
5. Empleo en la elaboración de estrategias de búsqueda para consulta bien en línea bien en CD-ROM.
6. Compatibilización de metodologías de ediciones impresas.

Organigrama general del diseño

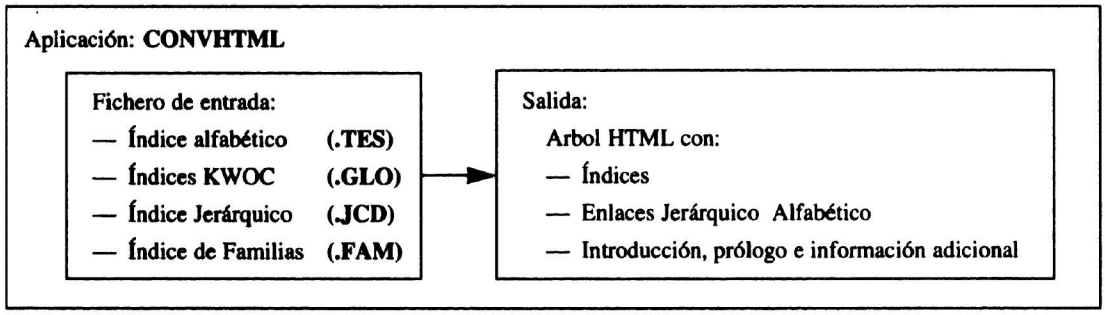
Fase A) Validación



Fase B) Preparación de Ficheros



Fase C) Conversión a HTML



Agradecimientos

Este artículo recoge parte de los resultados obtenidos en el subproyecto 5.º, en el marco del proyecto «Ciencia e Internet», financiado por el Programa Nacional de Aplicaciones y Servicios Telemáticos de la CICYT con el código TEL97-0670.

Bibliografía

1. LANCASTER, F. W. *Indexing and abstracting in theory and practice*. London. Library Association. 1991.
2. ROWLEY, J. The controlled versus natural indexing languages debate revisited: a perspective on information retrieval practice and research. *Journal of Information Science*, 1994 20(2):108-119.
3. SHONFELDT, R. Matematische eigenschaften fur thesaurusrelationen. *Nachrichten fuer Dokumentation*, 1994 45 (4):203-212
4. ARNZT, R. y PICHT, H. *Einfüchrung in die Terminologiarbeit* (1991).
5. CABRÉ, M. T. *La terminologia. Teoría, metodología, aplicaciones*. 1993.
6. INFOTERM. Selected Readings in Terminology, I Terminologi-wissenschaft. 1990.
7. *Inventaire des travaux de terminologie recents*. Québec OLF, 1994
8. JONES, S.; GATFORD, M.; ROBERTSON, S.; HANCKOCK-BEAULIEU, M.; SECKER, J. Interactive thesaurus navigation: intelligence rules ok? *Journal of the American Society for Information Science*. 1995, 46 (1):52-59
9. MELBY, A.; BUDIN, G. y WRIGHT, S. E. Terminology interchange format (TIF). *TermNet News*. 40-1993. págs. 3-64.
10. SAGER J.C. *A practical course in terminology processing*. 1990.
10. VALLE BRACERO, A.; FERNÁNDEZ GARCÍA, J.A. Traducción automática de títulos de artículos científicos del ruso al castellano. *Revista Española de Documentación Científica*, 5, 3, 231-43 (1982).
12. VALLE BRACERO, A.; y FERNÁNDEZ GARCÍA, J.A. Automatización de la indización y coordinación de descriptores. *Revista Española de Documentación Científica*. 6, 1 (1983), págs. 9-16.
13. LAGUNA SERRANO, E.; IRAZAZÁBAL NERPEL, A. y VALLE BRACERO, A. Confección automática de tesauros. *Revista Española de Documentación Científica*. 12, 2, págs. 129-140. 1989.
14. HESLOP, B. y BUDNICK, L. *HTML Publishing on the internet*. Ventana. USA 1995.
15. SIMPSON, A. *El libro oficial de desarrollo con Microsoft Internet Explorer 4* (Site Builder). Mc Graw Hill. Aravaca (Madrid). 1998.
16. WEXLER, S. *Official Microsoft HTML Help Authoring Kit*. Microsoft Press. Redmond, Washington. 1998.
17. ISAACS, S. *Inside dynamic HTML*. Microsoft Press, Redmond, Washington. 1997.
18. GOODMAN, D. *Programación en JavaScript*. Ediciones Anaya Multimedia. S.A. 1997.
19. Norma UNE 50-106-90 (equivalente a la ISO 2788-1986) de *Directrices para el establecimiento y desarrollo de tesauros monolingües*. AENOR, 1990.
20. Norma UNE 50-125 (equivalente a la ISO 5964-1985) de *Directrices para la creación y desarrollo de tesauros multilingües*. AENOR, 1997.
21. UNESCO: Programa General de Información. *Directrices para el establecimiento y desarrollo de tesauros monolingües*. París. 1984.
22. CHAPMAN, D. *Building Internet applications with Delphi 2*. Que Corporation. Indianapolis. 1996.
23. CHARTE, F. *Programación con Delphi*. Anaya. Madrid 1996.

24. AITCHISON, J.; GILCHRIST, A. *Thesaurus construction. A practical manual.* London. Aslib. 1987.
25. *Headings for tomorrow: public access display of subject headings*. Chicago. American Library Association. 1992.
26. VALLE BRACERO, A.; FERNÁNDEZ, J.A.; y MORALES FERNÁNDEZ, R. Separación automática de lexemas, sufijos y morfemas y su aplicación a la traducción automática. *Revista Española de Documentación Científica.* 7, 3 (1984), págs. 185-192.