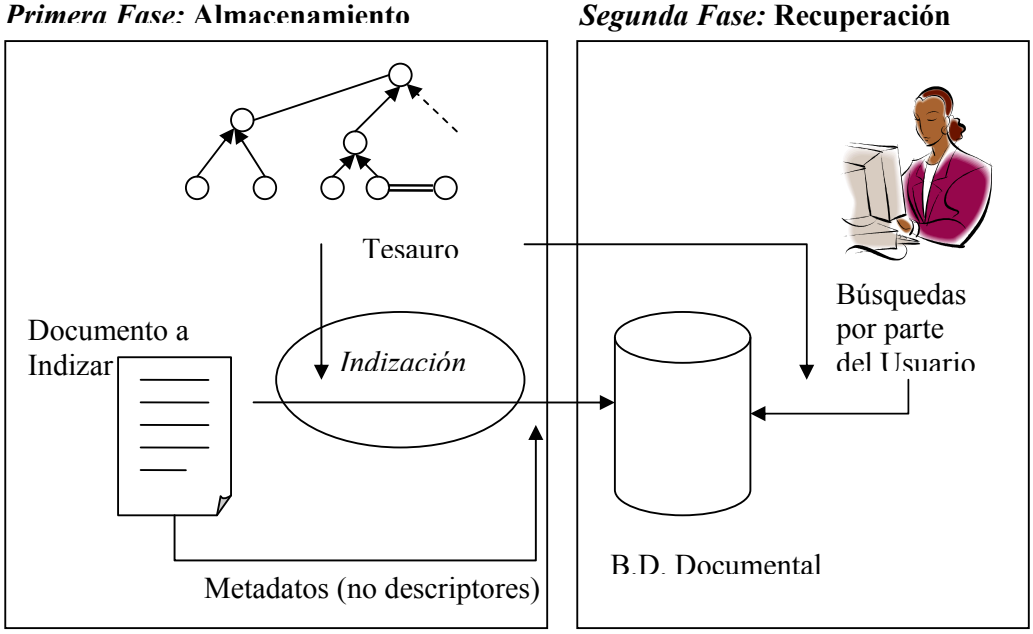
1. ¿CÓMO SE UTILIZA EL TESAURO PARA RECUPERAR: LA INDIZACIÓN?

La INDIZACIÓN es el proceso (automático o manual) que asigna descriptores (términos descriptivos) a un documento para mejorar su posterior recuperación.
Por documento se puede entender casi cualquier cosa: un libro, una noticia, un artículo, una página Web, una botella de vino, una obra de teatro, una edición concreta de un festival de teatro, una película, aplicaciones informáticas, juegos,….



Por tanto, la creación de un tesauro es anterior a la indización (esto es, insertar el documento con los términos del tesauro en la base de datos). La BD documental no sólo contiene los descriptores del documento, sino también otros metadatos como título, idioma, localizadores, origen del documento…
La indización puede ser automática o manual. En la práctica los sistemas automáticos solo suelen implementarse cuando se tienen objetos textuales como recurso a recuperar.
En un sistema idóneo el usuario que busque información podrá utilizar opcionalmente los descriptores del tesauro para recuperar información, eliminando así problemas de sinonimia y polisemia.

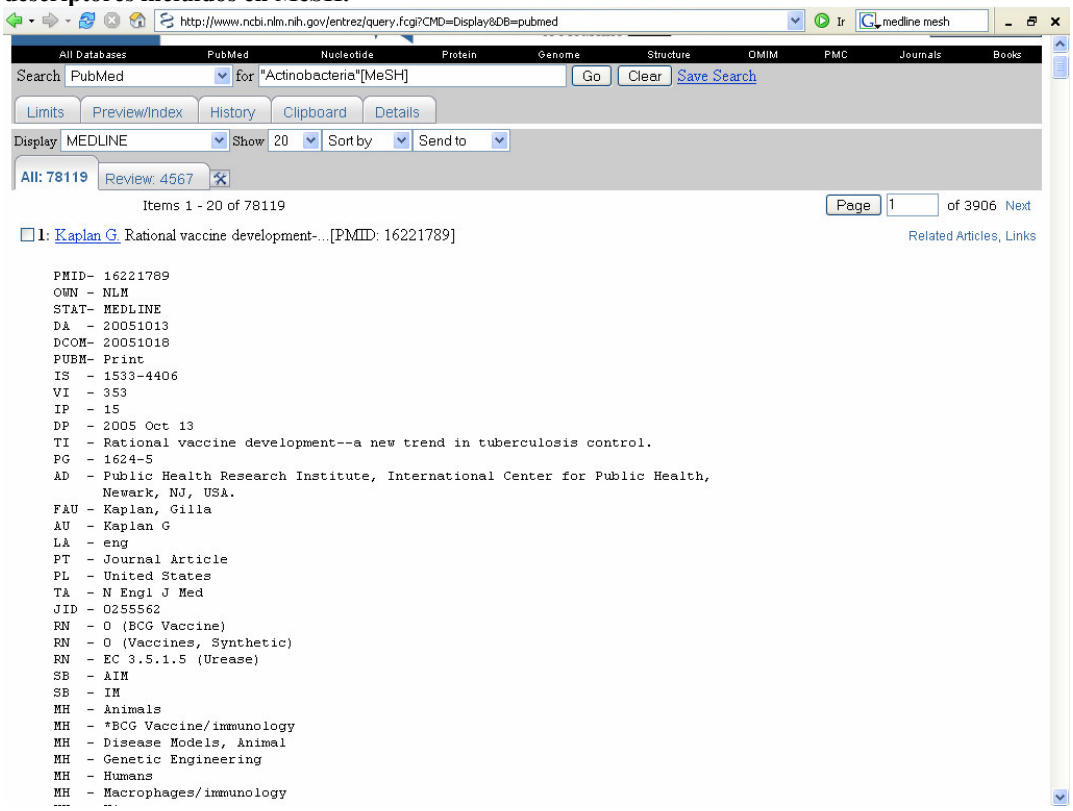
EJEMPLOS DE DOCUMENTOS INDIZADOS

I. Documento procedente del ICYT-CSIC. El objetivo de esta base de datos documental es recuperar artículos científicos y el perfil del usuario son científicos especializados en el dominio. Los descriptores están presentes en el tesauro del CINDOC de biología animal http://pci204.cindoc.csic.es/tesauros/Biol_Ani/Biol_Ani.htm, la información de topónimos (se corresponde con una faceta ajena al dominio de biología animal) procede del tesauro de topónimos http://pci204.cindoc.csic.es/tesauros/Toponimo/Toponimo.htm . El apartado de clasificación procede de la Clasificación de la UNESCO Existen facetas como lengua y tipo de documento. Si bien en el tesauro de biología animal estos términos no aparecen por no ser propios del dominio de biología sino de la descripción del objeto de búsqueda (recuperación de artículos científicos)

Núm. Registro: 173059
Autores: Zabala, Jab;Zuberogoitia, Iñigo
Título: Estado actual del conocimiento del visón europeo (Mustela lutreola) en Bizkaia.
Título en inglés: Status of the knowledge on the european mink (Mustela lutreola) in Biscay.
Lugar de trabajo: Zool. Anim., Bilbao, España;Logroño, España
ISSN: 0214-915X
Revista: Estudios del Museo de Ciencias Naturales de Alava (Datos revista)
Datos fuente: 2003-2004, 18-19: 187-192, 29 Ref
Tipo documento: Artículo de revista
Lengua: Español
Localización: ICYT
Descriptores: Mammalia;Visón europeo (Mustela lutreola);Mustelidae;Hábitat;Ecología;Dieta alimentaria;Mortalidad;Conservación de especies;Distribución espacial;Distribución geográfica
Topónimos: Vizcaya;España
Clasificación: 240118 Mamíferos

II. La siguiente pantalla es de la bases de datos documental LISA, para seleccionar los términos tiene una pestaña denominada tesauros.

III. Esta pantalla de PubMed permite buscar mediante el tesauro MeSH artículos aparecidos publicaciones médicas. La búsqueda remite a "Actinobacteria[MeSH]" indicando que debe estar indizado el artículo por ese término. En la referencia que aparece en la parte inferior se ven otros descriptores incluidos en MeSH.



2. ¿EXISTEN DIFERENCIAS EN LA INDIZACIÓN SI SE HA UTILIZADO UN TESAURO FACETADO O SE HA UTILIZADO UN TESAURO NO FACETADO?

Facetas y familias son en ocasiones difícilmente distinguibles. La diferencia en ocasiones es sutil en cuanto a la construcción de tesauros, pero son diferencias marcadas cuando se trata de aplicarlas en indización.

Un tesauro facetado es aquel en el que se hacen clasificaciones paralelas para indizar el recurso objetivo. Normalmente coincide con los atributos de una clase de UML, o con los elementos que responden a las cinco preguntas básicas o metadatos no temáticos sino descriptivos del recurso a recuperar.
Ejemplos
Un tesauro facetado para describir objetos del museo arqueológico podría ser:

FORMA	MATERIAL	ESTILO/ PERIODO	Conservación	LUGAR DE ORIGEN
Ajuar doméstico >Vasija >Plato Vestimenta >Fibula >Hebilla …	Arcilla Cerámica Metal >Hierro >Bronce …	Helénico Romano Renacentista …	Deteriorado Restaurado Buena Conservación …	Europa >Grecia >Italia Asia >China …

Así determinado objeto debería ser descrito por uno o más de un descriptor de cada faceta: Fíbula-Bronce-Helénico-Restaurada-Grecia

Una clasificación parecida se puede ver funcionando como AAT http://www.getty.edu/research/conducting_research/vocabularies/aat/

.... Associated Concepts Facet
..... Associated Concepts
.... Physical Attributes Facet
..... Attributes and Properties
..... Conditions and Effects
..... Design Elements
..... Color
.... Styles and Periods Facet
..... Styles and Periods
.... Agents Facet
..... People
..... Organizations
.... Activities Facet
..... Disciplines
..... Functions

..... Events
..... Physical and Mental Activities
..... Processes and Techniques

Estas clasificaciones facetadas suelen ser muy útiles en sitios Webs, sobre todo para organización de sites en Internet

La diferencia con las familias es que en estas no se necesita necesariamente tomar un descriptor de cada faceta para describir un objeto. Además, los descriptores en cada familia tienen sentidos de forma aislada cuando describen un objeto, esto no siempre ocurre en los tesauros facetados. Una familia es un tema independiente perteneciente a un dominio.

Un ejemplo es el tesauro anteriormente mencionado de Biología Animal del CINDOC, aquí se muestran las principales familias (identificadas por tener un número delante del término) y sus específicos inmediatos

01 Anatomía y Morfología
NT: ABDOMEN
NT: CABEZA
NT: EXTREMIDADES
NT: FLUIDOS CORPORALES
NT: GINANDROMORFISMO
NT: ORGANOS DEL CUERPO
NT: POLIMORFISMO
NT: QUETOTAXIA
NT: SEXO
02 Biometría
NT: DATOS BIOMETRICOS
03 Biología celular
NT: CICLO CELULAR
NT: CELULAS
04 Ciencias
NT: BIOLOGIA
NT: VETERINARIA
NT: PALEONTOLOGIA
05 Ecología
NT: ASOCIACIONES DE ORGANISMOS
NT: DINAMICA DE POBLACIONES
NT: DIVERSIDAD ECOLOGICA
NT: ECOSISTEMAS
NT: NICHO ECOLOGICO
06 Etología
NT: COMPORTAMIENTO ANIMAL
07 Filogenia
NT: EVOLUCION BIOLOGICA
NT: TEORIAS EVOLUTIVAS
08 Fisiología
NT: EFECTOS FISIOLOGICOS
NT: FERTILIDAD
NT: METAGENESIS
NT: PROCESOS FISIOLOGICOS
09 Histología
NT: EFECTOS HISTOPATOLOGICOS
NT: TEJIDOS BIOLOGICOS
10 Ontogenia
NT: CICLO BIOLOGICO
NT: CICLO CELULAR
NT: ESTADOS DE DESARROLLO
11 Paleontología
NT: BIOZONAS
NT: FOSILES
NT: FOSILIZACION
12 Producción animal
NT: CRIA ANIMAL
NT: EXPLOTACIONES GANADERAS
NT: CONEJOS
NT: COLMENAS
NT: AVES DE CORRAL
NT: GALLINEROS
NT: GANADO
NT: PISCIFACTORIAS
NT: PRODUCCION ANIMAL
NT: REDILES
13 Taxonomía y sistemática
NT: EUCARIOTAS
NT: CLAVES (TAXONOMIA)
NT: CATALAGO (TAXONOMIA)
NT: DIFERENCIACION DE ESPECIES
NT: IDENTIFICACION DE ESPECIES
NT: NUEVA CITA
NT: PRIMERA CITA
NT: PROCARIOTAS
NT: TAXONES
NT: VIRUS
14 Técnicas analíticas e instrumentales
NT: ESTUDIO BIOLOGICO
NT: TRABAJOS DE CAMPO
15 Zoología

Para la utilización de este tesauro basta con seleccionar los descriptores del tesauro, no siendo necesario, como era el caso de las facetas, seleccionar un término de cada familia.

Así un estudio sobre "Medidas estadísticas del cráneo del Iguanodon", podría tener de descriptores: "Cabeza, Datos Biométricos, Paleontología, Fósiles"

3. ¿QUÉ ELEMENTOS DEL DOMINIO DEBE CONTENER EL TESAURO?

Los términos del tesauro contienen los términos que describen al objeto cuya recuperación se pretende mejorar. El tesauro no debería contener los objetos que pretende describir sino el vocabulario para describirlo.

> Ejemplo:
> Si mi propósito es crear un sistema de recuperación que recupere obras de teatro como "la vida es sueño" no tendré que poner el nombre de esta obra en el tesauro, sino el vocabulario para describirla "teatro, siglo de oro, Pedro Calderón de la Barca, castellano".
> Pero si lo que quiero describir son festivales de teatro, y en concreto tengo "Festival Internacional de Teatro Clásico de Almagro 2005" si que podré tener "La Vida es Sueño" como descriptor en mi tesauro entre otros. Así en este festival en la edición del 2005 tendré como descriptores: "La Vida es Sueño, Calderón de la Barca, Centro Dramático de Aragón, El Quijote,…".

La violación de esta norma conduce a un excesivo número de polijerarquías.

4. ¿PUEDE TENER UN TESAURO HERENCIA MÚLTIPLE?

Si aparece herencia múltiple, lo cual no es deseable, puede ser una clasificación correcta pero frecuentemente son debidos a errores.

Posibles errores:

- Se trata de un homógrafo, es decir son DOS CONCEPTOS DISTINTOS que se escriben igual (algo prohibido en el estándar)
Solución: En este caso se debería añadir una frase clasificatoria o cambiar la grafía.
Ejemplo: si quiero representar en un mismo tesauro "planta" como un edificio industrial y como parte del cuerpo puedo o bien introducir una frase clasificatoria: planta (industria) y planta (anatomía) lo cual está permitido pero está desaconsejado en el estándar por dificultades en automatización. Tb. se pueden modificar los términos originales por "planta industrial" y "planta del píe", esta es la solución más acertada.

- Los términos se pueden repetir por cambios en el criterio de clasificación
Solución: poner un indicador clasificatorio (desaconsejado por el estándar pero admitido) o crear una nueva familia o una nueva faceta (depende del caso pero el resultado es idéntico)

Ejemplo:
Teatro
>Drama
>Comedia
Cine
>Drama
>Comedia

En el ejemplo se ha cambiado el criterio de clasificación provocando herencia múltiple. Ya que teatro, cine, novela referencian al formato de representación, pero no al contenido de lo que se representa (esto es el género). La misma obra "la vida es sueño" puede tener una película, una obra teatral y un libro y no por eso cambia de género. Lo voy a escribir con indicadores clasificatorios (no recomendado).
Obra Artística
>Escenificada
>Textos
Escenificada
>Cine
>Teatro
Textos
>Novela
Cine
Según genero: (esto es el indicador clasificatorio)
>Drama
>Comedia
Según época:
>Clásico
>Actual
Teatro
Según genero:
>Drama
>Comedia
Según época:
>Clásico
>Actual

Esta solución además es errónea pq no doy a entender polijerarquía sino que sugiero que "drama en teatro" es diferente a un "drama en cine" (no es polijerarquía es repetición de dos grafías que representan diferentes conceptos con lo que estaríamos en el caso anterior, es decir tendría que cambiar el término a "drama teatral" y "drama cinematográfico" respectivamente). Hay que recordar que el estándar taxativamente prohíbe representar con el mismo término diferentes conceptos

Obra Artística
>Escenificada
>Textos
Escenificada
>Cine
>Teatro
Textos
>Novela
Cine
Según genero:
>Drama cinematográfico
>Comedia cinematográfica
Según época:
>Cine Clásico
>Cine Actual

Teatro
Según genero:
>Drama teatral
>Comedia teatral
Según época:
>Teatro Clásico
>Teatro Actual

Una solución consiste en llevarlo a diferentes familias/facetas, esto es:
Obra Narrada

>Género
>Formato
>Épocas
Género
>drama
>comedia
Formato
>cine
>teatro
>novela
Época
>Renacentista
>Decimonónica
>Moderna
>Contemporánea

No puedo representar que el cine no puede ser renacentista (como en UML las relaciones negativas tienen poca cabida). Pero si que puedo asociar cine como término relacionado con moderno y contemporáneo e incluso concretarlo en la nota de alcance.

En general, si aparecen varios casos de herencia múltiple se pueden deber a dos causas:
- Se trata de un dominio que se le quiere dotar de una semántica facetada o no se han creado las familias pertinentes

- Se está intentando insertar en el tesauro los recursos a indizar con el tesauro. Si intentará insertar en el tesauro la vida es sueño o Romeo y Julieta aparecería poli jerarquía, ya que ambos son teatro y drama. El error aquí es no tener claro cuál es el objetivo del tesauro esto es indizar un recurso. En este caso una base de datos para recuperar obras de teatro.

Aún así la herencia múltiple se puede dar, pero en general es evitable (ciertamente esta restricción, como cualquier otra, quita semántica al resultado final pero mejora la claridad y su utilización posterior)

5. Si al crear un tesauro con tmCAKE en un ordenador quiero abrirlo en otro ¿QUÉ DEBO HACER?

Este error ocurre por falta de la conexión ODBC. Para solucionarlo se debe hacer lo siguiente:
1.- Cambiar el nombre del fichero .mdb
2.- En el tmCAKE crear un nuevo tesauro con el nombre del Tesauro
3.- Sobrescribir la base de datos .mdb del nuevo tesauro creado con la que se ha cambiado de nombre en el paso 1.

6. ¿QUÉ ENTREGAR EN EL CUADERNILLO?

Organización: Debe ser individual, por cuanto supone una referencia individual para la evaluación del examen en sustitución de la asignatura.

Descripción:
Se tendrá que explicar a alguien los fundamentos de la asignatura, y para eso le planteáis una serie de problemas o dificultades que nos hemos encontrado y la resolución que se ha propuesto con ello, así como problemas y conclusiones encontradas.
De este modo, el cuaderno de Ingeniería de la Información consiste en presentar los problemas y la resolución de los ejercicios y prácticas que se plantearán durante el curso, en concreto por los momentos tenemos **(25/10/2005)**:

1. Práctica de mini tesauro de vinos que hicimos en clase y después cómo eso ha servido para hacer un tesauro mayor.
2. Práctica primera de Tesauros de tema elegido por cada grupo.

<A LO LARGO DEL CURSO ESTE APARTADO CRECERÁ>

Notas:

- ***IMPORTANTE: Se irán incluyendo las prácticas que deben estar en el cuadernillo en posteriores actualizaciones periódicas de este documento. Por tanto debéis estar pendientes.***
- No se trata de que sea algo muy extenso, sino un comentario justificando de lo que se ha visto con las prácticas y ejercicios de clase, y la visión personal interrelacionando en medida de lo posible los diferentes temas tratados. (Es decir, imaginad que fuera un examen y que se os preguntara algo tan general como "justificación de los contenidos de la asignatura" y que se os pidiera que pusierais ejemplos (ejercicios como comparativa TopicMap-mapas conceptuales-tesauros, tesauro vinos, etc.))
- Se trata de que con la ayuda de los ejercicios saquéis vuestras propias conclusiones, siendo claros y concisos.
- Se entregará en papel y en formato electrónico (Disquete o CD)

- Se valorará el acierto en la resolución de los ejercicios planteados, así como la claridad y la concisión en el planteamiento de los argumentos y los temas tratados a lo largo del curso. También se valorará que en las conclusiones, se planteen propuestas de aplicación para los temas tratados en el curso.

7. ¿CÓMO ENTREGAR LA PRIMERA PRÁCTICA DE TESAUROS?

Se enviarán por correo electrónico tres ficheros con los índices jerárquico, alfabético y la base de datos del tmCAKE. Adicionalmente, la guía de uso y explicaciones de creación del tesauro, tema seleccionado, metodología que se ha usado, etc. Se enviará a las direcciones afraga@ie.inf.uc3m.es y jorge@ie.inf.uc3m.es. El asunto del mensaje debe ser PRÁCTICA CREACIÓN DE UN TESAURO. El nombre de todos los ficheros comenzará por GrupoXYZ, donde XYZ será el número de grupo que se asigne. Por favor enviar los archivos comprimidos en formato .zip o .rar

8. ¿ROLES EN LA METODOLOGÍA CAKE?

A continuación tenéis gráficamente como sería la colaboración entre roles y conocimientos.