Ingeniería de la Información

Tema 1:

Organización de la información y Tesauros

Jorge Morato
Anabel Fraga
Dpto. de Informática
Universidad Carlos III de Madrid

Leganés 2006

Organización de la información

¿Qué estamos utilizando para la organización y la recuperación?

=> El Lenguaje

El sistema de comunicación entre los humanos actualmente es el lenguaje

Problema

- En una fiesta, apunto todos los Teléfonos en servilletas: si las guardo en un cajón sin más...
  - ¿Cuánto tiempo necesito para recuperar un teléfono? [necesitaré mucho tiempo]
- Si los guardo en una agenda...
  - ¿cuánto tiempo tardaré en recuperarlos? en poco tiempo
  - ¿y en escribirlos? tardaré más tiempo en escribir el teléfono.

- El tiempo dedicado en clasificar disminuye el de recuperar. Si no clasifico tardaré en recuperar más tiempo.
- Información→clasificación→recuperación de la información

¿Qué es clasificar?

- Es la acción de **distribuir** en varias **clases**, generalmente **disjuntas**, un conjunto de **objetos**.
- También, es el producto resultante de la operación precedente cuando ésta desemboca en un sistema **coherente** y **estructurado**

Ordenar y Clasificar

- Ordenar y Clasificar
  - Libros
  - Fotos
- Es un acto mental:
  - Selección de productos
  - Selección de información

Lenguajes de clasificación o Lenguajes documentales

- Un lenguaje de clasificación divide un dominio de la realidad en una serie ordenada de clases o de subclases (clasificación Dewey, la CDU)
- Un Lenguaje documental está destinado a usos bibliotecológicos. Es un sistema de control de vocabulario para representar el contenido de los documentos.

Lenguaje documental

- Un **lenguaje documental** es un sistema de signos que permita representar el contenido de los documentos pertinentes en respuesta a consultas que tratan sobre ese contenido (Van Slype, 1991)

Tipos de Lenguajes

- Lenguajes de clasificación, representan el contenido de forma sintética
- Lenguajes de indización o combinatorios, permiten representar el contenido de los documentos y las consultas de forma analítica
  - **Lenguajes controlados, construidos a priori, antes de indizar los documentos de una colección**
  - Lenguajes libres, construidos a posteriori, basándose en las palabras en lenguaje natural de los documentos
  - Lenguajes codificados

Clasificaciones documentales

- Monojerárquico
- Facetadas
- Híbridas

¿Por qué lenguajes controlados?

- Clasificación de información, a partir de una lista de vocabulario consensuado
  - Sistemas de Clasificación Bibliográfica
  - Listas de Encabezamiento de Materias
  - Tesauros

¿Qué es la indización?

Es la actividad por la cual se representa el contenido de un documento o de una consulta de forma analítica, es decir, enumerando conceptos y/o las palabras.

IA-1 Indizar este documento

Variantes de Bagle en la 'sopa de virus' de Internet

EFE SAN FRANCISCO (EEUU).- La aparición de **cuatro nuevas variantes** del virus Bagle tiene en jaque a los expertos en seguridad. Las variantes de este gusano amenazan con acabar con las letras del abecedario: se trata del **Bagle.Q** (al parecer, la más extendida), **Bagle.R, Bagle.S y Bagle.T**, hermanos menores del original, que hizo su aparición en enero.

La peligrosidad del virus, que afecta sólo a los sistemas que utilizan el sistema operativo Windows, de Microsoft (y no el Macintosh, de Apple) radica en que no necesita que el usuario abra el fichero incluido en el correo para infectar su computadora, informa la corresponsal de EFE en EEUU Natalia Martín Cantero tras entrevistar a Graham Cluley, consultor de la compañía Sophos. Sin embargo, algunos expertos creen que **los parches que lanzó Microsoft para tapar ese agujero podrían ser insuficientes**, ya que aún así podrían infectarse.

Los comentarios escondidos en el código de programación hicieron pensar a los investigadores que piratas informáticos podrían estar compitiendo entre ellos. Así, Bagle.J contenía una línea en su código que decía: "Hey, Netsky... no arruines nuestro negocio, *quieres entrar en guerra?"

- Variante de Bagle
- "Sopa de virus de Internet
- San Francisco
- variante del virus Bagle
- Expertos en seguridad
- Variante de este gusano
- letra del abecedario
- Peligrosidad del virus
- Sistema operativo Windows
- corresponsal de EFE
- Natalia Martín Cantero
- Grahan Cluley
- Consultor de la compañía Sophos
- Código de programación
- Pirata Informático

agujero	enero	negocio
aparición	peligrosidad del virus	Netsky
Apple	experto	nueva
Bagle.J	experto en seguridad	original
Bagle.Q	fichero	parche
Bagle.R	Graham Cluley	pirata informático
Bagle.S	guerra	SAN FRANCISCO
Bagle.T	hermano	sistema
código	menor	sistema operativo Windows
código de programación	insuficiente	sopa de virus de Internet
comentario	investigador	EFE
computadora	letra del abecedario	usuario
corresponsal de EFE	Línea	Variante de Bagle
consultor de la compañía Sophos	Macintosh	variante de gusano
correo	Microsoft	variante de virus Bagle
EEUU	Natalia Martín Cantero	

IA-2 Indización Manual

Seguridad en Internet

Virus

Gusano

Bagle

Variante gusano

S.O. Windows

Pirata Informático

Correo electrónico

Netsky

Recuerda que:

- **La indización manual es por asignación intelectual de conceptos, la automática por extracción de palabras**
- **La automática suele ser más exhaustiva (muchos descriptores) pero también da más ruido**
- **Consistencia entre ambas baja**

Indización Automática

"... un ordenador reconoce los términos que figuran dentro del título, del resumen, del texto completo [...] empleando estos términos tal cual, o bien después de transformarlos en otros términos, equivalentes o conceptualmente próximos, con el fin de convertirlos en elementos que se incorporan al fichero de búsqueda y quedan disponibles para recuperar el documento"

(Van Slype, 1991)

¿Qué es un Tesauro?

- "Es un **vocabulario** de un lenguaje de indización **controlado**, **organizado formalmente** con objeto de hacer explícitas las **relaciones** a priori entre **conceptos**" (ISO 2788-1986)

  – Seleccionar descriptores de un dominio
  – Establecer relaciones entre descriptores

Relaciones de un Tesauro

- Equivalencia → Para los sinónimos y cuasi-sinónimos
- Jerarquía → **Genérica**, **Parte todo**, **Enumerativa**
- Asociación → Disciplina, Instrumental, Causa, Atributivas, Medición

Tipo de relación	Subtipo de relación	Ejemplo
Jerarquía	Órganos del cuerpo	Esqueleto-articulaciones
	Lugares geográficos	Andalucía-Cadiz
	disciplinas	Biología-botánica
	Estructuras sociales	Ejércitos-divisiones-regimientos
Sinónimos	Jergas	Dolor de cabeza-migraña
	Variaciones idiomáticas	México-Méjico
	Cuasi-sinónimos	Ab intestato-Sucesión AB Intestato
Asociación	Disciplina	Silvicultura-bosques
	Instrumento	Termostato-control de temperatura
	Operación	Proceso de datos-sistema automatizado
	Acción y su sujeto	Reclusión-reclusos
	Concepto y propiedad	Venenos-toxicidad
	Origen	Roma-romanos
	Contraagentes	Plantas-herbicidas
	Unidades de medida	Corriente eléctrica-amperio

Tipos de tesauros

- Los tesauros pueden ser:
- monolingües 2788:1986[i]
- multilingües. ISO 5964: 1985[ii].

[i] Norma UNE 50-106-90. Directrices para el establecimiento y desarrollo de tesauros monolingües. Madrid. AENOR. 1990.

[ii] Norma Une 50-125-1997. Directrices para el establecimiento y desarrollo de tesauros multilingües. Madrid. AENOR.1997.

La norma 2788

Sujeta a las siguientes restricciones:

1) Trata de la presentación y organización de los términos que constituyen un subconjunto controlado del lenguaje natural.
2) Se basa en el concepto de términos preferentes
3) Se limita a los centros que emplean indizadores humanos para analizar documentos y expresar su contenido mediante un lenguaje de indización controlado. No para técnicas automáticas.
4) Tiene como finalidad, indizar colecciones de documentos incluidos en catálogos o bibliografías.

Contradicciones de la norma 2788

- Norma /no obligado cumplimiento
- Control de vocabulario/ no aparece reflejado en el texto de la norma
- Control de vocabulario/ utilizar símbolos para tesauros iguales o se puede utilizar otro idioma (en anexo, 4.1 y 4.2)
- Tesauro monolingüe/ se puede utilizar otro idioma (en 4.1 y 4.2)
- Tesauros manuales/ dan indicaciones para utilizar equipos de tratamiento automático de datos.

Tesauros multilingües

Los tesauros multilingües pueden estar formados por una **relación idiomática**. Para permitir la implementación en otros idiomas en un mismo tesauro.

- la norma define una **lengua fuente y una serie de lenguas objetivos**.
- Los conceptos de la lengua frente se traducen a la lengua objetivo mediante una serie de reglas que dependen de la existencia de **equivalencias exactas, inexactas, parciales, compuesta o no equivalencia**. (1997)

Elementos de un Tesauro

- Términos preferentes - Descriptores
- Términos no preferentes - No descriptores (¿Coches? ¿ChupaChups?)
- Relaciones semánticas (Asociaciones)
- Aplicaciones, *Scope Notes*
- *Facetas*

Reglas en general que deberían tenerse en cuenta para los Términos de los Tesauros

- Nombres o frases nominales
- No suele incluir nombres propios
- No deberían ser generales y representar diferentes áreas temáticas del tesauro

Tratamiento de términos

Directrices	**Ejemplos**
Plural para sustantivos contables	"TUBES"
Singular para materias	"WOOD"
Singular para procesos, propiedades y condiciones	"REFRIGERATION" "WEIGHT", "POVERTY"
No cambiar el orden	"RADAR ANTENNAS" (mejor que "ANTENNAS, RADAR")
Quitar preposiciones	"CARBOHYDRATE METABOLISM" (mejor que "METABOLISM OF CARBOHYDRATES")
Quitar signos de puntuación, abreviaturas y caracteres especiales	"COOPERATIVE PROGRAMS" (mejor que "CO-OPERATIVE PROGRAMS" o "COÖPERATIVE PROGRAMS") "MUSICAL NOTES" (mejor que "(MUSICAL) NOTES" o "MUS. NOTES"

Objetivos de un Tesauro

- Crear un **mapa** de un campo de **conocimiento**
- Crear un **vocabulario estándar** en dicho campo, asegurando la consistencia en la indización y recuperación
- Asegurar que para un concepto sólo se utilizará un término y no sus sinónimos
- **Facilitar** a los usuarios la **localización de nuevos conceptos** mediante las relaciones del sistema
- Servir como **referencia** a los usuarios para la **selección de un término correcto**
- **Expansión o acotación de términos** respecto a **búsquedas** mediante las **relaciones**

Ejemplo de un Tesauro (Orden Alfabético)

AB INTESTATO
  USE SUCESION AB INTESTATO
ABADIA TERRITORIAL
  CL 06020501040501
  BT1 IGLESIA PARTICULAR
  BT2 OBISPOS DIOCESANOS
  BT3 GOBIERNO DIOCESANO
ABALIZAMIENTO
  CL 0906030601
  SN3 01Colocación de señales en el mar para indicar
  SN3 02cualquier peligro.
  BT1 NAVEGACION MARITIMA
  BT2 DERECHO MARITIMO
  BT3 DERECHO DE LA NAVEGACION

Ejemplo Tesauro (Índices)

Índice jerárquico

02 CLASIFICACIÓN DEL VINO
- Vinos añejos
- Vinos claret
- Vinos elegantes
- VINOS ESPECIALES
- . MISTELA
- . MOSTO
- . . YEMA
- . VINOS AROMATIZADOS
- . VINOS ENVERADOS
- . VINOS ESPUMOSOS
- . . CAVA (VINO)
- . . VINOS DE AGUJA
- . . VINOS GASIFICADOS
- . VINOS FINOS
- . . VINOS SALINOS
- . VINOS MANZANILLA
- . VINOS NOBLES
- . VINOS PALO CORTADO
- VINOS SEGÚN EL AZÚCAR
- . VINOS ABOCADOS
- . VINOS DULCES

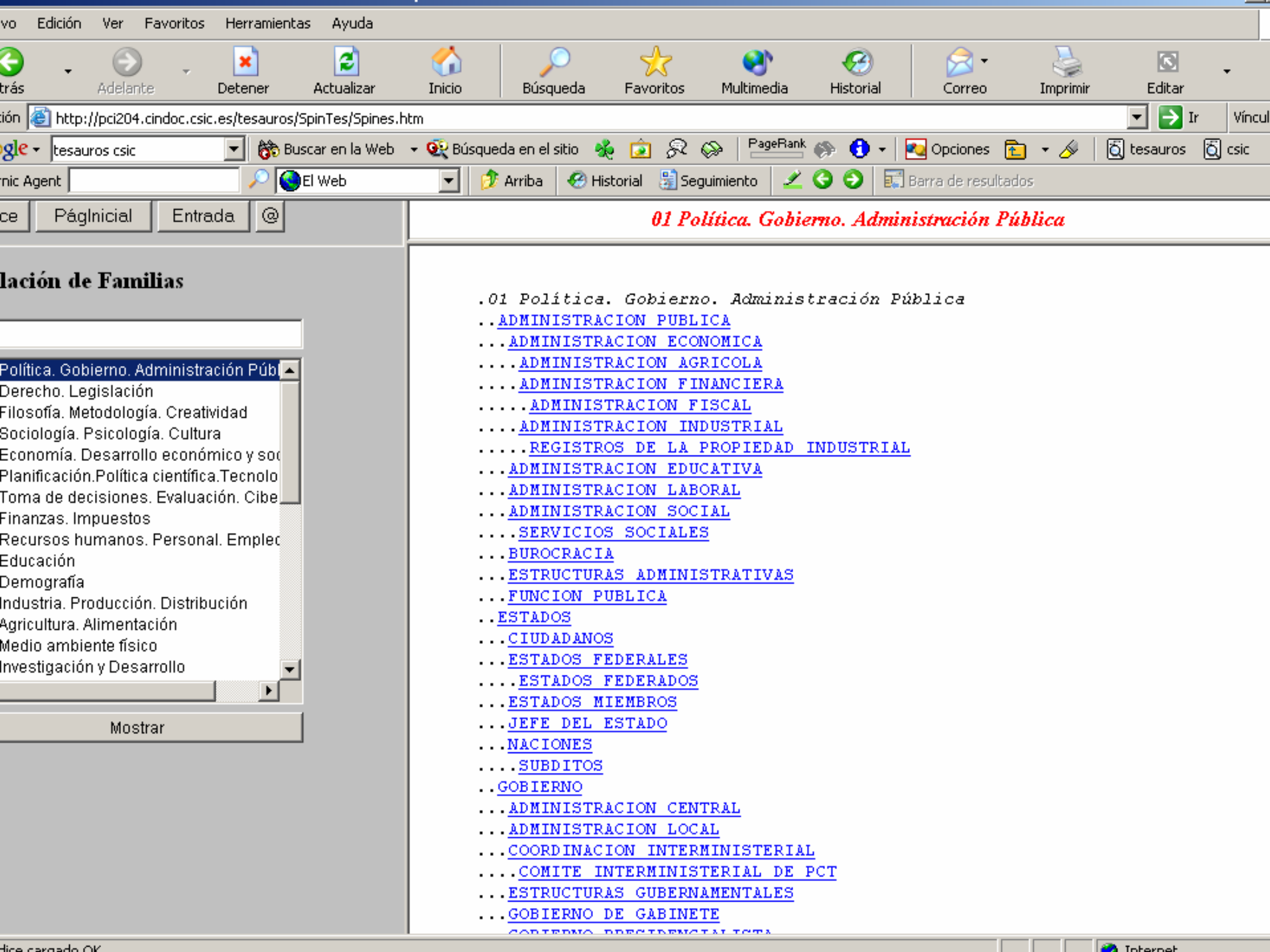
Índice Alfabético

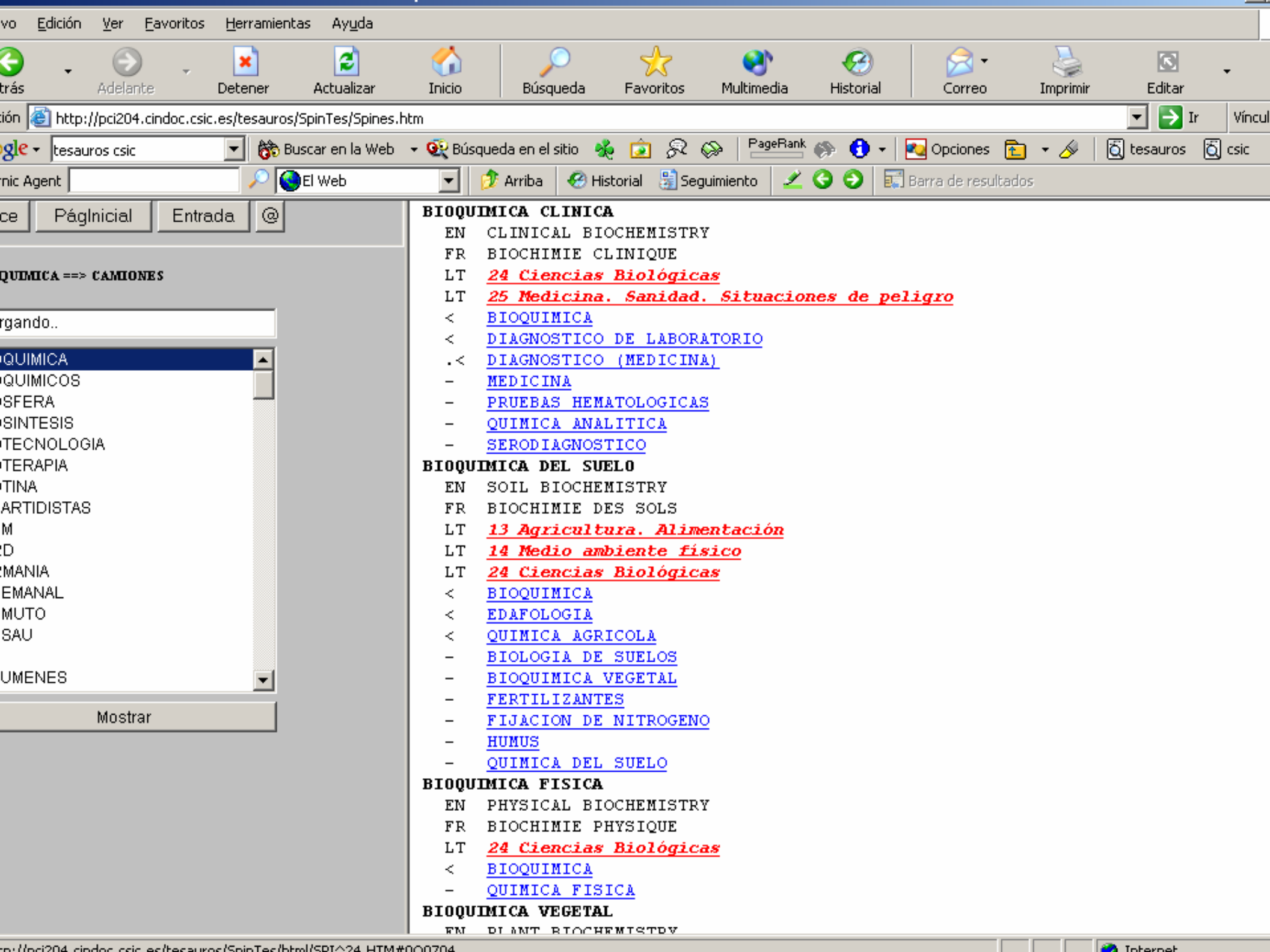
ACIDEZ DEL VINO
- CP: 01 CARACTERÍSTICAS DEL VINO
- TR: FASE GUSTATIVA

AEROBIA
- CP: 04 ENOLOGÍA
- TG: DESCUBE DEL VINO
- TR: AIREADORA
  MERCAPTANO

AIREADORA
- CP: 05 CATA DEL VINO
- TG: INSTRUMENTOS DE SERVICIO DEL VINO
- TR: AEROBIA

ALCOHOL
- CP: 04 ENOLOGÍA
- TG: COMPONENTES DEL VINO
- TR: ALDEHÍDICO
  AROMAS ETÉREOS
  CHAPTALIZACIÓN
  ENCABEZADO DEL VINO
  LÁGRIMA


vo Edición Ver Favoritos Herramientas Ayuda
trás Adelante Detener Actualizar Inicio Búsqueda Favoritos Multimedia Historial Correo Imprimir Editar
ción http://pci204.cindoc.csic.es/tesauros/SpinTes/Spines.htm Ir Vínculos
gle tesauros csic Buscar en la Web Búsqueda en el sitio PageRank Opciones tesauros csic
rnic Agent El Web Arriba Historial Seguimiento Barra de resultados
ce PágInicial Entrada @
lación de Familias
Política. Gobierno. Administración Públ
Derecho. Legislación
Filosofía. Metodología. Creatividad
Sociología. Psicología. Cultura
Economía. Desarrollo económico y soc
Planificación.Política científica.Tecnolo
Toma de decisiones. Evaluación. Cibe
Finanzas. Impuestos
Recursos humanos. Personal. Emplec
Educación
Demografía
Industria. Producción. Distribución
Agricultura. Alimentación
Medio ambiente físico
Investigación y Desarrollo
Mostrar
01 Política. Gobierno. Administración Pública
.01 Política. Gobierno. Administración Pública
..ADMINISTRACION PUBLICA
...ADMINISTRACION ECONOMICA
....ADMINISTRACION AGRICOLA
....ADMINISTRACION FINANCIERA
.....ADMINISTRACION FISCAL
....ADMINISTRACION INDUSTRIAL
.....REGISTROS DE LA PROPIEDAD INDUSTRIAL
...ADMINISTRACION EDUCATIVA
...ADMINISTRACION LABORAL
...ADMINISTRACION SOCIAL
....SERVICIOS SOCIALES
...BUROCRACIA
...ESTRUCTURAS ADMINISTRATIVAS
...FUNCION PUBLICA
..ESTADOS
...CIUDADANOS
...ESTADOS FEDERALES
....ESTADOS FEDERADOS
...ESTADOS MIEMBROS
...JEFE DEL ESTADO
...NACIONES
....SUBDITOS
..GOBIERNO
...ADMINISTRACION CENTRAL
...ADMINISTRACION LOCAL
...COORDINACION INTERMINISTERIAL
....COMITE INTERMINISTERIAL DE PCT
...ESTRUCTURAS GUBERNAMENTALES
...GOBIERNO DE GABINETE
Internet


vo Edición Ver Favoritos Herramientas Ayuda
trás Adelante Detener Actualizar Inicio Búsqueda Favoritos Multimedia Historial Correo Imprimir Editar
ción http://pci204.cindoc.csic.es/tesauros/SpinTes/Spines.htm Ir Vínculos
gle tesauros csic Buscar en la Web Búsqueda en el sitio PageRank Opciones tesauros csic
rnic Agent El Web Arriba Historial Seguimiento Barra de resultados
ce PágInicial Entrada @
QUIMICA ==> CAMIONES
rgando..
QUIMICA
QUIMICOS
SFERA
SINTESIS
TECNOLOGIA
TERAPIA
TINA
ARTIDISTAS
M
D
MANIA
EMANAL
MUTO
SAU
UMENES
Mostrar
BIOQUIMICA CLINICA
EN CLINICAL BIOCHEMISTRY
FR BIOCHIMIE CLINIQUE
LT 24 Ciencias Biológicas
LT 25 Medicina. Sanidad. Situaciones de peligro
< BIOQUIMICA
< DIAGNOSTICO DE LABORATORIO
.< DIAGNOSTICO (MEDICINA)
- MEDICINA
- PRUEBAS HEMATOLOGICAS
- QUIMICA ANALITICA
- SERODIAGNOSTICO
BIOQUIMICA DEL SUELO
EN SOIL BIOCHEMISTRY
FR BIOCHIMIE DES SOLS
LT 13 Agricultura. Alimentación
LT 14 Medio ambiente físico
LT 24 Ciencias Biológicas
< BIOQUIMICA
< EDAFOLOGIA
< QUIMICA AGRICOLA
- BIOLOGIA DE SUELOS
- BIOQUIMICA VEGETAL
- FERTILIZANTES
- FIJACION DE NITROGENO
- HUMUS
- QUIMICA DEL SUELO
BIOQUIMICA FISICA
EN PHYSICAL BIOCHEMISTRY
FR BIOCHIMIE PHYSIQUE
LT 24 Ciencias Biológicas
< BIOQUIMICA
- QUIMICA FISICA
BIOQUIMICA VEGETAL
EN PLANT BIOCHEMISTRY
Internet

Concepto de Faceta

- Concepto de faceta: está directamente ligada a las insuficiencias de las clasificaciones monojerárquicas: rigidez, pesadez y relativa pobreza de las relaciones semánticas.
- Un documento concierne a cinco aspectos de la realidad: personalidad, materia, energía, espacio y tiempo (PMEST).

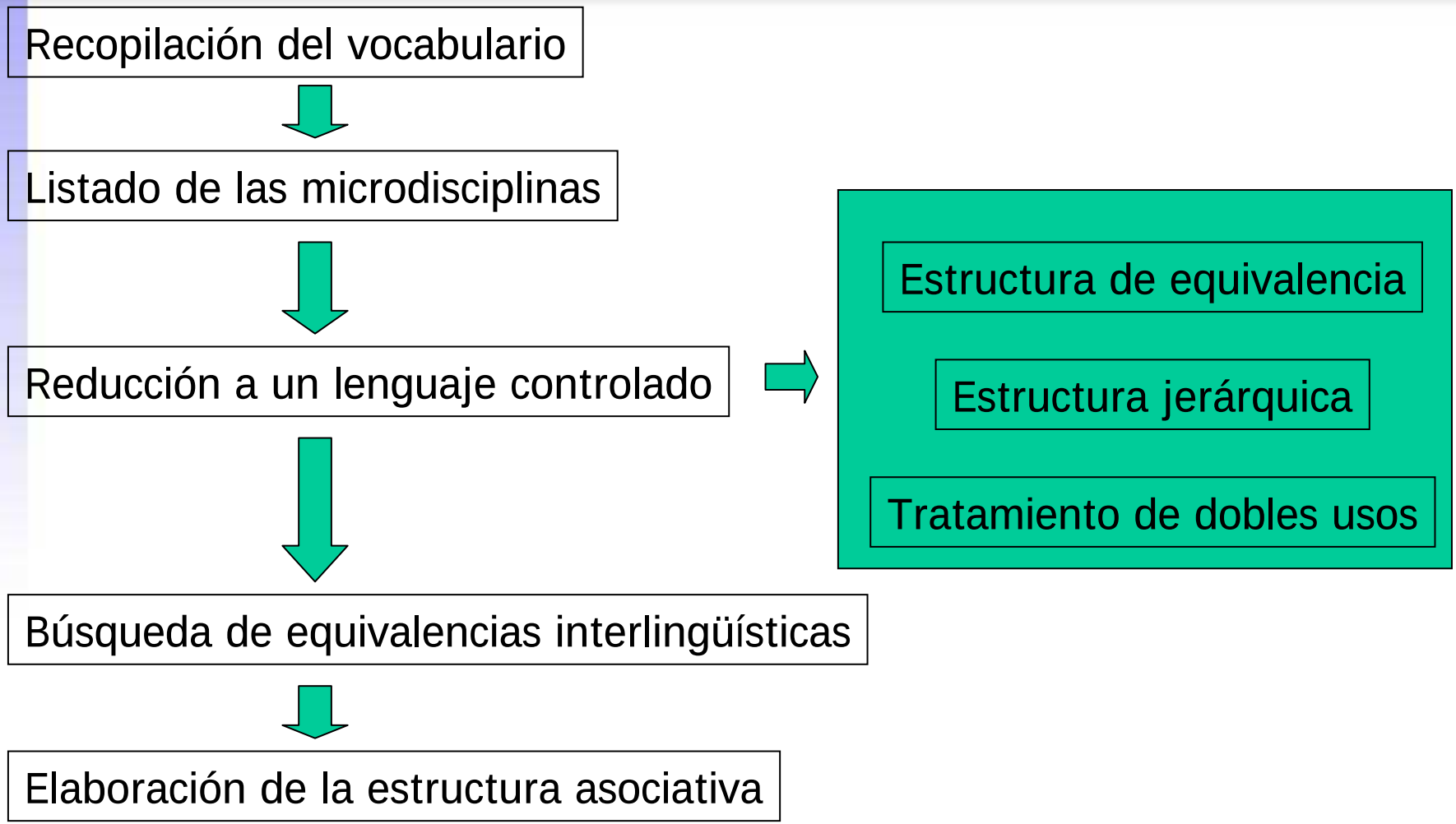
Ejemplos de Faceta

- Robo (energía) de coche (personalidad) a mano armada (materia) el sábado (tiempo) en Vincennes (espacio)
- Análisis gramatical: han robado (verbo) un coche (objeto) a mano armada (comp. De modo) el sábado (comp.. de tiempo) en Vicennes (comp. De lugar)

Facetas

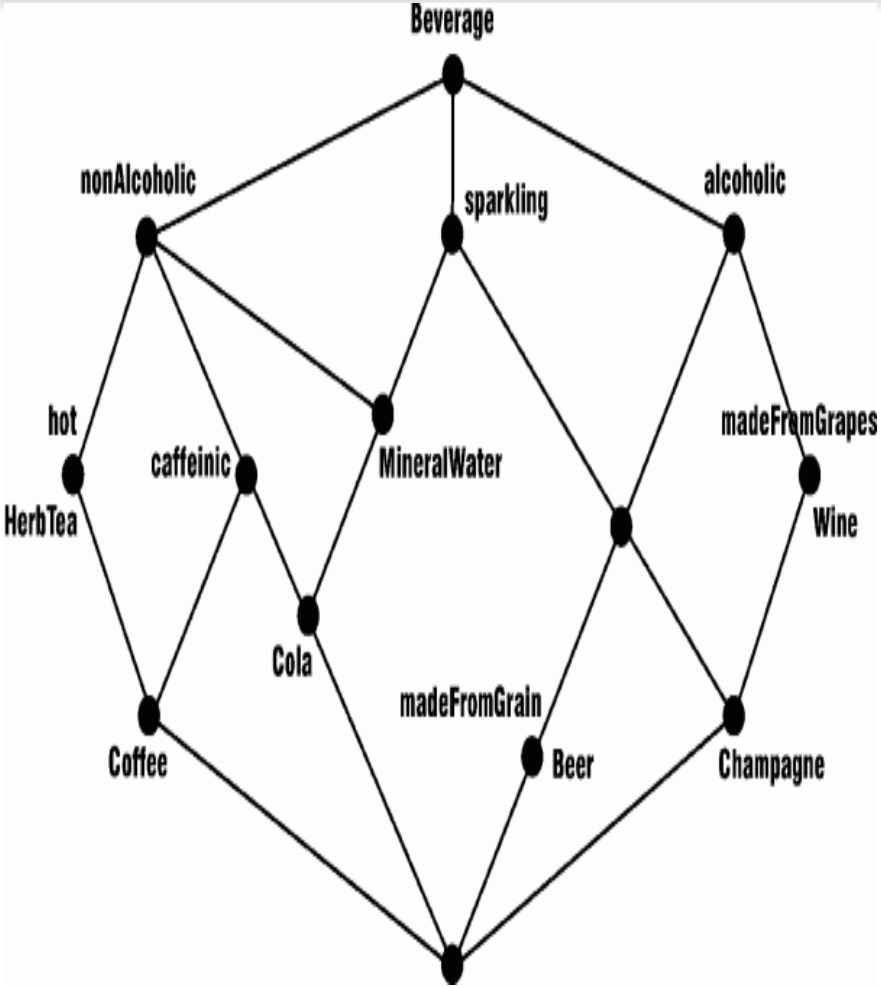
Facetas	Temas			
	Biología	Artes decorativas	Construcción metálica	Transporte
Fenómeno	Digestión	Reflejo	Gravedad	Retraso
Procesos	Manipulación genética	Pintura	Ensamblado	Viaje
Organización	Macromolécula	Museo	Cadena de montaje	Sociedad de transporte
Ser vivo	Bacteria	Decorador	Mecánico	Camionero
Materiales	Proteína	Papel pintado	Herramienta	gas-oil
Equipamiento	Microscopio electrónico	Pincel	Robot de soldadura	Locomotora
Propiedad	Biomagnetismo	Funcionalidad	Flexibilidad	Rapidez
Disciplina	Citología	Estilística	Construcción asistida por ordenador	Estudio del trafico

¿Cómo construir un Tesauro?



Creación automática de Tesauros

- **Objetivo:** Generar una red
- **Área:** Semántica léxica
- **Técnicas:** Relación de oposición, redes de co-ocurrencia léxica, reconocimiento de patrones léxicos
- **Colaboración:** Tesis doctorado en Ciencias de la Información



Beneficios creación automática de Tesauros

- Mayor actualización
- Menor coste, y menor tiempo de construcción
- Reutilización del conocimiento y software
- Mayor consenso
- Buenas perspectivas de importación y exportación de resultados
- Posibles mejoras en la indización y recuperación automática de nuevos documentos.

Dificultades para la creación de Tesauros automáticos

- Ausencia de expertos del dominio
- Indeterminación en la discriminación del dominio
- Tratamiento documental automático
- Tipología del texto, tipología en los tipos de lenguajes

Gestores de Tesauros

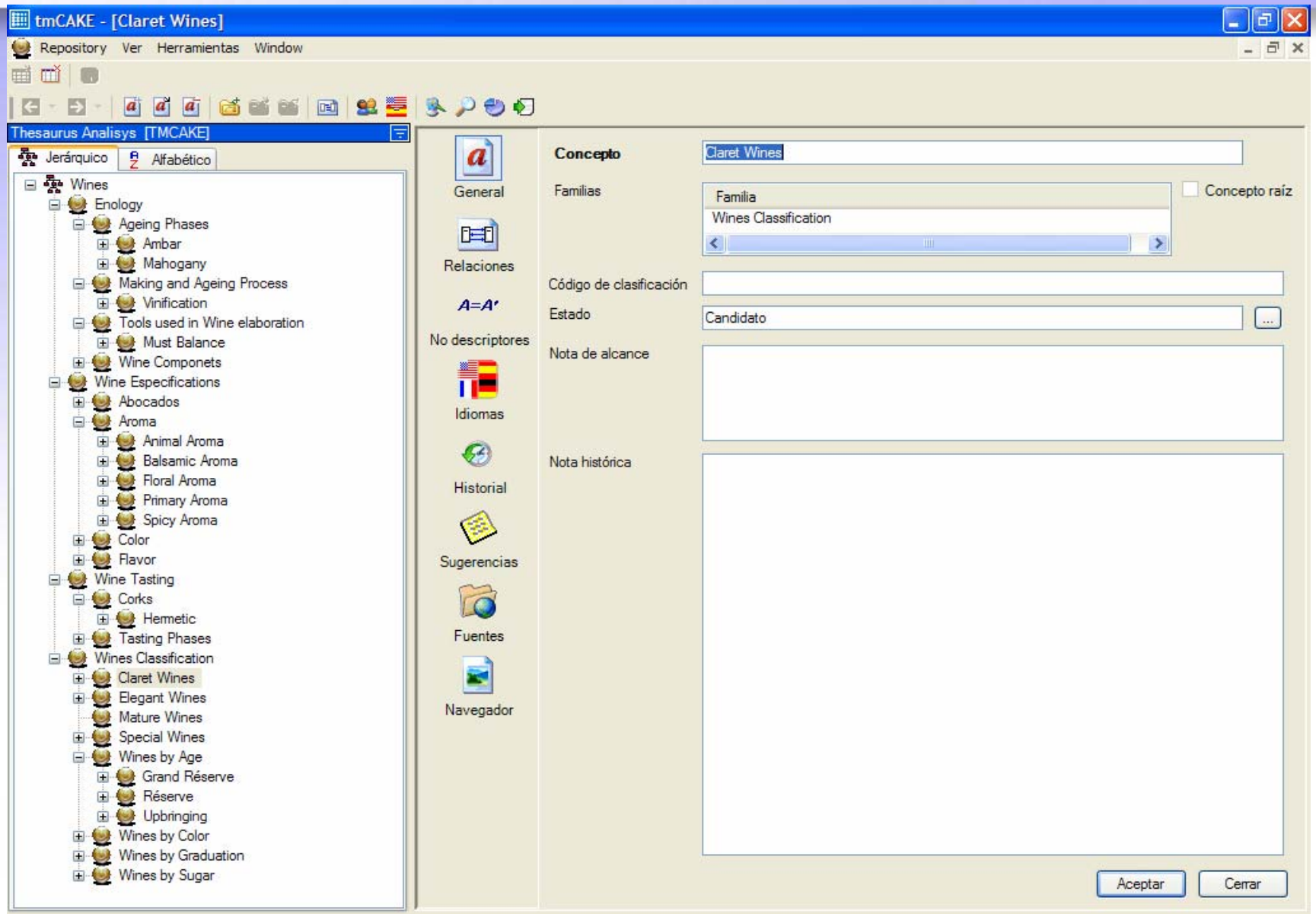
- A.k.a
- B.E.A.T
- CONCEPTO
- INDEX
- LIDOS
- PfleSaurus
- MIDOS
- MultiTes
- TAT
- Term Tree
- ThesMain
- THeW
- THSRS
- Noch mehr
- Thesauruss Software

Thesaurus Management Software

↓

http://www.fbi.fh-koeln.de/institut/labor/bit/thesauri_new/thsoften.htm

Herramienta para generación de Tesauros (tmCAKE / domainREUSER)



Ejemplos tesauros

- **Alcoholismo**
  http://etoh.niaaa.nih.gov/dbtw-wpd/exec/dbtwpub.dll

- **MeSH**
  http://www.nlm.nih.gov/mesh/MBrowser.html

- **Arte**
  http://www.getty.edu/research/tools/vocabulary/aat/index.html

- **cdu agencia isbn**
  http://www.mcu.es/libro/plantilla?id=261&area=libro

- **código unesco mec**
  http://wwwn.mec.es/ciencia/jsp/plantilla.jsp?area=ayudaid&id=41

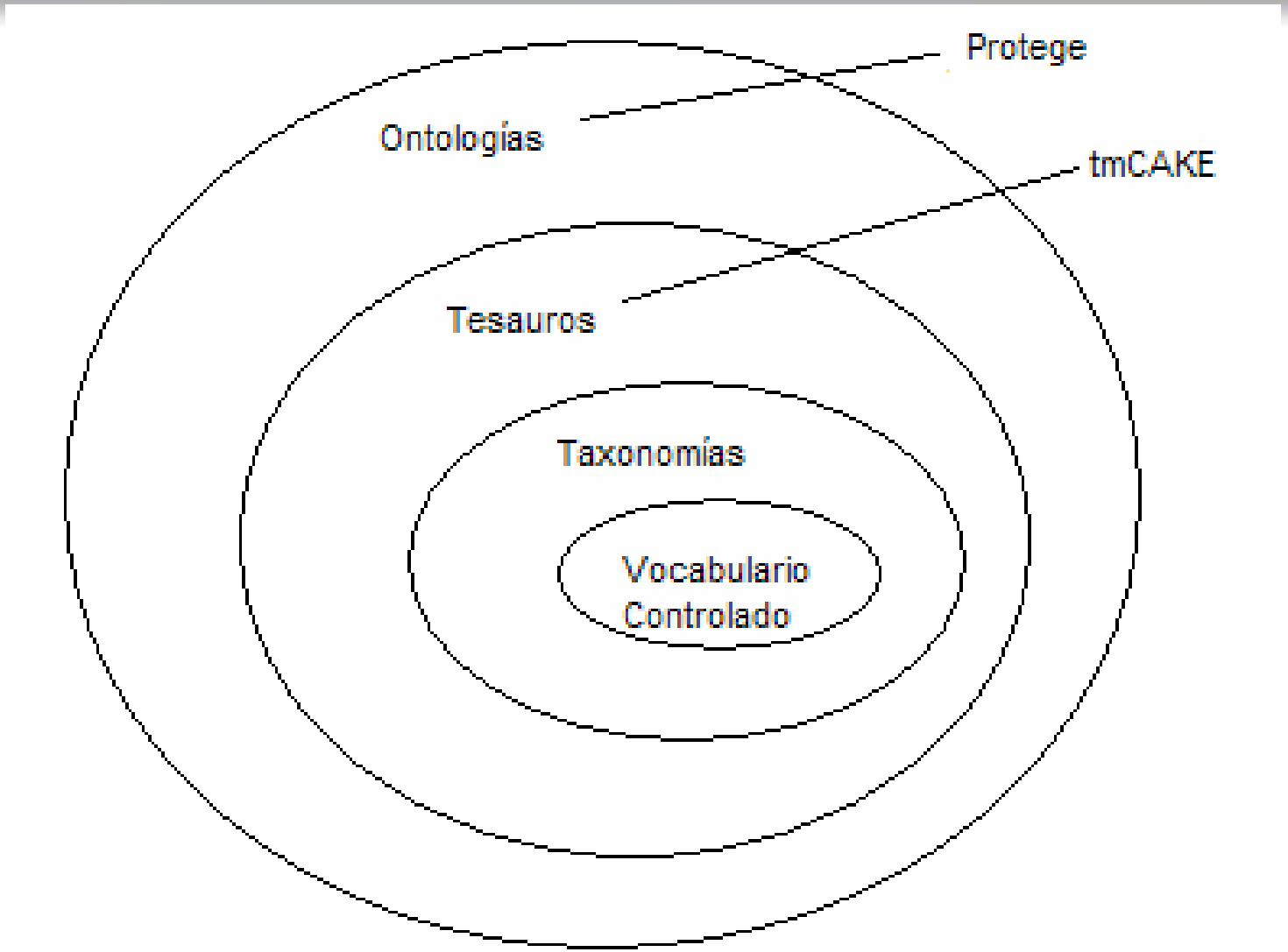
- **visual thesaurus**
  http://www.visualthesaurus.com/

- **wordNet- Web MCR interface Meaning**
  http://nipadio.lsi.upc.edu/cgi-bin/wei4/public/wei.consult.perl

Conclusiones

- Los sistemas de clasificación pueden constituir fuentes externas de conocimiento, para generar nuevas clasificaciones (de forma manual pero preferentemente automáticas)
- Los sistemas de clasificación permiten ordenar y clasificar de forma física y conceptual elementos y conceptos
- Los sistemas de clasificación pueden contribuir a una mejor recuperación de la información

Estructura de conocimiento



Bibliografía

- http://instruct.uwo.ca/677/thesaur/main04.htm (Tim Craven)
- Georges Van Slype. Los lenguajes de Indización. 1987
- ISO 2788: 1986
- NISO Z39:19 Estándar for Structure and Organization of Information Retrieval Thesauri. 1998
- Comparación de tesauros http://willpower.demon.co.uk/thestabl.htm
- Gestores de tesauros Freeware http://publish.uwo.ca/~craven/freeware.htm
- Gestores de tesauros http://www.fbi.fh-koeln.de/institut/labor/Bir/thesauri_new/thsoften.htm
- Tesauros del Cindoc (Centro de Información y Documentación Científica) http://pci204.cindoc.es/tesauros/Index.html

Ejercicio

- Construir un Tesauro de Vinos manualmente, clasificando términos relevantes al dominio.
- Definir el dominio y el propósito del Tesauro.