Tesaurus, llista de descriptors lliures i indexació automàtica

Manela Juncà Campdepadrós

PID_00193277

Universitat Oberta de Catalunya

Índex

Introducció

Aquest mòdul tracta dels tesaurus, les llistes de descriptors lliures i les llistes de paraules clau o indexació automàtica.

Els tesaurus són un llenguatge documental que ha sabut unir els avantatges de tots els llenguatges anteriors: és una classificació en la seva presentació jeràrquica, és una llista de relacions semàntiques en la seva presentació alfabètica i es pot recuperar per paraules clau en els seus índexs permutats.

En el primer apartat s'expliquen tres operacions amb aquest llenguatge: com s'indexa, com es crea un tesaurus nou i com es recupera. L'activitat final consisteix a crear un tesaurus en tres presentacions diferents.

L'apartat sobre els descriptors lliures tracta especialment de la seva aplicació al Web, on ha representat una revolució el fet de convertir cada internauta en autor, editor i documentalista alhora.

Així mateix, tracta de les etiquetes o *tags*, els seus inicis i tipus. I també de la indexació social o *tagging* i dels factors que els han fet adients per al Web, de les motivacions de l'indexador (egoista, amiguista, altruista o populista) i del resultat final de tot plegat conegut com a *folksonomia* o *classificació feta pel poble*.

La recuperació amb aquest llenguatge planteja sospesar els avantatges que proporciona una gran comunitat de persones indexant davant els desavantatges de la manca de control del vocabulari.

Finalment, el darrer apartat tracta sobre la indexació automàtica, explica com funciona un programa d'aquest tipus i quines opcions s'usen en l'actualitat: decidir quina part del text s'indexa automàticament, mantenir o eliminar signes de puntuació i nombres, què s'ha de fer amb les paraules buides (eliminar-les des del començament, contextualitzar-les, mantenir-les per a fer operacions més endavant), aplicació de mètodes estadístics, lingüístics o semàntics.

La recuperació amb aquest llenguatge passa ineludiblement pels cercadors del Web, ja que és la base dels seus robots. L'apartat s'acaba amb els canvis que s'espera que comportarà el Web semàntic sobre això, ja que passarem a recuperar en un web estructurat.

Objectius

Els objectius que ha d'assolir l'estudiant amb aquest mòdul didàctic són els següents:

1. Indexar amb tesaurus de manera específica.

2. Construir un tesaurus a mida.

3. Conèixer el procés de recuperació amb tesaurus i saber usar les referències semàntiques del llenguatge.

4. Identificar les llistes de descriptors lliures al Web: marcadors socials, webs per a compartir imatges i vídeos, etc.

5. Aprendre què són les *etiquetes*, la *indexació social* i les *folksonomies*.

6. Ser conscient dels avantatges i inconvenients en la recuperació per etiquetes.

7. Aprendre quines opcions d'indexació automàtica hi ha en l'actualitat i en quines línies de treball s'està investigant.

8. Conèixer el paper dels llenguatges documentals en la recuperació amb cercadors generals i en un web estructurat amb metadades i ontologies.

9. Adquirir prou elements de judici i coneixement per poder estar al corrent de les noves investigacions que vagin sorgint en l'entorn dels llenguatges documentals i del Web semàntic.

1. Indexació i recuperació amb tesaurus

Indexar amb un tesaurus, igual que amb tots els llenguatge documentals postcoordinats, és molt senzill. Són llenguatges en què no hi ha sintaxi; per tant, la dificultat no està en la composició, l'ordre i la sintaxi del terme d'indexació, sinó en la selecció dels descriptors.

Tesaurus

Els tesaurus són llenguatges naturals, controlats, postcoordinats, jeràrquics i alfabètics i que indexen per conceptes.

1.1. Tesaurus al Web

Hi ha un gran nombre de tesaurus en línia i gratuïts a la Xarxa. Trobem tesaurus d'agricultura, astronomia, biblioteconomia, biologia, art, etc. A continuació, n'oferim una selecció classificada per temes.

Llista de tesaurus en línia

Temàtica	Nom del tesaurus
Agricultura	AGROVOC
Astronomia	The Astronomy Thesaurus
Biblioteconomia	IEDCYT - Tesauro de Biblioteconomía y Documentación DOCUTES Universitat de León
Biologia	IEDCYT - Tesauro de Biología Animal
Ciència	IEDCYT - Tesauro SNIPES
Demografia	Population Multilingual Thesaurus
Economia	EUROVOC Thesaurus IEDCYT - Tesauro ISOC de Economía
Educació	EUROVOC Thesaurus
Empresa	EUROVOC Thesaurus IEDCYT - Tesauro de Propiedad Industrial
Geografia	EUROVOC Thesaurus Getty Thesaurus of Geographic Names IEDCYT - Tesauro de Topónimos
Geologia	IEDCYT - Tesauro de Geología
Història	IEDCYT - Tesauro de Historia Contemporánea de España Història de Catalunya
Llengua i literatura	Traces. Base de dades de llengua i literatura catalanes - Tesaurus
Matemàtiques	BUCM Tesamat Biblioteca Complutense
Propietat industrial	CSIC - Tesauro de Propiedad Industrial
Psicologia	IEDYCT - Tesauro ISOC de Psicología

La majoria dels tesaurus són especialitzats, però alguns són genèrics com l'EUROVOC o els darrers de la llista.

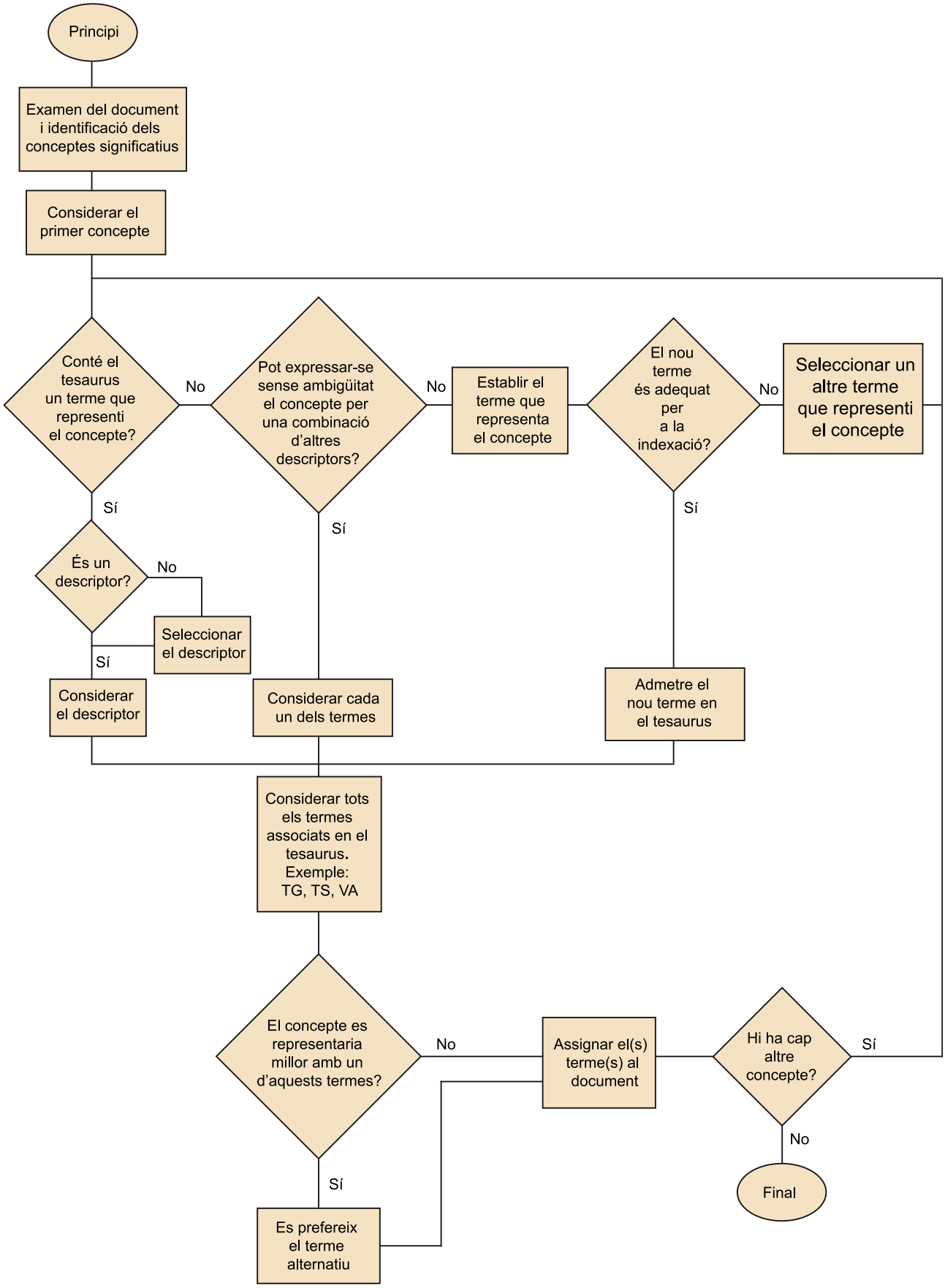
Temàtica	Nom del tesaurus
Sociologia	EUROVOC Thesaurus IEDCYT - Tesauro de Sociología
Topònims	CSIC - Tesauro de Topónimos
Urbanisme	IEDCYT - Tesauro de Urbanismo
Genèrics	UNESCO Història de Catalunya Microtesaurus temàtics de la UB SPINES del IEDCYT ERIC

La majoria dels tesaurus són especialitzats, però alguns són genèrics com l'EUROVOC o els darrers de la llista.

1.2. Com s'indexa amb un tesaurus?

El procés per a indexar amb tesaurus i per extensió amb qualsevol llenguatge documental postcoordinat el trobem gràficament explicat en la norma UNE-50-121-91, annex A, pàg. 7.

Descripció del procés d'indexació amb llenguatges postcoordinats



L'indexador examinarà el document i n'extraurà conceptes, que després traduirà a descriptors del tesaurus. Primer se cerca en la presentació alfabètica i després es comprova en la presentació jeràrquica (aquesta segona consulta ajuda a visualitzar la posició del descriptor en tot l'arbre). Els descriptors que li interessen poden ser en diverses microdisciplines i en diferents nivells de sagnia.

Exemple de descriptors en diferents microdisciplines

Document	Indexació
Keefer, Alice. "Los repositorios digitales universitarios y los autores" [en línia]. *Anales de Documentación*. No 10 (2007) pag. 205-214 Disponible a http://revistas.um.es/analesdoc/article/viewFile/1151/1201	Biblioteques universitàries Fonts d'informació Documents electrònics Universitats Documentació Bases de dades

Hem indexat amb el Tesaurus d'Història de Catalunya (http://sdhlc.uab.cat/Tesaurus.htm). Els tres primers descriptors són de la microdisciplina [Documentació i informació]. El que diu Universitats és d'[Educació]. Els dos últims són de [Ciència i Tecnologia].

Exemple de descriptors en diferents nivells de sagnia

Document	Indexació
Programa electoral presentat per Convergència i Unió de Sant Andreu de Llavaneres a les eleccions municipals de 2007 i que també conté la llista de candidats d'aquest partit.	Partits polítics Programa electoral Eleccions municipals 2007 Candidatures electorals Convergència i Unió (provinent de LENOTI) Sant Andreu de Llavaneres (provinent de la GEC)

En aquesta ocasió hem necessitat només una microdisciplina, la de política, perquè el document no fa referència a altres temes.

Noms propis i geogràfics

Recordem que els noms propis i el geogràfics no es troben en el tesaurus, sinó que provenen de llistes d'autoritats com els de l'exemple (LENOTI i Gran Enciclopèdia Catalana).

1 [POLÍTICA]

[Acció política]
. Vida política
.. Oposició política
... Atemptats
... Bullangues NA: [1835-1842]
... Conspiracions NA: useu-lo acompanyat de la data de la conspiració.
... Guerrilles
.... Guerrillers
.... Maquis
... Insurrecció
.... Insurrecció federal NA: useu-lo acompanyat de la data.
... Oposició parlamentària
... Rebel·lions militars
.... Cop d'estat
.... Pronunciament
... Terrorisme

.. Partits i grups polítics NA: no l'useu com a descriptor.
... Associacions polítiques
... Partits polítics

.. Repressió política
... Depuracions polítiques
... Exili
... Persecucions polítiques
... Presos polítics

.. Sistema electoral
... Eleccions
.... Abstencionisme electoral
.... Campanyes electorals
.... Candidatures electorals
.... Comportament electoral
.... Eleccions autonòmiques NA: useu-lo acompanyat de la data de celebració.
.... Eleccions generals NA: useu-lo acompanyat de la data de celebració.
.... Eleccions municipals NA: useu-lo acompanyat de la data de celebració.
.... Eleccions Parlament Europeu NA: useu-lo acompanyat de la data de celebració.
.... Eleccions provincials NA: useu-lo acompanyat de la data de celebració.
.... Frau electoral
.... Programa electoral
.... Referèndum NA: useu-lo acompanyat de la data de celebració.
... Insaculació
... Sufragi universal

2

En primer lloc, convé fixar-se que els descriptors seleccionats formen part de cadenes jeràrquiques diferents. Un error seria indexar *Eleccions* perquè és el terme ampli (TA) de *Candidatures electorals*, *Eleccions municipals*, *Programa electoral*. No podem indexar el descriptor (o terme) específic (TE) i el seu TA alhora.

Reflexió

Aquesta és l'única regla que necessitem saber per a indexar amb tesaurus: no indexar el TA i el TE a la vegada.

En segon lloc, convé fixar-se que cal ajustar l'enunciat al descriptor aprovat i admès en el tesaurus: *llista de candidats* per *Candidatures electorals*.

En el procés de manteniment d'un tesaurus és possible que conceptes no recollits en un primer moment s'hi acabin afegint posteriorment, però això és una tasca que correspon a l'administrador del tesaurus i no al documentalista; en tot cas, el documentalista pot proposar la necessitat d'un descriptor nou en un camp que es diu *Descriptors candidats*.

1.3. Creació d'un tesaurus

Els tesaurus tenen les presentacions bàsiques de tot llenguatge documental: la jeràrquica, l'alfabètica, la gràfica i la permutada.

Recordem que les fases de construcció d'un tesaurus són vuit en els monolingües i nou en els multilingües.

1) Recollida del vocabulari en llenguatge natural dins el domini que inclourà el tesaurus.

2) Subdivisió del conjunt dels dominis que es tindran en compte en una sèrie de microdisciplines.

3) Transformació del vocabulari lliure en un llenguatge controlat, establiment de les relacions de pertinença, d'equivalència semàntica, de jerarquia i redacció de les notes explicatives.

4) Cerca de les equivalències interlingüístiques (si es tracta d'un tesaurus multilingüe).

5) Enriquiment del tesaurus per mitjà de relacions associatives.

6) Elaboració de l'esborrany del tesaurus.

7) Formació dels indicadors.

8) Test del tesaurus.

9) Revisió final i primera edició.

Els descriptors de cada microdisciplina poden estar ordenats de tres maneres diferents:

- Cronològicament
- Alfabèticament
- Segons el procés

Els dos primers criteris són clars, el tercer es refereix a processos que ja tenen un ordre lògic intern com, en l'exemple, l'ordre dels estudis: primer preescolar, després primària, secundària i superior.

Reflexió

Si sabem construir un tesaurus, sabem construir tots els llenguatges documentals. A més, en ser especialitzat, és el llenguatge perfecte per a fer-nos-el a mida de les nostres necessitats. Per tots aquests motius, doncs, és convenient saber construir un tesaurus.

Lectures recomanades

Per a més informació sobre el procés i les fases, recomanem les lectures següents: Aitchison (1987), Lancaster (2002), Slype, van G. (1991) i les normes UNE 50-106 (ISO 2788-1986) i UNE-50-125 (ISO 5964-1985).

Tres tipus d'ordenacions

Cronològicament	Alfabèticament	Segons el procés
[Història contemporània]	. Medi ambient	[Nivells d'ensenyament]
.Edat contemporània	.. Residus	.. Preescolar
.. Crisi de l'Antic Règim NA: [1808-1833]	... Aigües residuals	.. Ensenyament primari
… Guerra del Francès NA: [1808-1814]	... Residus industrials	.. Ensenyament secundari
….Batalles del Bruc NA: [1808]		.. Ensenyament superior

Finalment, apuntem que les facetes d'un tesaurus es poden ordenar segons la conveniència dels constructors amb la finalitat que siguin més entenedores, com, per exemple, aquestes facetes de la microdisicplina d'[ECONOMIA], en què veiem que *Economia general* precedeix la resta.

- [Història econòmica]
- [Economia general]
- [Economia agrària]
- [Economia pesquera]
- [Economia industrial]
- [Comerç]
- [Hoteleria i turisme]
- [Finances]
- [Economia de l'empresa]

1.4. Recuperació amb tesaurus

La recuperació amb un llenguatge analític i postcoordinat com els tesaurus és més senzilla que la de llenguatges precoordinats, perquè no hi ha sintaxi i s'hi poden afegir tants descriptors com es consideri oportú.

Igual que en la indexació, és molt important que l'indexador conegui fil per randa el tesaurus que indexa la base de dades, les microdisciplines i l'abast conceptual de cadascuna. I també que conegui les llistes d'autoritats del seu SID, tant per noms geogràfics, com personals, títols o entitats.

1.4.1. Procés de cerca amb un tesaurus

El procés de cerca amb tesaurus té tres parts:

- Recollida de conceptes
- Traducció al llenguatge

- Formulació de la cerca

Exemplificarem una cerca a la base de dades ISOC – Biblioteconomia i documentació, a partir del tesaurus de Biblioteconomia de l'IEDCYT (IEDCYT – Tesauro de Biblioteconomía y Documentación).

Recollida de conceptes

El tesaurus és un llenguatge documental analític i, com a tal, permet demanar tants descriptors com calgui. És important que la demanda d'informació es formuli de manera exhaustiva a fi de recollir tots els conceptes interessants per a l'usuari i que podem trobar idèntics o no en el tesaurus.

L'usuari demana documentació sobre documents d'arxiu d'oficina a l'empresa i el documentalista acota la demanda als descriptors que coneix del seu tesaurus.

Quin tipus d'empresa, pública o privada? De quin sector? Documents comptables? Normatives? Com classificar-los? Política d'esporgada? De quins anys? Tot tipus de documentals, tots o només un segment? Etc.

Traducció al llenguatge

Un cop el documentalista tingui els conceptes, la segona tasca és localitzar-los al tesaurus per a traduir-los. Aquí el documentalista jugarà amb les tres presentacions bàsiques de tot tesaurus: l'alfabètica, la jeràrquica i la permutada.

El documentalista es pot trobar en dues situacions: troba el concepte expressat més o menys de la manera que pensava o bé no el troba.

1) Per a localitzar el descriptor, cal consultar la **presentació alfabètica** del tesaurus. En un primer moment es consulta aquesta presentació i no la jeràrquica pels motius següents:

a) Perquè la presentació alfabètica té les relacions d'equivalència entre el no-descriptor i el descriptor acceptat.

En l'expressió de l'usuari era *Arxius d'oficina* que és un no-descriptor que remet a *Archivos de gestión*.

b) Per a comprovar com s'escriu el descriptor, és a dir, quina és la forma acceptada.

En l'expressió de l'usuari era *Arxius d'oficina a l'empresa* i en el tesaurus el concepte es formalitza en *Archivos de empresas*; *Archivos de gestión*.

c) Perquè el documentalista no sap a quina microdisciplina o faceta pertany el descriptor.

Archivos de empresas i *Archivos de gestión* no pertanyen a [*Archivística*] sinó a la microdisciplina de [*Unidades de información*].

d) Si el busqués per la sistemàtica hauria de fullejar el tesaurus sencer per localitzar-lo; en canvi, amb l'alfabètica els trobarà a la primera.

Si el documentalista no troba el descriptor, llavors li serà més útil la presentació jeràrquica i la permutada.

2) Consultar la **presentació jeràrquica**. La seva utilitat rau en el fet que l'arborescència li pot suggerir descriptors paral·lels, genèrics i específics. En posarem un exemple de cada.

Exemple de termes paral·lels

El documentalista busca algun concepte que expressi la cadena documental a *arxius*. No hi és en el tesaurus i tampoc no és un no-descriptor. Encara que no hi sigui, s'adona que totes les fases de la cadena es troben sistematitzades sota el descriptor *Proceso documental*. En una segona opció podria obrir el descriptor en termes més específics i cercar per fases i subfases concretes de la cadena; per exemple, *Adquisiciones*; *Análisis de contenido*.

Exemple de termes genèrics

L'usuari ha demanat pel concepte *unitermes* que no és al tesaurus i tampoc no hi ha cap altre terme que pugui usar. En aquest cas, seleccionaria el descriptor immediatament superior conceptualment a altres descriptors paral·lels, és a dir, si uniterme és al mateix nivell que *descriptor* i que *paraula clau*, escolliria *Términos*, que engloba tots els tipus de termes d'indexació. Un altre cas es dóna quan el documentalista troba el descriptor correcte; per exemple, *Reglamentos de archivos*, però la base de dades li retorna zero resultats, per la qual cosa decideix consultar la jeràrquica i reformular la cerca aquesta vegada amb el terme genèric de *Reglamentos de archivos*, que és *Política archivística*.

Exemple de termes específics

L'usuari ha demanat pel tema *llenguatges documentals*. El tesaurus recull aquest concepte com a descriptor, però el documentalista consultant la presentació jeràrquica veu que també pot cercar pels termes específics que són en aquest tesaurus:

TE Clasificaciones

TE Lenguajes de indización

Observació

Recordem que el documentalista no haurà indexat amb el TA i el TE a la vegada. Per tant, un manual general sobre llenguatges documentals estarà indexat com a *Lenguajes documentales* i no amb el descriptor de cada llenguatge concret.

3) Consultar els **índexs permutats**. Els índexs permutats (KWIC o KWOC) permeten localitzar altres descriptors que continguin la paraula clau que cerquem **en qualsevol posició del descriptor**.

Si busquem *archivos*, a més de la lletra *A* de *archivos*, si consultem l'índex KWIC podem trobar:

Automatización de archivos

Historia de los archivos

Sistemas nacionales de archivos

...

Formulació de la cerca

Finalment formularà la cerca distribuint els conceptes en els camps de la base de dades (per matèria, abast cronològic, format, idioma, etc.) fent ús d'operadors booleans si cal.

1.5. Activitats

Seguidament us proposem un seguit d'activitats perquè pugueu posar en pràctica els tesaurus.

1.5.1. Indexació del contingut d'articles

A l'aula, a la secció de Recursos i fonts, hi trobareu una bateria d'articles a text complet de temes molt diferents. Seguiu les instruccions de l'aula per a saber quin article i quin tesaurus heu d'usar.

1.5.2. Construcció manual i automàtica de tesaurus

Creació d'un petit tesaurus sobre anàlisi de contingut a mà i en segon terme un programa de programari lliure amb les característiques següents:

- Dues microdisciplines [Cadena documental], [Indexació]
- Tres presentacions: jeràrquica, alfabètica i permutada KWIC
- Tres tipus d'ordenacions: històrica, alfabètica i procedimental
- Es faciliten els descriptors ordenats per microdisciplines

La presentació jeràrquica com a base dels tesaurus

La presentació jeràrquica és la base d'un tesaurus, a partir de la qual neixen les altres presentacions alfabètiques, gràfiques i permutades.

A l'hora de crear el tesaurus cal recordar que les relacions es posen en un ordre predeterminat i que les sigles que hem d'usar són les que hi ha recollides en aquesta taula.

Taula resum de sigles en català, castellà i anglès

		Català	Castellà	Anglès
Domini (no és obligatori)		DOM	DOM	DOM
Notes d'aclariment		NA/NE (aclariment/explicativa)	NA	SC (*scope note*)
Equivalència	**Empreu**	EM	USE	USE
	Emprat per	EP	UP	UF
Terme capçalera (no és obligatori)		TC	TC	TT (*top term*)
Jerarquia	**Terme genèric**	TA	TG	BT (*broad term*)
	Terme específic	TE	TE	NT (*narrow term*)
Relació associativa		TR	TR	RT

Descriptors de la microdisciplina [Cadena documental]. Nota: no s'obre en facetes. Aquests descriptors ja estan controlats en la forma.

- Accés directe al fons
- Accés lliure
- Adquisició
- Anàlisi de contingut
- Anàlisi documental
- Anàlisi formal
- Bases de dades
- Catàlegs
- Causes de degradació externes
- Causes de degradació internes
- Cercadors
- Compra
- Descripció bibliogràfica
- Dipòsit legal
- Directoris
- Donació
- Emmagatzematge i conservació
- Fase d'anàlisi i tractament
- Fase de sortida
- Fase d'entrada
- Formació d'usuaris
- Guies butlletins
- Indexació
- Instruments de cerca
- Intercanvi
- Inventaris
- Ordenació
- Ordenació altament significativa
- Ordenació amb significat limitat
- Ordenació no significativa
- Política de selecció
- Portals
- Preparació del material
- Préstec
- Processament tècnic
- Recepció
- Registre
- Reprografia
- Resum
- Resum automàtic
- Resum indicatiu
- Resum informatiu
- Resum selectiu
- Selecció
- Serveis de difusió
- Serveis de referència

- Transferència

Descriptors de la microdisciplina [Indexació]. Nota: s'obre en tres facetes: *[Evolució històrica], [Llenguatges documentals]* i *[Llenguatges naturals].* Us donem els descriptors ja classificats en les tres facetes i controlats en la forma.

[Evolució històrica]

- Bilindex [1983]
- Guía para los encabezamientos de materia [1934]
- Indexació automàtica [1957]
- Library of Congress subject headings [1909]
- List of subject headings for small libraries [1923]
- List of subject headings for use in dictionary catalogs [1895]
- Lista de encabezamientos de materia para bibliotecas [1967]
- Répertoire d'autorité-matière encyclopedique et alphabetique unifié RAMEAU [1980]
- Répertoire de vedettes-matière RVM [1946]
- Rules for a dictionary catalog [1876]
- Segle XIX
- Segle XX (1900-1950)
- Segle XX (1950-1999)

[Llenguatges documentals]

- Autoritats
- Descriptor
- Descriptors controlats
- Descriptors lliures
- Encapçalaments de matèria
- Llenguatges codificats
- Llenguatges controlats
- Llenguatge de descriptors
- Llenguatges de paraules clau
- Llenguatges documentals
- Llenguatge lliure
- Llenguatges de matèria
- Llenguatges de postcoordinació
- Llenguatges precoordinats
- Llista d'autoritats
- Llista d'encapçalaments de matèria
- Llista de descriptors lliures
- Llista de paraules clau
- Notacions
- Paraules clau
- Segons el nivell d'anàlisi
- Segons el nivell de control

- Segons el nivell de coordinació
- Segons la naturalesa dels termes
- Sistema de classificació
- Termes d'indexació
- Tesaurus
- Tipologies de llenguatges documentals

[Llenguatges naturals]

- Ambigüitat del llenguatge natural
- Homofonia
- Homografia
- Homonímia
- Polisèmia
- Sinonímia

1.5.3. Recuperació amb tesaurus

Atén les demandes d'usuari següents a la base de dades ISOC de Biblioteconomia i Documentació, indexada amb el tesaurus de Biblioteconomia i Documentació:

- Informació sobre la indexació i recuperació amb tesaurus en centres de documentació.
- Informació sobre tractament de la documentació dels museus militars i l'atenció als usuaris.
- Els dispositius de biblioteques mòbils, els tipus de fonts que porten i mesures de seguretat.
- Opcions laborals per a bibliotecaris i arxivers.
- Indexació automàtica i llei de Zipf.

1.6. Solució

Seguidament recollim les solucions dels exercicis que us hem plantejat més amunt.

1.6.1. Indexació d'articles

La solució es treballarà a l'aula.

1.6.2. Construcció manual i automàtica de tesaurus

Aquesta és la presentació jeràrquica dels descriptors proposats:

[Cadena documental]
. Fase d'entrada
.. Selecció
... Política de selecció
.. Adquisició
... Accés lliure
... Compra
... Intercanvi
... Donació
... Dipòsit legal
... Transferència
.. Recepció
... Preparació del material
... Registre

. Fase d'anàlisi i tractament
.. Anàlisi documental
... Anàlisi formal
.... Descripció bibliogràfica
... Anàlisi de contingut
.... Indexació
.... Resum
..... Resum automàtic
..... Resum indicatiu
..... Resum informatiu
..... Resum selectiu
.. Processament tècnic
... Ordenació
.... Ordenació altament significativa
.... Ordenació amb significat limitat
.... Ordenació no significativa
... Emmagatzematge i conservació
.... Causes de degradació externes
.... Causes de degradació internes

. Fase de sortida
.. Instruments de cerca
... Catàlegs
... Inventaris
... Bases de dades
... Guies butlletins
... Cercadors
... Directoris
... Portals
.. Serveis de difusió
... Accés directe al fons
... Serveis de referència
... Préstec
... Reprografia
... Formació d'usuaris

[Indexació]

[Evolució històrica]
. Segle XIX
.. Rules for a dictionary catalog [1876]
.. List of subject headings for use in dictionary catalogs [1895]
. Segle XX (1900-1950)
.. Library of Congress Subject headings [1909]
.. List of subject headings for small libraries [1923]
.. Guia per a encapçalaments de matèria [1934]
.. Répertoire de vedettes-matière RVM [1946]
. Segle XX (1950-1999)
.. Indexació automàtica [1957]
.. Llista d'encapçalaments de matèria per a biblioteques [1967]
.. Répertoire d'autorité-matière encyclopedique et alphabetique unifié RAMEAU [1980]
.. Bilindex [1983]

[Llenguatges documentals]
. Llenguatges documentals
.. Llista d'encapçalaments de matèria
.. Llista d'autoritats
.. Llista de descriptors lliures
.. Llista de paraules clau
.. Sistema de classificació
.. Tesaurus

. Tipologies de llenguatges documentals
..Segons el nivell d'anàlisi
... Llenguatges de descriptors
... Llenguatges de paraules clau
... Llenguatges de matèria
.. Segons el nivell de control
... Llenguatges controlats
... Llenguatges lliures
.. Segons el nivell de coordinació
... Llenguatges postcoordinats
... Llenguatges precoordinats
..Segons la naturalesa dels termes
... Llenguatges codificats
... Llenguatges naturals
. Termes d'indexació
.. Autoritats
.. Descriptors
... Descriptors controlats
... Descriptors lliures
.. Encapçalaments de matèria
.. Notacions
.. Paraules clau
NA Paraula clau entesa com a uniterme provinent de la indexació automàtica

[Llenguatges naturals]
. Ambigüitat del llenguatge natural NA Per *llenguatge natural* entenem el llenguatge que usem quotidianament: català, castellà, basc, gallec, francès, etc.

.. Polisèmia
... Homonímia
.... Homofonia
.... Homografia
.. Sinonímia

Solució en part a la presentació alfabètica

Presentem cinc exemples corresponents a totes les posicions que pot tenir un descriptor en aquest tesaurus. I totes les sigles que defineixen les relacions semàntiques existents.

- Amb un punt al davant: Fase d'entrada
- Amb dos punts al davant: Selecció
- Amb tres punts al davant: Política de selecció
- Amb quatre punts al davant: Ordenació altament significativa
- Amb cinc punts al davant: Resum automàtic
- D'un no-descriptor al descriptor acceptat: *Extracts*

Observació

Fixeu-vos que els descriptors van en ordre alfabètic.

Extracts	
EM	Resum automàtic
Fase entrada	

TC	Cadena documental
TE	Selecció
	Adquisició
	Recepció
Ordenació altament significativa	
TC	Cadena documental
TA	Ordenació
TR	Llenguatges codificats
	Sistemes de classificació
Política de selecció	
TC	Cadena documental
TA	Selecció
Resum automàtic	
EP	Extracts
TC	Cadena documental
TA	Resum
Selecció	
TC	Cadena documental
TA	Fase d'entrada
TE	Política de selecció

Solució en part a l'índex KWIC

Exemple lletra *L* del KWIC. En aquesta solució podem observar que tots els unitermes dels descriptors que comencin per *L* s'ordenen alfabèticament. Trobarem índexs KWIC que marquen la paraula en qüestió en negreta (com en l'exemple) i d'altres que la situen en una columna central del tipus:

Dipòsit	legal	
	Llenguatge	Lliure
Accés	lliure	

Fixeu-vos que descriptors com *llenguatges lliures* apareixeran dues vegades al KWIC, tant per la *Ll* de *llenguatges* com per la *ll* de *lliures*. Els descriptors que tenen algun article com *les* no s'indexen, ja que es consideren paraules buides (per aquest motiu no apareix, per exemple, Segons la naturalesa dels termes).

Observació

Els índexs permutats actuen com a llistes de paraules clau. Ja que de fet estem descomponent el descriptor en unitats soltes.

Dipòsit **legal**

List of subject headings for small **libraries** [1923]

Library of Congress subject headings [1909]

Ordenació amb significat **limitat**

List of subject headings for small libraries [1923]

List of subject headings for use in dictionary catalogs [1895]

Lista de encabezamientos de materia para bibliotecas [1967]

Ambigüitat del **llenguatge** natural

Llenguatge lliure

Llenguatges codificats

Llenguatges controlats

Llenguatges de descriptors

Llenguatges documentals

Tipologies de **llenguatges** documentals

Llenguatges de matèria

Llenguatges de paraules clau

Llenguatges de postcoordinació

Llenguatges precoordinats

Llista d'autoritats

Llista d'encapçalaments de matèria

Llista de descriptors lliures

Llista de paraules clau

Accés **lliure**

Llenguatge **lliure**

Descriptors **lliures**

2. Indexació amb llistes de descriptors lliures: etiquetes i indexació social

La llista de descriptors lliures és un llenguatge que es crea dinàmicament, en temps real, a mesura que l'indexador va llegint i assignant un terme. Els termes del vocabulari no consten en cap full previ. L'indexador no comprova que el terme existeixi. No comprova com s'escriu. Hi ha plena llibertat.

Llistes de descriptors lliures

Les llistes de descriptors lliures són llenguatges naturals, lliures, postcoordinats, alfabètics i analítics per conceptes.

2.1. Descriptors lliures al Web

Al Web hi ha moltes iniciatives d'indexació amb descriptors lliures, les més meritoses són els marcadors socials (Delicious), webs per a compartir imatges (Tagzania, Flickr, Youtube) i aplicacions del Web 2.0 com blogs (Blogger), xarxes socials i webs (Buzzillions), que recullen l'opinió de consumidors sobre marques de tota mena de productes.

- Delicious (https://www.delicious.com): Diigo (http://www.diigo.com), Mr Wong (http://www.mister-wong.com) són serveis de gestió d'adreces d'interès mitjançant el Web. Permeten guardar i recuperar a la Xarxa les adreces d'interès, que clàssicament s'emmagatzemaven des del navegador localment a l'ordinador, de manera que són consultables en línia i no solament localment.

- Tagzania (http://www.tagzania.com): és un sistema que usa folksonomies sobre l'API del potent Google Maps. És un *mashup* de geolocalització de fotografies similar a Panoramio (http://www.panoramio.com), que ofereix altres funcionalitats de valor afegit als mapes.

- Flickr (http://www.flickr.com): és un lloc web de Yahoo per a organitzar fotografies digitals que funciona com una xarxa social. És un servei molt utilitzat pels usuaris de blogs com a dipòsit de fotos.

- Youtube (http://www.youtube.com): és un lloc web per a compartir vídeos, clips de pel·lícules, clips de televisió, vídeos musicals, i també contingut amateur. Els usuaris no registrats poden veure vídeos, i els usuaris registrats poden pujar un nombre il·limitat de vídeos

- Blogger (https://accounts.google.com): és un servei per a crear i publicar un blog fàcilment.

- Buzzillions (www.buzzillions.com): és un lloc web que recull prop de disset milions de crítiques de productes d'una àmplia gamma de categories (electrònica, moda, salut, etc.). Les recomanacions provenen de persones

reals (no es paguen per les revisions) amb la intenció d'assessorar compradors nous a partir del grau de satisfacció dels productes.

2.2. Etiquetes i indexació social

Cada usuari indexa els descriptors lliures que li semblen millors. Milions d'usuaris indexen els seus descriptors. Entre tots creen un espai d'aportacions sense una intervenció centralitzada ni més autoritat que la que fan els usuaris, no hi ha descriptors predeterminats.

James Surowiecki

James Surowiecki (2004) ho anomena *la saviesa de les masses* (*the wisdom of crowds*).

Aquesta manera d'indexar, no professional i sense llenguatge documental controlat, es coneix com a ***indexació social***. Hi intervenen les etiquetes o *tags*, el *tagging* o acció d'indexar lliurement i les folksonomies o conjunt total de totes les etiquetes assignades pels usuaris.

És una revolució en el món del Web perquè s'ha invertit el paradigma: abans pocs autors escrivien per a molts lectors, i ara molts autors no sols escriuen sinó que també editen i descriuen els seus documents.

Com diu Mari Carmen Marcos (2009):

"cadascú és autor, editor i documentalista alhora".

Terminologia

Trobarem diversos termes per a cada concepte:

- Per als termes d'*indexació*: *descriptors lliures* o *etiquetes* o *tags*. Del conjunt de *tags* se'n diu *núvol de tags*, que seria el més semblant a un llenguatge documental.
- Per a l'acció d'*indexar lliurement*: *tagging* o *etiquetatge social* i, més específicament, *social bookmarking* o *website bookmarking* quan es tracta de descriure els recursos web.
- Per al conjunt de *tags de tots els usuaris*: *folksonomies* o *classificació feta pel poble*.

2.2.1. Etiquetes

Una etiqueta o *tag* és un terme d'indexació que s'afegeix a un objecte digital com un web, un vídeo o una foto, per tal de descriure'l en forma i contingut.

Exemple

Per exemple, enciclopedia_art: enciclopèdia (forma) d'art (contingut). No és un descriptor controlat, és un descriptor lliure.

Les primeres etiquetes van aparèixer als blogs, i proporcionaven enllaços i comentaris sobre recursos, tipus "recomano el web tal per a tal tema". Es considera que van ser les primeres metadades, encara que molt mancades d'estructura. Avui dia, els usuaris indexen amb etiquetes els seus webs preferits, les localitzacions de les fotos, les emocions d'unes imatges, el grau de satisfacció d'un rentaplats, etc.

Les etiquetes són funcionals perquè són les autoritats dels usuaris. Lancaster ja observava l'any 1995 que els termes s'havien d'obtenir dels usuaris potencials i que n'havien de representar els interessos concrets. O més reculat en el temps, Cutter ja postulava que els termes d'indexació havien de representar l'ús comú i posar el focus en el lector.

Les etiquetes poden ser unitermes o descriptors compostos, és a dir, poden estar formades per una sola paraula (tesaurus) o per dues paraules (per exemple, Llenguatges_documentals).

Separació amb guió

Les paraules s'acostumen a separar amb guió perquè l'espai és el signe que marca el final de l'etiqueta.

Ros-Martin (2008) va classificar les etiquetes en aquests grups:

1) Les basades en el contingut temàtic.

Exemple: Capítol_indexació_social

2) Les basades en el context o emmagatzematge.

Exemple: Mòdul3_cap2

3) Les subjectives.

Exemple: Útil

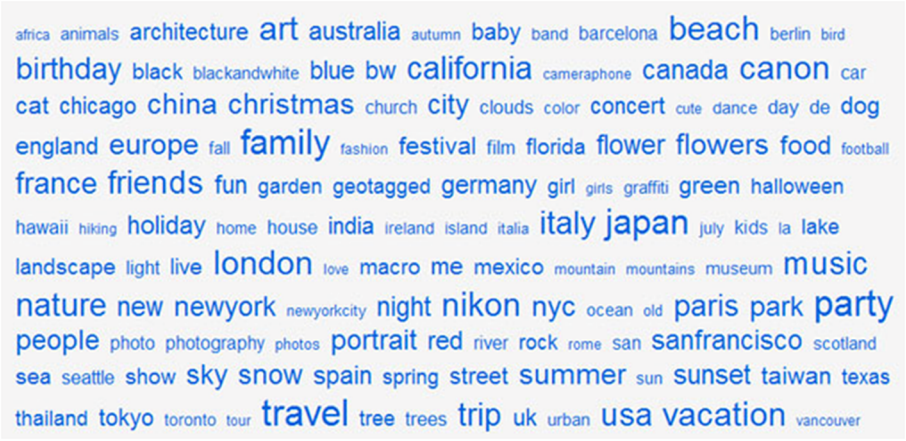
4) Els atributs que no es derivin del contingut.

Exemple: UOC

5) Les d'organització o de recordatori de tasques.

Exemple: Guardar, Relacionar_amb_Recuperació, Per_Joan

El conjunt d'etiquetes es coneix com a **núvol d'etiquetes**. Aquest núvol és un espai pla en què les etiquetes no tenen relacions de parentiu ni de jerarquia entre elles, però que permeten la compartició de categories entre usuaris. Es presenten en ordre alfabètic i destacades amb tipografia més grossa segons la freqüència d'ús.



Font: imatge presa de Flickr.

2.2.2. Indexació social

Els descriptors lliures són el llenguatge ideal per a indexar el Web pels factors següents:

1) Perquè és un llenguatge lliure. El Web no es pot indexar amb els llenguatges controlats, perquè el temps i l'esforç econòmic que se'n derivarien serien inassolibles. Els llenguatge documentals controlats no són adequats en entorns en què les metadades són una opció millor. Les metadades poden ser de diversos tipus: creades per un documentalista, per l'autor del document o per un robot. Amb les etiquetes podem afegir una altra via, la de les metadades creades pels usuaris (Mathes, 2004).

2) Perquè no necessiten formació documentalista prèvia: les característiques d'aquest llenguatge el fan ideal per a qualsevol col·lectiu no professional de la documentació, com els internautes del Web.

3) El grup d'usuaris és tan nombrós que assumeixen quantitats enormes de documents (ja no és un indexador, sinó una comunitat d'indexadors).

4) Permeten indexar documents com ara una imatge o un vídeo que no van acompanyats de text o peus de foto, que fins ara només eren indexables per humans i no per robots.

5) Les etiquetes són properes als usuaris; no són termes escollits per tècnics sinó que són termes intuïtius. La comunitat actua com un sedàs que filtra les paraules realment més útils.

6) Són eficaços individualment –a nivell d'usuari– perquè organitzen la informació personal i, socialment, perquè tota la comunitat virtual es beneficia de la indexació que han fet els altres.

Lectures recomanades

S'han fet diversos estudis sobre la consistència d'indexar amb etiquetes entre indexadors a l'hora d'indexar imatges i fins i tot emocions amb resultats molt bons de coherència entre usuaris (emocions identificades de manera homogènia). Un exemple el teniu a Knautz and Stock (2010) i a Ransom and Rafferty (2011):

Kathrin Knautz; Wolfgang G. Stock (2010). "Collective indexing of emotions in videos". *Journal of Documentation* (vol. 67, núm. 6, pàg. 975-994).

N. Ransom; P. Rafferty (2011). "Facets of user-assigned tags and their effectiveness in image retrieval". *Journal of Documentation* (vol. 67, núm. 6, pàg. 1038-1066).

Els professionals de la informació també usen la indexació social o *tagging* per a indexar els recursos web. S'utilitzen en intranets, sistemes corporatius, bases de dades i biblioteques per a donar valor afegit a les seves bases de dades (per exemple, la base de dades Complured de la Universidad Complutense de Madrid), també per a compartir els marcadors seleccionats amb altres usuaris i per a reutilitzar els continguts en altres aplicacions com xarxes socials tipus Twitter i així donar més visibilitat a la institució.

Organització de les col·leccions a les biblioteques universitàries

La majoria de biblioteques universitàries organitzen les col·leccions de la manera següent:

- **Col·lecció pròpia:** catàleg indexat de manera controlada (sistemes de classificació + llistes d'encapçalaments + llista d'autoritats / tesaurus + llista d'autoritats) i automàtica (llista de paraules clau).

- **Recursos electrònics del Web:** directoris temàtics o guies temàtiques (sistemes de classificació) + Delicious (llista de descriptors lliures o *tags*).

Podeu comprovar que les etiquetes d'un Delicious són descriptors lliures fent la comparació següent: busqueu una llista d'encapçalaments de matèria que s'usi o es creï en una biblioteca, llavors consulteu el Delicious d'aquesta biblioteca.

Per exemple, la Biblioteca de Catalunya, autora de la LEMAC, indexa en el catàleg amb l'encapçalament Art – Història, però Delicious indexa Història de l'art, que és un terme més pròxim a l'usuari.

Només cal consultar les biblioteques d'universitats que imparteixen Informació i Documentació per adonar-se que, a més del catàleg, tenen Delicious.

- Delicious de la Universitat de Barcelona, CRAI (http://www.delicious.com/CRAIUBreferencia)

- Delicious de la Universidad Nacional de Educación a Distancia (UNED) (http://delicious.com/brelreferencia20)

- Delicious de la Universitat Complutense de Madrid (http://delicious.com/bibliotecacps)

Els indexadors tenen diverses motivacions per a fer indexació social, ja que obtenen diversos beneficis socials. Javier Cañada (2006) els va classificar tal com queda recollit en la taula que hi ha a continuació.

Tipologia de motivacions de les persones a l'hora d'etiquetar

Tipus d'etiquetatge	Benefici social	Motivació
L'etiquetatge egoista: etiquetar en benefici propi, acostumen a ser etiquetes molt significatives per a l'usuari, però no per a la comunitat. Ex.: "per_llegir".	Si les etiquetes són més personals, es crea molt de soroll. A mesura que l'usuari indexa etiquetes més consistents, augmenta el benefici social.	Alta, per benefici propi.
L'etiquetatge amiguista: etiquetar per a compartir en un grup reduït (amics, companys, família). Usen etiquetes identificatives dins el grup però desconegudes per d'altres. Ex.: Tinet.	Molt útil dins el grup, però aporta poc a la resta de comunitats.	Alta, per a compartir i reforçar el sentiment de comunitat dins un grup.
L'etiquetatge altruista: etiquetar per a compartir amb tothom. S'escullen etiquetes generalment comprensibles i conegudes. Ex.: música_funky.	Molt alt. És la que més contribueix, la més generosa.	Baixa. No hi ha un benefici directe associat, tret de la satisfacció personal.
L'etiquetatge populista: etiquetar per a fer una cosa més atractiva i que tingui més visites. Ex.: Molt_interessant.	Cap. És correu brossa (*spam*).	Alta. Qui indexa així busca un benefici directe i evident.

Font: basat en Javier Cañada (2006).

La indexació resulta barata, ràpida, fàcil d'usar i té tot l'espectre possible de la terminologia, des dels termes més generals fins als més específics i actualitzats (si el document tracta de Tagzania l'usuari l'indexa Tagzania sense necessitat que un llenguatge documental controlat l'hagi recollit prèviament).

Ara bé, l'exhaustivitat no és homogènia, ja que els objectes no són descrits amb el mateix grau:

- Hi pot haver un recurs amb moltes etiquetes (exhaustivitat alta) i recursos amb poques etiquetes (exhaustivitat baixa).

- Hi pot haver documents indexats per a moltes persones que ens donaran enfocaments diferents sobre el mateix document o hi pot haver documents sense indexar.

2.2.3. Folksonomia

La indexació social és el procés distribuït en què els recursos es descriuen mitjançant etiquetes. El resultat agregat es coneix com a *folksonomia*[1], que significa 'classificació feta pel poble'. Són sistemes simples i eficients. La seva utilitat es deriva de la capacitat d'emparellar les necessitats dels usuaris amb un vocabulari habitual. No busquen la precisió.

(1)*Folksonomia*, de l'anglès *folksonomy*, és un neologisme. *Volk* (alemany) = 'del poble' + *Taxis* (grec) = 'ordenació' + *nomia* (grec) = 'regles'. Classificació feta pel poble.

Les folksonomies tenen dues dimensions relacionades (Hassan Montero, 2006): la personal i la col·lectiva.

- En la **personal**, *personomia*, cada usuari confecciona el seu propi índex d'etiquetes.

- En la **col·lectiva**, cada usuari comparteix les seves etiquetes i contribueix a generar un índex global d'etiquetes o folksonomia. Aquest aspecte resulta molt interessant en indexació, perquè un document descrit per cent usuaris amb etiquetes coincidents és una indexació més fiable (cal entendre recuperable) que la que faria l'autor. Hassan Montero parla d'*indexació per agregació*.

Podem classificar les folksonomies en dos grups (Hernández Quintana, 2008, i Weller, 2007):

- Les folksonomies estretes o *narrow*, que són del tipus "un document, un indexador", és a dir, només l'autor en pot etiquetar el contingut; seria el cas de Flickr.

- Les folksonomies generals o *broad*, en què un document pot ser etiquetat per diverses persones, com és el cas dels marcadors socials.

La tecnologia que fa possible les folksonomies s'activa el 2003 amb programes com Delicious i Flickr, i tenen un augment imparable fins al 2006, moment en què aquests programes ja ofereixen opcions de clusterització de les etiquetes (per exemple, Flickr etiquetes agrupades per categories). Tots dos són propietat de Yahoo.

Reflexió

L'any 2010 Yahoo, propietària de Delicious, va fer un informe en què anunciava que el web arribava a la seva posta de sol (*sunsetted*). Molts ho van interpretar com el tancament del web i la comunitat social va esclatar per por de perdre tot els marcadors que havia guardat al Delicious. La qüestió es va saldar amb la revenda de Delicious a l'empresa Avos System. Com a documentalistes seria bo que hi reflexionéssim, i que ens adonéssim de la indefensió dels usuaris davant les decisions empresarials de productes gratuïts com aquest. La recomanació dels experts és que exportem els nostres marcadors en paral·lel a altres programes com Diigo, Mr Wong.

Milers de persones que indexen etiquetes representa un volum considerable. És evident que contenen molta informació no solament sobre el contingut del document en qüestió, sinó sobre els usuaris del sistema i les seves rutines de cerca. Què se'n fa, de tantes etiquetes? Bàsicament hi ha dos enfocaments:

1) Aprofitar tot el coneixement de les folksonomies per tal de crear més coneixement (Navoni i González, 2009):

a) Usar les folksonomies com a complement d'altres sistemes d'indexació, que exerceixi algun control sobre les etiquetes. Es tracta d'aplicar tècniques d'indexació automàtica a les etiquetes, és a dir, aplicar mètodes estadístics sobre freqüència d'ús i coocurrència de les paraules.

b) Combinar les folksonomies amb sistemes controlats com ontologies. Es tractaria que un llenguatge documental controlat[2] proporcionés més noms d'etiquetes, que en el mateix context serien útils a l'etiqueta *x* introduïda per l'usuari.

(2)També hi hauria una sinergia positiva a la inversa, el llenguatge documental controlat es podria beneficiar de l'aportació continuada i actualitzada de vocabulari, que en definitiva és el que usa l'usuari.

Per exemple, l'usuari introdueix l'etiqueta *moneda* i l'ontologia li suggereix indexar, a més a més, *bancs*, *diners*, *encunyació*, *finances*, *or*, *plata*, *riquesa*.

2) Millorar la qualitat de la indexació. Es proposen dues línies:

a) Sistemes de recomanació d'etiquetes. L'usuari introdueix el web que vol etiquetar i el sistema li respon amb les etiquetes que altres usuaris han indexat en el mateix web, per si li són útils. D'aquesta manera, s'aconsegueix un cert control sobre el vocabulari i s'eviten alguns casos de sinonímia. El suggeriment és un suggeriment; l'usuari sempre els pot obviar. Podem classificar els llocs web que permeten la indexació social en dos grups: els que permeten posar etiquetes lliurement (Flickr o Youtube) i els que les suggereixen (Delicious). Suggerir etiquetes beneficia la recuperació, perquè augmenta la coherència entre internautes però empobreix l'espontaneïtat de l'usuari (Marcos, 2009).

b) Alfabetitzar l'usuari. Són diversos els autors (Hernandez Quintana, 2008; Noruzzi, 2006, i Spiteri, 2007) que proposen alfabetitzar l'usuari donant-li instruccions per a indexar. Apunten que les folksonomies han estat un canvi en la metodologia per la distribució i descentralització de la indexació, i podrien assolir més fites si s'organitzés la manera d'indexar i classificar la informació. Algunes de les propostes que es fan són la redacció de normes sobre:

- L'ús de substantius quantitatius i no quantitatius.
- L'elaboració d'etiquetes compostes (per exemple, amb un espai o guió entre unitermes).
- L'avaluació de la qualitat o aplicacions de cada ítem.
- L'ús d'enllaços a diccionaris que actuïn com a autoritats i controlin la forma de l'etiqueta.
- L'addició de noms personals provinents de llistes d'autoritats i afegir el rol que té amb el concepte que s'etiqueta.
- L'addició de tota mena de facetes (*faceted tagging*): geogràfiques (noms geogràfics provinents de llenguatges controlats com tesaurus), de temps, de forma, de gènere.

Les propostes que fan referència a copiar l'etiqueta des d'un vocabulari controlat (diccionari, tesaurus o classificació) són les més interessants i hi ha força articles que proposen usar la LCSH o la CDU o tesaurus, però també hi ha

la proposta d'indexar a partir dels articles de la Viquipèdia (creats de manera col·laborativa i amb el mateix esperit intuïtiu de les etiquetes) com a vocabulari controlat.

Observació

Fixeu-vos que si l'internauta escull un terme suggerit, vingui de la Viquipèdia, del WordNet o d'un càlcul estadístic del Delicious, ja està indexant de manera controlada i no lliure. Amb tot, el canvi no rau en la tipologia lliure respecte de la controlada, sinó en una tipologia nova, el que en anglès s'anomena ***user vocabulary*** (o provinent de la col·laboració social) davant el ***controled vocabulary*** (vocabulari fet per professionals).

2.3. La recuperació amb descriptors lliures

La indexació amb descriptors lliures, que tothom ha fet de manera individual (persona que indexa la seva biblioteca personal), pren una altra dimensió quan milers de persones fan el mateix. Malgrat els inconvenients de la manca de control sobre el vocabulari, que són evidents, és tan gran la seva aportació en el món del Web, que, malgrat ser imperfecta, resulta molt útil en la recuperació.

Avantatges i inconvenients de la recuperació amb descriptors lliures

Avantatges	Inconvenients
1) La comunitat es beneficia d'un volum immens de documentació mitjanament descrita. La qualitat pot ser discutible però està operativa, accessible. **2)** Es trenca la subjectivitat d'un únic indexador. **3)** Els punts d'accés són més diversos. **4)** No necessita traducció dels conceptes del llenguatge natural dels documents a un llenguatge artificial. **5)** Es tracta d'un tipus de llenguatge ràpid i fàcil d'actualitzar. **6)** S'adapta perfectament al nivell d'usuaris i tipus de SID, ja que és un llenguatge fet a mida. **7)** No cal una formació prèvia dels analistes. Precisament l'absència de regles i principis fan innecessària la formació. **8)** Indexen text però també imatge fixa (foto) i en moviment (vídeo, pel·lícula). **9)** Vocabulari amb autoritat d'usuari. **10)** El nombre d'indexadors augmenta la taxa de consistència.	**1)** Tots els que es deriven del llenguatge natural: • Sinònims • Polisèmics • Manca de termes relacionats que amplïin la cerca • Sigles o acrònims • Paraules sense significat en determinats contextos (ex.: la paraula *tuya*, que només té significat en Botànica). **2)** *Ego-centered tag* o etiquetes amb termes buits per a la comunitat, ja que només tenen sentit individualment. **3)** Nivell d'exhaustivitat divers, no tots els documents estan indexats amb el mateix grau.

En resum:

La indexació social participa en les característiques de les llistes de descriptors lliures en la filosofia de la indexació, ja que cada participant indexa uns descriptors lliures seleccionats per un procés intel·lectual a partir de l'examen del recurs sense verificar si els descriptors proposats existeixen o no en un llenguatge controlat. A mesura que han passat els anys, el volum d'etiquetes ha permès anar més enllà i crear un vocabulari de termes amb autoritat d'usuari (*user vocabulary*). Sobre els seus termes es poden fer càlculs estadístics i seleccionar les etiquetes amb la taxa de coherència entre indexadors més elevada o fer clusterització. El pas següent serà importar les etiquetes d'altres llenguatges, aquesta vegada controlats, com llistes d'autoritats (per als noms propis), tesaurus (per a noms geogràfics), etc. El Web semàntic permet als descriptors lliures crear sistemes basats en llenguatge natural i lliure, que de mica en mica s'aniran estructurant i controlant. La meta és un Web semàntic amb ontologies.

2.4. Activitats d'indexació amb descriptors lliures

1) Creeu un Delicious i introduïu-hi deu webs que indexareu amb etiquetes. Analitzeu les etiquetes que us suggereix Delicious. Són les que utilitzaríeu o en proposaríeu de noves?

2) Calculeu la taxa de coherència entre indexadors a Delicious.

3) Creeu un compte a Tagzania o Flickr. Quines opcions d'etiquetes us ofereixen?

3. Indexació automàtica

La indexació automàtica és el mètode per al qual un ordinador aplica un algoritme (o programa) a un document electrònic per tal d'identificar els termes que puguin representar la matèria i ser usats com a termes d'indexació i recuperació en un índex o llista.

La indexació automàtica és, juntament amb la indexació social, l'alternativa més viable per a indexar el Web.

Indexació automàtica

La indexació automàtica és un llenguatge natural, lliure, post-coordinat, alfabètic i analític per a paraules clau.

3.1. Com s'indexa automàticament?

El primer pas és llegir el text. Per a fer-ho, cal que el document es trobi en format electrònic i sigui accessible. Aquesta afirmació tan senzilla implica:

- Deixar fora la documentació audiovisual, imatge fixa (fotografies) o en moviment (vídeo) que habitualment no va acompanyada de text.
- També en queda fora tota la documentació que pertanyi a intranets (on cal contrasenya) i tota la que es generi dinàmicament (continguda en bases de dades), el que coneixem com a *internet invisible* i que es calcula que supera en cinc vegades el Web visible.

Després es prenen tot un seguit de decisions.

1) El document electrònic pot ser un text pla amb algun camp tipus *resum* i *paraules clau* o pot estar estructurat amb metadades, tant per al contingut com per a la forma. **Cal decidir si el programa s'aplicarà en el text complet o en camps determinats del document**; per exemple, només en el camp *paraules clau*. La qualitat del resultat serà molt diferent en un cas o en un altre: en el primer cas serà el programa que amb càlculs estadístics seleccionarà les paraules més representatives –per repetides– del text, mentre que en el segon cas els termes d'indexació ja han estat seleccionats per un procés intel·lectual.

Recordem que les metadades són dades formalitzades i que són una peça clau del Web semàntic, juntament amb el llenguatge XML i el format RDF.

Observació

L'XML és un llenguatge que té les propietats de l'HTML i la possibilitat d'incloure en el nivell de codi una infraestructura de metadades que expliciti la informació del recurs.

2) Què s'ha de fer amb els termes que contenen nombres, signes de puntuació, guions, majúscules/minúscules i accents? Habitualment són caràcters que no aporten significat, però que en determinats contextos poden ser determinants.

Nombre: N2, TV1

Punts, guions, signes (www.uoc.edu), Fonts_Informació (és una etiqueta pròpia de Delicious)

Accents (útils per a diferenciar diacrítics): en català, os/ós; en castellà, te/té

Observació

L'RDF és un marc de descripció de recursos (*resource description framework*, RDF) per a metadades desenvolupat pel World Wide Web Consortium (W3C).

3) Què s'ha de fer amb les paraules buides (articles, pronoms, preposicions, conjuncions, adverbis, numerals)? Són paraules molt freqüents, però que aporten poc valor de contingut. Es coneixen com a *llistes de paraules buides* en català, *listas de detención* en castellà i *stopword list* en anglès. Els programes d'indexació automàtica tenen un fitxer amb les paraules buides que han d'obviar. Ara bé, aquest fitxer pot estar implementat de tres maneres diferents:

Exemple de metadades

Feu clic a la icona *Indización* de la base de dades de revistes de la Universitat de Múrcia: http://revistas.um.es.

a) Predeterminat. El sistema disposa de bon començament de la llista de paraules buides del seu idioma o idiomes. De fet, la seva realització és fàcil, ja que només cal afegir les categories buides d'una base de dades de terminologia en l'idioma volgut. Els articles sempre són els mateixos, les conjuncions també, fins i tot els verbs es poden arribar a comptabilitzar i flexionar en tots els temps verbals.

b) Contextualitzat (*stop word context-dependent*). Cada sistema elabora la llista de paraules buides segons el seu àmbit temàtic. Contextualitzar la llista permet evitar dos inconvenients greus:

- Paraules amb significat que esdevenen buides.

  En un centre especialitzat en medicina de l'esport tots els documents faran referència a *medicina de l'esport* i, per tant, aquesta paraula serà buida en aquell context.

- Paraules buides que esdevenen importants en la indexació.

  En un text d'història els nombres (1319-1387), numerals (Pere III) i els adjectius poden tenir molta càrrega significativa (el Cerimoniós). En aquest exemple podem veure que *Pere III el Cerimoniós 1319-1387* podria quedar indexat com a *Pere* si no es mantenen algunes paraules buides.

c) Evitat expressament per a possibilitar al sistema la cerca per frases i sintagmes.

Per a recuperar un concepte com el nom del diari *El País*, en el qual l'article té un paper important. Els sistemes que els eviten, disposen d'altres eines per a reduir significativament el nombre de paraules indexades, com, per exemple, tècniques de *stemming* o *lematització*. En aquest sentit, més endavant parlarem dels marcadors discursius on veurem com paraules en principi buides ajuden molt en la decisió de quins termes seleccionar.

4) Aplicar mètodes estadístics. Un cop eliminades les paraules buides ens queda un conjunt d'unitermes amb significat, però així i tot el nombre pot ser molt elevat. El pas següent consisteix a seleccionar les de més rellevància

en la descripció del document. Aquest pas es resol aplicant diversos mètodes estadístics (o lingüístics i semàntics que veurem més endavant), bé en un ordre seqüencial, bé alternant els mètodes.

Els mètodes estadístics han estat la primera aproximació a la indexació automàtica i encara avui en dia en són una part consubstancial. La teoria de fons és el càlcul del pes (ponderació) de les paraules: ni les paraules més repetides (per buides) ni les menys repetides (per específiques) són adequades per a ser seleccionades. Els mètodes estadístics aplicats en PLN són de tres tipus (es poden usar sols o en combinació):

PLN

El processament del llenguatge natural (PLN o NLP del seu nom en anglès, *natural language processing*) és la disciplina informàtica que s'encarrega de tractar computacionalment les llengües naturals o llenguatges humans.

Les principals aplicacions o àrees de treball del PLN en l'actualitat són les següents:

- Recuperació de la informació
- Extracció de la informació
- Cerca de respostes
- Traducció automàtica
- Generació de resums
- Reconeixement de la parla

a) Freqüència. Hans Meter Luhn (1957) aplica la llei de Zipf al camp de la indexació automàtica. Luhn proposa els passos següents: calcular la freqüència de totes les paraules del text o col·lecció. Ordenar-les en ordre decreixent. Eliminar les de freqüència més alta. Eliminar les de freqüència més baixa. Indexar amb la resta.

b) Freqüència inversa. Sparck Jones (1972) va posar de manifest la capacitat de discriminació d'un terme enfront d'un altre. Aquesta discriminació ha de ser vista en el conjunt de la col·lecció, no en un sol document. Cal comparar les paraules clau entre els documents del fons per detectar quines són realment discriminatives.

c) Discriminació. G. Salton (1989), a partir de la idea que les paraules d'un text es classifiquen segons la seva capacitat per a discriminar uns documents dels altres en una col·lecció, va idear un sistema d'indexació, conegut com el **model de valor de discriminació**, que atribueix el pes o valor més alt a aquells termes que causen la màxima separació possible entre els documents d'una col·lecció. És a dir, el valor d'un terme depèn de com varia la separació mitjana entre els documents. Per tant, les millors paraules són les que aconsegueixen la distància més gran. L'anàlisi del **valor de discriminació** consigna una funció específica en l'anàlisi de contingut a les paraules simples, a les juxtaposades, a les frases i a grups de paraules.

5) Mètodes lingüístics. Els primers analitzadors lingüístics són de les dècades de 1960 i 1970. La seva aportació a l'anàlisi del contingut és cabdal, ja que permeten analitzar el text en tres nivells de profunditat: paraula, frase i text.

Cadascun d'aquests nivells és analitzat per mòduls del programa basats en diferents disciplines:

Paraula	Morfologia
Paraula dins la frase	Sintaxi
Paraula dins el text	Semàntica

Amb aquestes operacions s'aconsegueix un fitxer invers en què consten els unitermes i els documents en què apareixen. Cada uniterme va associat a un document i a una posició dins el document (per exemple, al títol).

6) Mètodes semàntics. La semàntica és la ciència que estudia el significat de les paraules i és una peça clau dins el PLN i el Web semàntic, valgui la redundància. Algunes de les propostes són els marcadors discursius i la participació de llenguatges controlats en tasques d'indexació automàtica.

a) Els marcadors discursius

El PLN encara és lluny d'oferir sistemes capaços d'entendre semànticament un text, com ho faria una persona, però està treballant en una línia molt interessant, que són els marcadors discursius. Es tracta de dotar l'algoritme del robot de les relacions semàntiques que es deriven de cinc grups de marcadors i d'aquí inferir un coneixement.

Els marcadors discursius són unitats lingüístiques invariables, per la qual cosa són automatitzables. Els cinc grans grups són els marcadors (Portolés).

Lectura complementària

Per a més informació sobre cada marcador discursiu podeu consultar el *Diccionario de partículas discursivas del español* de Briz, Pons, Portolés (http://textodigital.com/P/DDPD/).

Exemples d'alguns marcadors discursius

Marcadors	Exemples
Estructuradors de la informació	Primer, segon D'una banda, de l'altra Després, llavors
Connectors	Fins i tot, és més Així, doncs; per tant Tot i així, emperò
Reformuladors	És a dir, a saber, en altres termes En tot cas, en qualsevol cas
Operadors argumentadors	En realitat, en el fons En concret, en particular
Marcadors conversacionals	Naturalment, sens dubte Veritat? Eh?

Un dels marcadors estructuradors són els marcadors ordenadors que agrupen diversos ítems com si fossin parts d'un de sol, com ara

- Numèricament: primer, segon...
- En l'espai: d'una banda, de l'altra...
- En el temps: després, llavors, a la fi...

Si el programa té aquests marcadors podrà inferir un discurs més elaborat a partir del document i controlarà millor les parts discursives (introducció, cos, conclusions) i les parts orgàniques del text.

El programa mantindrà unit el conjunt d'ítems que d'una manera o d'una altra estaven ordenats amb els marcadors anteriors.

Així, si el text deia "primer Namíbia, segon Veneçuela, tercer Nepal..." el programa indexarà els tres noms i no un i prou, i els mantindrà relacionats.

Si el text deia "[...] el que investigava en el fons era el sodi", el programa detectarà un marcador argumentador (*en el fons*) i indexarà la primera paraula amb significat que vagi darrere (*sodi*).

Observació

Fixeu-vos que qualsevol d'aquests marcadors discursius es podria haver catalogat com una paraula buida, ja que són adjectius, conjuncions i adverbis, i el programa hauria perdut una informació molt valuosa a l'hora de mantenir indexades parts del text.

b) La participació de llenguatge documental controlat

Es tracta d'una indexació semiautomàtica, a diferència de les anteriors purament automàtiques.

El funcionament a grans trets consisteix en el fet que el robot detecta les paraules més significatives del document i les compara amb un vocabulari controlat, com un tesaurus o algun tipus de classificació, que a partir de les seves referències proposa un terme controlat per indexar.

En alguns sistemes aquest darrer pas és automàtic i en d'altres és una persona qui valida la decisió. De sistemes semiautomàtics de categorització n'hi ha de tres tipus:

- Categorització basada en regles.
- Basada en l'aprenentatge automàtic a partir de documents exemplars.
- Combinació dels models anteriors. És l'opció que més bons resultats dóna, però cal dedicar un temps al disseny de les regles i l'entrenament de documents exemplars.

7) La indexació automàtica no és tan sols una manera d'indexar i, per tant, un llenguatge documental en si, sinó que **també és una aplicació** de la qual tots els llenguatges documentals es beneficien.

Al llarg de cada llenguatge s'ha tractat com l'automatització dels processos d'indexació i recuperació pot agilitar tot el procés. Així, hem vist com es pot classificar de manera automàtica o semiautomàtica, com es pot descompondre un encapçalament de matèria controlat en una successió de paraules clau, com es poden crear tesaurus o indexar amb un tesaurus de manera automatitzada, el paper rellevant de les etiquetes i els càlculs estadístics que s'hi poden fer per suggerir etiquetes noves.

De cara al futur el més interessant és veure com els llenguatges documentals més potents i també més experimentats estan al dia del Web semàntic i ja els tenim en format SKOS:

- Ex CDU en SKOS (http://www.udcc.org/udcsummary/exports.htm),
- LCSH en SKOS (http://id.loc.gov/techcenter/metadata.html),
- la classificació Dewey, (http://oclc.org/developer/documentation/dewey-web-services/using-api).

3.2. La recuperació d'informació indexada automàticament

3.2.1. Cercadors

Al Web s'hi pot cercar de dues maneres: **navegant** o amb **cercadors**. És a dir, podem arribar a trobar una dada saltant d'una pàgina a una altra pels enllaços o bé posant els termes que volem en una caixeta d'un buscador. El primer sistema no implica cap tasca d'indexació, el segon sí i és una indexació automàtica.

Exemple

L'autor d'un blog penja un apunt sobre unes vacances a Sicília. L'autor no ha indexat el contingut de l'article, però nosaltres hi podem arribar bé saltant d'una pàgina que tenia enllaçada, o bé buscant al Google.

Els **algoritmes dels cercadors** comparen la paraula de la cerca amb les paraules contingudes en els textos de la seva base de dades. Funciona bé per a textos, però no per a material gràfic i audiovisual que no porti text o peu de fotografia.

L'usuari té la sensació que el cercador rastreja tota el Web buscant els termes que ha demanat com si fos en temps real, però això és una il·lusió perquè seria mecànicament impossible (milers d'usuaris cercant en paral·lel al Google i rebent respostes en temps real). En realitat, els cercadors no rastregen el Web en el moment de la consulta, sinó en el moment de la indexació. Rastregen i creen els seus fitxers inversos, que es van actualitzant.

Quan l'usuari fa una cerca, el programa no consulta el Web, sinó la seva base de dades del fitxer invers, per això la cerca es resol en segons.

La indexació automàtica no planteja gaires problemes, tret d'un, que és en quin ordre presenta els milers de resultats que troba. Les solucions han anat evolucionant en el temps: primer eren els documents que contenien els termes, després les cerques acotades amb els operadors booleans, després Google introdueix el concepte de *rellevància de la font* en funció dels enllaços que té i que rep, és a dir, ja no és solament la qualitat interna de la font, sinó també la qualitat externa que li atribueixen altres fonts.

Observació

Fixeu-vos que és el mateix criteri d'avaluació de la qualitat que es fa amb les publicacions periòdiques i el factor d'impacte, com el JCR d'ISI web of knowledge, In-recs, RESH, etc.

3.2.2. Recuperació en un web estructurat

La recuperació tal com l'entenem avui en dia patirà una **revolució** per l'ús d'ontologies i els motors d'inferència.

Estadístiques de buscadors

Els tres buscadors més usats segons les estadístiques són, per ordre, Google, Yahoo i Bing (AOL ho és a Amèrica).

El futur es presenta més enfocat cap a les cerques en context més apropiades per a aquests nous usuaris-editors-documentalistes. Es pretén usar les metadades per a fer càlculs sobre la rellevància del Web, la navegació per facetes (per lloc, temps, forma o qualsevol altra faceta pròpia d'un tema), cercar per fórmules que altres usuaris hagin usat reiteradament.

Com diu Mendez citant Witten, Gori i Numerico, anem cap a una "diversitat descentralitzada", en què interrogarem el Web de diverses maneres, i en què coexisteixen amb una anarquia organitzada de dades entrellaçades (documents, opinions, relacions, etc.).

Un dels avantatges de les metadades, és a dir, de partir de documents estructurats, és que l'usuari podrà cercar al Web com cerca en una base de dades, **per camps**.

> Això significarà que podrà acotar la cerca, per exemple, demanant documents en què es parli de *Bedrich Smetana* com a tema i no recuperar tota l'obra d'aquest músic (equivaldria a un catàleg demanar *Bedrich Smetana* com a matèria o *Bedrich Smetana* com a autor).

Una altra aplicació són els **sistemes de cerca de respostes**, que respondran directament a la pregunta, no oferiran un conjunt de documents en què aparegui el terme de la consulta, sinó que sortirà directament el fragment amb la resposta.

Des del punt de vista de la recuperació i llenguatges documentals, són interessants dues tècniques d'aquesta "diversitat descentralitzada", que són els vocabularis postcontrolats i les tècniques de clusterització. Totes dues tècniques parteixen d'un vocabulari lliure que el programa acabarà per controlar.

1) Els vocabularis postcontrolats (Lancaster). Es constata que els usuaris fan cerques curtes d'un o dos termes, que bolquen molts resultats. L'usuari no fa cerques llargues i elaborades amb operadors booleans, però els cercadors poden emmagatzemar les cerques d'altres usuaris i suggerir a l'usuari que busqui per aquest concepte i aquest altre. D'alguna manera, el cercador està indexant la pregunta i guarda la fórmula per a altres usuaris. El vocabulari és lliure però el robot el controla.

Exemple

Els usuaris acostumen a demanar *monovolums* però el programa ha emmagatzemat la fórmula (Monovolums) and (Seat or Volkswagen or Nissan...), que recuperarà de manera més exhaustiva. De fet, el programa està recollint els TE i TR (termes específics i termes relacionats) de monovolums.

2) Sistemes de clústers. La clusterització de dades és una tècnica molt comuna en l'anàlisi estadística de dades. Bàsicament, és la classificació d'objectes similars en diferents grups. Els clústers són carpetes classificades, segons la coaparició dels termes en el text. Se suposa que com més sovint apareguin junts els termes d'un tema determinat, més probable serà que els seus significats estiguin relacionats. El programa presenta les carpetes o els clústers en què apareix el tema que es busca, així l'usuari pot escollir l'enfocament que li interessi més.

Exemple

Un usuari busca el terme *llista de paraules buides* al cercador yippy (http://search.yippy.com) i aquest dóna noranta registres classificats en deu carpetes inicials (algunes carpetes s'obren) perquè l'usuari esculli: Search, My SQL Manual, Tools, Download, etc. En aquest cas, el programa ha sintetitzat el contingut dels resultats en forma de taxonomia.

3.2.3. Web semàntic: indexació i recuperació

El Web semàntic és un conjunt d'iniciatives destinades a promoure un futur Web amb pàgines organitzades, estructurades i codificades de tal manera que els ordinadors siguin capaços d'efectuar inferències i raonar a partir dels seus continguts.

Serà una **gran base de dades** capaç de suportar un processament sistemàtic i coherent de la informació (Codina, Pedraza, 2007).

El Web semàntic es basa en un llenguatge XML i uns formats comuns (RDF) que permeten la interoperatibilitat (*linked data*) amb independència de la plataforma des de la qual es treballi.

La indexació al Web semàntic es fonamentarà en la informació estructurada: els recursos web estaran descrits, és a dir, indexats en forma i contingut amb metadades (que poden haver estat generades manualment o automàticament), se cercarà amb agents intel·ligents que s'adaptaran a la nostra situació i els termes d'indexació s'interrelacionaran a partir d'ontologies.

Sembla que el més sensat és pensar que la indexació al Web semàntic consistirà en una combinació de tots els sistemes actuals, així:

- Es continuarà indexant de manera intel·lectual amb llenguatges controlats (classificacions, encapçalaments de matèria, autoritats i tesaurus) les fonts d'informació prou valuoses perquè el resultat no estigui condicionat per la inversió econòmica. Per exemple, bases de dades d'articles en ciències de la salut com ara MESH.

- L'ús de vocabularis controlats altament formalitzats i un PLN cada cop més potent propiciaran la implementació d'ontologies. Es crearan ontologies automàticament i manualment, i s'indexarà automàticament i manualment a partir d'ontologies.

- S'indexarà de manera semiautomàtica o semiassistida la gran majoria del Web, que per la seva mida no albira altres possibilitats. I s'espera que cada cop més els documents electrònics vinguin de sèrie amb metadades. Aquestes metadades, al seu torn, poden haver estat generades de manera intel·lectual o per un robot automàtic.

- S'indexarà socialment amb llenguatges lliures com els descriptors lliures o etiquetes, sobretot la informació audiovisual que no és fàcilment indexable de manera automàtica per no portar text. En aquest sentit, s'està investigant en robots que reconeguin formes simples en les imatges; de tota manera, fins que no siguin una realitat, la millor opció són les etiquetes dels internautes.

Un cas interessant: els wikis i les ontologies

Podem trobar dos enfocaments: el que considera un wiki una ontologia en la qual les pàgines són tractades com a conceptes i els enllaços que en surten i hi van es consideren relacions. A mesura que es crea el wiki, es crea l'ontologia. I el segon enfocament, que parteix de l'existència prèvia d'una ontologia a partir de la qual etiqueta semànticament les pàgines i relacions del wiki.

La recuperació en el Web semàntic consistirà, com diu Berners-Lee, no en una intel·ligència artificial màgica que permeti als ordinadors entendre les paraules dels usuaris, sinó en l'habilitat d'una màquina per a resoldre problemes ben definits, a partir d'operacions ben definides que es duran a terme sobre dades ben definides (W3C, 1999).

Webs recomanats

Buscador al Web semàntic http://swoogle.umbc.edu

Sobre metadades: http://ca.wikipedia.org/wiki/Metadades

Bibliografia

Manuals i articles de revista

AENOR (1990). *Documentación: Directrices para el establecimiento y desarrollo de tesauros monolingües*.

AENOR (1996). *UNE-50-125 (ISO 5964-1985). Documentación: Directrices para la creación y desarrollo de tesaurus multilingües*.

Aitchison, J.; Gilchrist, A.; Bawden, D. (2000). *Thesaurus construction and use: a practical manual* (4a. ed.). Chicago: Fitzroy Dearborn.

Bonilla, S. (2007). "Web Semántica y Agentes Metarrepresentacionales basados en Marcadores Discursivos" [en línia]. *Hipertext.net* (núm. 5) <http://www.hipertext.net>

Cañada, J. (2006). *Tipologías y estilos en el etiquetado social* [en línia]. <http://www.terremoto.net/tipologias-y-estlos-en-el-etiquetado-social/>

Codina, L.; Marcos, M. C.; Pedraza, R. (2009). *Web semántica y sistemas de información documental*. Gijón: Trea.

Currás, Emilia (2005). *Ontologías, taxonomía y tesauros: manual de construcción y uso*. Gijón: Trea.

Gómez Díaz, Raquel (2005). *La lematización en español: una aplicación para la recuperación de información*. Gijón: Trea.

Knautz, Kathrin; Stock, Wolfgang G. (2010). "Collective indexing of emotions in videos". *Journal of Documentation* (vol. 67, núm. 6, pàg. 975-994).

Lancaster, F. Wilfrid (2002). *El control del vocabulario en la recuperación de información*. València: Universitat de València.

Naumis, C. (2007). *Los tesauros documentales y su aplicación en la información impresa, digital y multimedia*. Mèxic: Alfagrama.

Noruzzi, Allreza (2006). "Folksonomies: (un)controlled vocabulary?". *A Knowledge Organization* (vol. 33, núm. 4, pàg. 199-203).

Ransom, N.; Rafferty, P. (2011). "Facets of user-assigned tags and their effectiveness in image retrieval". *Journal of Documentation* (vol. 67, núm. 6, pàg. 1038-1066).

Slype, van G. (1991). *Los lenguajes de indización: concepción, construcción y utilización en los sistemas documentales*. Madrid: Pirámide / Fundación Germán Sánchez Ruipérez ("Biblioteca del Libro").

Spiteri, Louise (2007, setembre). "The structure and form and folksonomy tags: the road to the public library catalogue". *Information Technology and Library*.

Trant, Jennifer (2009). "Studying Social Tagging and Folksonomy: A Review and Framework" [en línia]. *Journal of Digital Information* (vol. 10, núm. 1). < http://dlist.sir.arizona.edu/2595/>.