Anàlisi de contingut: resum i indexació

Manela Juncà Campdepadrós

PID_00143940

Universitat Oberta de Catalunya
www.uoc.edu

Índex

Introducció

Aquest mòdul us introdueix en els processos documentals de la segona fase de la cadena documental, anomenada *anàlisi de contingut*: el resum i la indexació.

Itinerari d'estudi

El mòdul comença amb un capítol dedicat a l'anàlisi de contingut, per a situar l'estudiant en les dues operacions esmentades: el resum i la indexació.

L'apartat dedicat al **resum** està dissenyat per a respondre les qüestions següents: què és un resum, qui el redacta, quines utilitats té i quants tipus de resums hi ha. Finalment, es presenten els resums automàtics i se n'explica l'evolució i el funcionament.

La **indexació** és el gruix d'aquesta assignatura i, en aquest mòdul, té tres apartats. El primer tracta de donar resposta a les preguntes següents: què és indexar, qui indexa, per què calen els llenguatges documentals i com s'indexa. El segon, titulat "Llenguatges documentals", respon les preguntes següents: què són els llenguatges, quants n'hi ha, què són els termes d'indexació, com han evolucionat, quan són necessaris i quan s'han d'usar en solitari o combinats. El darrer apartat, titulat "Tipologia", tracta dels diferents criteris usats per a classificar els llenguatges.

Aquest és un mòdul bàsic per a l'aprenentatge de la terminologia que s'usarà en la resta de mòduls.

Conceptes més importants

Concepte	Vegeu
Resum informatiu Resum indicatiu Resum selectiu Resum automàtic	1. El resum
Ambigüitat Llenguatge natural Exhaustivitat Especificitat Traducció Univocitat	2. La indexació

Concepte	Vegeu
Llenguatge documental Sistemes de classificació Llistats d'encapçalaments de matèria Llistes d'autoritats Tesaurus Llistats de descriptors lliures Llistats de paraules clau Notació Encapçalament Descriptor Identificador o autoritat Paraula clau	4. Els llenguatges documentals
Codificat Natural Lliure Controlat Precoordinat Postcoordinat Jeràrquic Combinatori Matèries Conceptes Paraules clau	5. Tipologia dels llenguatges documentals

Objectius

Amb l'estudi dels materials associats a aquest mòdul assolireu els objectius següents:

Quant al **resum**:

1. Aprendre a fer resums de manera intel·lectual: resums informatius, indicatius i selectius.

2. Aprendre a fer resums amb programes de resums automàtics.

Quant a la **indexació**:

1. Analitzar els factors necessaris perquè hi hagi una bona comunicació documental: entendre els problemes del llenguatge natural i la funció dels llenguatges documentals dins d'aquesta comunicació.

2. Conèixer els processos d'indexació: examen del document, selecció i traducció.

Quant als **llenguatges documentals**:

1. Conèixer les característiques principals dels llenguatges documentals.

2. Conèixer l'evolució històrica dels llenguatges documentals.

3. Aprendre a distingir i saber emprar la diferent tipologia dels llenguatges documentals: sintètic-analític, precoordinats-postcoordinats, controlats-lliures, jeràrquics-combinatoris, matèries-conceptes-paraules clau.

1. L'anàlisi de contingut

L'**anàlisi de contingut** se situa en la segona fase de la cadena documental i reuneix tot el conjunt d'operacions destinades a representar la matèria dels documents per a una recuperació posterior.

Són tasques de caire intel·lectual, en què la formació i l'habilitat de l'analista tenen un paper important.

"Representar la matèria" o "descriure el contingut" és respondre la pregunta "quin és el tema d'un document?".

Per representar el contingut d'un document l'analista ha de dur a terme dues operacions:

1) El **resum**, que condensa el contingut en un text més breu i manejable.

2) La **indexació**, que identifica els conceptes o temes principals. També es coneix com a *descripció característica*.

Aquestes dues operacions admeten una elaboració humana o automàtica. Per tant, hi haurà resums fets per documentalistes i resums fets per programes, i també indexacions fetes per analistes i indexacions fetes per un programari.

Operacions humanes i automatitzades

	Humà	Automatitzat
Resum	Resum informatiu Resum indicatiu Resum selectiu	Resum automàtic
Indexació	Sistemes de classificació Llistes d'encapçalaments de matèria Llistes d'autoritats Tesaurus Llistes de descriptors lliures	Llistat de paraules clau

Els dos sistemes tenen avantatges i inconvenients. La qualitat i la coherència que aporta un documentalista supera, en aquests moments, la que ofereixen els programes informàtics però, en canvi, els sistemes automàtics són instantanis, barats i capaços d'assumir quantitats de documents ingents.

La branca científica que estudia com emular el coneixement humà, quant a la identificació dels conceptes i les frases amb contingut rellevant per al resum i la indexació, és el processament en llenguatge natural.

El processament en llenguatge natural (PLN[1]) és una branca de la intel·ligència artificial i de la lingüística computacional que estudia els llenguatges que usen els humans per a interactuar amb els ordinadors en contexts escrits i orals.

A tall de conclusió

Per representar o descriure el contingut d'un document l'analista ha de dur a terme dues operacions:

- El resum, que condensa el contingut en un text més breu i manejable.
- La indexació, que identifica els conceptes o temes principals. També es coneix com a *descripció característica*.

Les dues operacions es poden dur a terme de manera humana o automàtica.

Vegeu també

Tractarem el processament en llenguatge natural en el subapartat 2.4 i en l'apartat 3.

(1)*PLN* és la sigla de *processament en llenguatge natural.*

Bibliografia

I. Gil Leiva; J. V. Rodríguez Muñoz (1996). "El procesamiento del lenguaje natural aplicado al análisis del contenido de los documentos". Revista general de información y documentación (vol. 6, núm. 2, pàg. 205-218).

2. El resum

Segons la norma UNE 50-103-90. *Preparació de resums*, un **resum** és la presentació abreujada i precisa d'un document, sense interpretació ni crítica, i sense menció expressa de l'autor del resum.

Vegeu també

Trobareu la norma UNE 50-103-90 en l'espai "Materials i fonts" de les aules.

Quan diem *document*, ens referim a tot tipus de document, sigui quin sigui el seu suport material. Podem resumir un text, la imatge d'una fotografia, un vídeo, àudios, informació en línia o hipertextos.

Els resums, com la indexació, poden ser d'elaboració humana o automàtica. En el primer cas hi ha quatre tipus de persones que poden redactar un resum. En el cas dels resums automàtics, es tracta d'un programari.

1) Resum humà:

a) L'**autor** del document. Els resums elaborats pels propis autors són molt habituals en el món de les comunicacions científiques i tecnològiques.

Resums per a revistes

Les revistes acostumen a donar directrius als seus autors per a l'elaboració de resums. Vegeu, per exemple, l'apartat "Instruccions per als autors" de la revista *EPI*.

b) Un **especialista** en la matèria de què tracta el document.

c) L'**editorial**. Són els resums que apareixen en la contraportada dels llibres impresos i que tenen una funció clarament publicitària.

d) Un **professional de la documentació**. Aporta el seu coneixement sobre la redacció de bons resums i l'elabora pensant en les utilitats futures.

2) Resum automàtic: els programes es coneixen com a programes resumidors de textos o *automatic text summarizer*.

Programes resumidors de textos

Un exemple de programes resumidors de textos és el Swesum, que fa una anàlisi estadística del text i elabora el resum amb els fragments que contenen les paraules més ponderades (les més repetides però amb significat).

La norma internacional ISO 214:1976, traduïda per AENOR (Norma UNE 50-103-90. *Preparación de resúmenes*), estableix les directrius que s'han de seguir per a presentar els resums en els documents. Posa un èmfasi especial en la preparació de resums per part dels autors dels documents primaris i en la mateixa publicació.

Redactar un resum és fàcil. En canvi, és força més difícil redactar un bon resum. El punt d'inflexió és la qualitat del resum, que el farà més o menys útil en un sistema documental. Un resum propagandístic no aportarà gaires conceptes principals per indexar, encara que hagi estat un bon reclam per a les vendes.

Exemple de resum elaborat per l'editorial amb finalitat publicitària

SAGAN, Carl. *Cosmos*. Traducció: Albert Santamaria i Martínez; pròleg: Ricard Guerrero. Barcelona: Publicacions i Edicions de la Universitat de Barcelona: Omnis Cellula, cop. 2006.

"Teniu a les mans una de les obres més destacades de la literatura internacional de divulgació científica, publicada per primera vegada en català. Una obra imprescindible d'un dels grans mestres de la divulgació, que ens endinsa en els grans enigmes que la humanitat ha tractat d'entendre i explicar des de temps immemorials, i pels quals ha nascut allò que anomenem ciència.

Des de la infinitud de l'Univers fins al món invisible dels àtoms, des del naixement de les estrelles fins a l'aparició de la vida, Carl Sagan aconsegueix transmetre els coneixements de la ciència actual d'una manera entenedora i apassionant."

Per a un analista només tindria utilitat el darrer paràgraf, en què surten termes com ara *univers*, *àtoms*, *estrelles*, *vida*.

El resum és útil en dues fases de la cadena, en els processos de selecció i d'adquisició que es dóna en la primera fase de la cadena i en la fase de sortida, en què és un instrument de recuperació excel·lent, ja que el resum ofereix més dades que la simple referència documental. La principal utilitat del resum és la de difondre la informació.

Difondre la informació

Cada cop més bases de dades referencials ofereixen el resum de les seves monografies i revistes, com per exemple Ebsco, Dialnet, Compludoc, CBUC, Eric database o *ISI current contents connect*. També ho fan les bases de dades de novetats editorials, com per exemple l'editorial Trea (en recomanem l'accés des de la Biblioteca de la UOC).

En tots els casos és indubtable el valor informatiu que aporta el resum per a difondre el contingut del document de la col·lecció. Però, a més, el resum té altres utilitats, tal com diu la norma UNE 50-103-90:

a) Determinar la pertinença: un resum ben elaborat capacita els lectors per a identificar de manera ràpida i precisa el contingut d'un document i decidir si cal llegir-lo en la seva totalitat.

b) Evitar la lectura del text complet en documents d'interès secundari. Un resum ben elaborat proporciona prou informació sobre temes que no siguin d'interès principal per al lector. Estalvia temps a l'usuari.

c) Ajudar en la cerca automatitzada. Els resums automatitzats incorporats als catàlegs són molt útils per a:

- Extreure termes d'indexació del seu text, és a dir, indexar a partir del resum.
- Fer cerques per paraules clau que no es troben en el títol.
- Servir de control bibliomètric, en comparar els termes usats en una equació de cerca amb els termes que apareixen en un resum i així establir la pertinença de la recuperació.
- Ajudar en la difusió des dels serveis d'alerta.

Segons Maria Pinto (1992), les **característiques d'un resum** són les següents:

Lectura complementària

Podeu ampliar la informació sobre el resum llegint l'obra següent:

M. Pinto Molina (1992). *El resumen documental: principios y métodos*. Madrid: Pirámide / Fundación Germán Sánchez Ruipérez ("Biblioteca del Libro", Y).

- Brevetat. S'han d'ometre dades preliminars o temes del coneixement comú.
- Pertinença. El resum s'ha d'adequar al missatge principal del document, sense obviar o interpretar les dades.
- Claredat i coherència. Frases completes, dotades de coherència lineal i global.
- Profunditat. Varia en funció del tipus de resum o de diferents nivells de detall que es persegueixin.
- Consistència lingüística. Un resum s'ha d'adaptar a les pautes lingüístiques en ús i ha de tenir en compte les regles morfològiques i sintàctiques corresponents.
- Proximitat cronològica entre les edicions del document original i el resum. És important que el temps transcorregut entre la publicació de l'original i el resum no sigui excessiu, especialment en àmbits científics i tècnics.

A tall de conclusió

- El resum és la presentació abreujada i precisa d'un document, sense interpretació ni crítica, i sense menció expressa de l'autor del resum.
- El resum pot ser redactat per l'autor del document, un especialista en la matèria, l'editorial, un documentalista o un programa informàtic.
- El resum és útil en dues fases de la cadena: en els processos de selecció i d'adquisició que es dóna en la primera fase de la cadena i en la fase de sortida, on és un instrument de recuperació excel·lent.
- La utilitat principal del resum és la de difondre la informació, però a més, el resum té altres utilitats, com són determinar la pertinença, evitar la lectura del text complet en documents marginals i ajudar en la cerca automatitzada.
- Els resums automatitzats incorporats als catàlegs són molt útils per a extreure termes d'indexació del text, per a fer cerques per paraules clau que no es troben en el títol, per a servir de control bibliomètric i ajudar en la difusió a través dels serveis d'alerta.

2.1. Tipus de resums

Hi ha diversos tipus de resums, segons la mida, els usuaris i l'aprofundiment en el contingut. Els tipus més habituals són els resums informatius, els indicatius i els selectius:

1) Resum informatiu

Redactarem el tema central, els temes addicionals, la naturalesa i l'objectiu del document, la metodologia, els resultats, les conclusions i els annexos. La idea de fons és que un resum informatiu pot substituir, en algunes ocasions, la lectura del document original. La norma UNE 50-103-90 recomana que l'esquema a seguir sigui el de:

objectiu + metodologia + resultats (o conclusions)

Tanmateix, no cal seguir forçosament aquest ordre, ja que hi ha entorns, com el tecnicocientífic, on es prefereixen els resums orientats als resultats (perquè la discriminació sigui més ràpida).

La mateixa norma dóna pautes pel que fa a l'extensió dels resums, però adverteix que en aquesta qüestió, haurà de ser més determinant el contingut del document que les pautes. Tanmateix, la norma ens suggereix:

- Monografies, informes, tesis: 500 paraules.
- Articles de revista, capítols de monografies: 250 paraules.
- Comunicacions breus: 100 paraules.

Exemple de resum informatiu

CONSUEGRA FERNÁNDEZ, Jesús: "El Ajedrez: evolución y claves de un juego milenario". En *Mundo antiguo*. Madrid: 2002. n° 3-4, año 1, p. 60-61.

"Article divulgatiu sobre el joc dels Escacs, estructurat segons el seus orígens, antiguitat, expansió, variants i simbolisme.

L'origen dels escacs és hindú, el primer representant conegut és el Ghaturanga aparegut entre el 3000 i el 2000 a.C., a Sri Lanka, tot i que no apareix documentat fins el segle VII d.C.

Del Ghaturanga procedeixen en cascada les diferents variants dels escacs: de la Índia va viatjar a Pèrsia en el segle VI d.C., on va passar dels 4 jugadors originals a 2 en la versió persa Shatranj. Des de Pèrsia es va estendre cap a Occident i cap a Orient.

Cap a Occident: paral·lel a l'expansió àrab el joc arriba a la península Ibèrica durant l'Alta Edat Mitjana, i des d'aquí s'expandeix a la resta d'Europa i la resta del món en l'època de les colonitzacions.

Cap a Orient: a la Xina, en el s.VII dC, els escacs prenen la forma de l'escac xinès Xiang qi; al Japó, el Shogi; a Indoxina, els escacs birmans i tailandès. Tant a Orient com Occident, presenten innumerables variacions locals.

El tauler i les fitxes semblen posseir un significat simbòlic. El tauler, amb l'alternança de caselles blanques i negres, forma un mandala. El simbolisme de les fitxes és menys esotèric i ha anat canviant segons els temps: bisbes, elefants, etc.

L'autor conclou que els escacs a més d'un joc és una eina educativa de primer ordre, quasi una ciència."

Com podeu comprovar, aquest resum té 241 paraules.

2) Resum indicatiu

Redactarem només les idees centrals del document. La seva lectura no pot substituir la lectura de l'original. Com el seu nom suggereix, el resum indicatiu presenta de manera abreujada i molt sintètica el contingut o la tipologia del document. La seva extensió pot oscil·lar entre una frase o quatre línies de text.

Exemple de resum indicatiu

CONSUEGRA FERNÁNDEZ, Jesús: "El Ajedrez: evolución y claves de un juego milenario". En *Mundo antiguo*. Madrid: 2002. n° 3-4, año 1, p. 60-61.

"Article divulgatiu sobre el joc dels Escacs, tracta del seu origen hindú, antiguitat, expansió històrica tant a Orient com Occident, variants nacionals i simbolisme del tauler i les fitxes."

3) Resum selectiu

Redactarem només una part concreta del document. El més habitual és el resum de conclusions, però també hi ha altres tipus, com la ressenya (*review*), que és una anàlisi del document amb elements crítics. Aquest tipus de resum s'adapta molt bé a les necessitats dels usuaris, per exemple investigadors o tècnics que necessiten una dada molt concreta sobre l'objectiu del document o les conclusions a què arriba.

Exemple de resum selectiu

CONSUEGRA FERNÁNDEZ, Jesús: "El Ajedrez: evolución y claves de un juego milenario". En *Mundo antiguo*. Madrid: 2002. n° 3-4, año 1, p. 60-61.

"Els escacs a més d'un joc és una eina educativa de primer ordre, quasi una ciència."

A tall de conclusió

Els resums més habituals són el resum informatiu, l'indicatiu i el selectiu:

- El **resum informatiu** consigna el tema central, temes addicionals, naturalesa i objectiu del document, metodologia, resultats, conclusions i annexos. La idea de fons és que un resum informatiu pot substituir en ocasions la lectura del document original.
- El **resum indicatiu** consigna només les idees centrals del document. La seva lectura no pot substituir la lectura de l'original.
- El **resum selectiu** consigna només una part concreta del document. El més habitual és el resum de conclusions, però també hi ha altres tipus com la ressenya (*review*).

2.2. Resum automàtic

Una de les necessitats més peremptòries davant de l'augment d'informació digital arrel del creixement exponencial d'Internet és manejar i filtrar el gran volum d'informació. Una de les solucions aportades pel PLN ha estat els programes de resum automàtic, que actuen sobre textos, imatges, webs i correu electrònic.

Els primers que van treballar en el camp de l'automatització dels resums van ser Hans Peter Luhn, l'any 1958, i Edmundson, el 1969, que varen aplicar tècniques com la freqüència de les paraules, o la posició d'una frase dins un document per a redactar resums sense intervenció humana.

Hans Peter Luhn

A partir d'aquestes primeres investigacions s'han perfeccionat moltes tècniques diferents basades en coneixement i recursos lingüístics (com les de Lin i Hovy, 2002; Gotti i altres, 2007) o les basades en mètodes estadístics i d'aprenentatge automàtic (Hirao i altres, 2002; Svore, 2007) (autors citats a Lloret i altres, 2008; i Mateo i altres, 2003).

Darrerament les investigacions giren al voltant del resum multidocument, és a dir, resumir més d'un document (Goldstein i altres, 2000; Qiu, 2007; Huo i Chen, 2008) de continguts afins o redundants (autors citats a Lloret i altres, 2008; i Mateo i altres, 2003).

Els resums automàtics es coneixen també com a *extracts*. La terminologia anglosaxona diferencia així els *extracts* i els *abstracts*. Els *extracts* són els resums formats a partir de l'extracció d'algunes frases del text prèviament seleccionades per un programa, mentre que els *abstracts* són els resums elaborats per una persona.

La base de totes les tècniques de funcionament d'un programa de resum automàtic és el còmput de la freqüència de les paraules.

Hi ha diverses eines per a fer aquests càlculs, com per exemple el WVTool. Es tracta de comptar quantes vegades surt una paraula no buida en el text.

Exemple de funcionament d'un programa de resum automàtic (extret de Lloret i altres, 2008)

"Tropical storm Gilbert formed in the eastern Caribbean and strengthened into a hurricane Saturday night. There were no reports of casualties."

Oració 1:	Tropical (2) storm (6) Gilbert (7) formed (1) in (0) the (0) eastern (1) Caribbean (1) and (0) strengthened (1) into (0) a (0) hurricane (7) Saturday (4) night (2).
Oració 2:	There (0) were (0) no (0) reports (1) of (0) casualties (1).

El primer que veiem és que les paraules buides, és a dir, les paraules que no tenen significat (preposicions, articles, verbs) no es computen.

Al costat de cada paraula amb significat veiem el nombre de vegades que surt en tot el text. Se sumen els valors, de manera que l'oració 1 té 3,2 punts i l'oració 2, 0,2. El programa seleccionarà la frase 1 com la més representativa per al resum automàtic.

Aquest sistema de resumir a partir de les frases amb les paraules més significatives en el text sembla simplista però té una certa justificació. Segons Kupiec i altres (1995) aproximadament el 80% de les frases en resums humans estan copiades literalment o amb petites modificacions del text original.

A partir d'aquesta base estadística, s'incorporen altres tècniques per a dotar el programa de més coneixement i pal·liar l'escassa coherència del resultat, com pot ser, per exemple, la resolució de l'anàfora o fer servir eines (per exemple, el WordNet) que proporcionin relacions com les de sinonímia o hiperonímia, o mecanismes per a detectar i eliminar la redundància.

Definim breument què són les anàfores i la hiperonímia:

Lectures complementàries

Podeu consultar els resultats de les investigacions d'aquests autors als articles següents:

E. Lloret; O. Ferrández; R. Muñoz; M. Palomar (2008). "Integración del reconocimiento de la implicación textual en tareas automáticas de resúmenes de textos". *Procesamiento del Lenguaje Natural* (núm. 41, pàg. 183-190).

P. L. Mateo; J. C. González; J. Villena; J. L. Martínez (2003). "Un sistema para resumen automático de textos en castellano".

Vegeu també

Trobareu l'explicació detallada sobre les paraules buides en el mòdul "Indexació automàtica i descriptors lliures".

a) Les **anàfores** són la relació de referència entre un element lingüístic i un d'anterior en el discurs.

Anàfora

"El Saló del Hobby ha tingut més de 60.000 visitants aquest any. Aquest saló ha esdevingut la fira d'oci familiar més visitada."

En aquest exemple, l'anàfora es dóna en "aquest saló", que fa referència al Saló del Hobby, expressat en la frase anterior. Com es pot comprovar, si en el resum automàtic apareix només la segona frase, el lector no sabrà a quin saló es fa referència.

b) Diem que una paraula és **hiperònima** quan té un camp significatiu que n'inclou un altre de menor extensió.

Hiperonímia

Color és un hiperònim. Els conceptes *groc, taronja* i *verd* són hipònims de l'hiperònim *color*.

Els experts consideren que la tecnologia actual no té problemes per a detectar les frases amb més significat, però sí que en té per a ordenar-les segons la seva importància.

Els programes funcionen, a grans trets, de la manera següent: es copia el text a resumir o bé s'escriu l'adreça del document. S'escull el tipus de document (acadèmic, periodístic, etc.) i el tant per cent de reducció del text.

A continuació teniu una selecció dels programes més coneguts:

[(2)]Podeu practicar amb el programa SweSum, que és gratuït i tradueix castellà.

- Connexor
- Daedalus
- Extractor
- FociSum
- InTEXT (Dinamic Summarizing)
- Inxight Summarizer
- IslandInText
- K-Site de Daedalus
- Pertinencer Summarizer
- Sinope Summarizer
- Summarizer
- SweSum[2]
- System Q
- TextAnalyst
- Trestle

El programa K-Site de Daedalus

Dels programes de resum automàtic esmentats, vegem el funcionament del programa K-Site de Daedalus. Aquest programa té cinc mòduls:

- **Mòdul 1: Anàlisi morfosintàctica**. En aquest mòdul es determina la categoria lèxica de cada paraula: substantiu, verb, adjectiu, article, preposició, etc. També es determina el lema. Aquestes operacions permeten destriar les paraules amb significat (substantius, adjectius, verbs) de les paraules buides (articles, preposicions, pronoms, etc.). El lema permet agrupar totes les paraules que són flexions d'una altra (info/informar/informació/informador/informacional/etc.). El producte final és una llista amb les paraules puntuades i una llista de frases candidates.

- **Mòdul 2: Ponderació de frases**. Aquest mòdul rep les paraules etiquetades pel mòdul anterior. La seva funció és escollir entre totes les frases candidates. Per a fer-ho s'ajuda de diversos submòduls que ponderen les frases segons els paràmetres següents: la freqüència, la presència de paraules indicatives (busquen paraules com ara *important, essencial, conclusions*, etc.), busquen frases que continguin paraules que apareguin al títol, o que tinguin noms propis, o que la tipografia sigui destacada (negretes, cursi-

ves, mida superior, etc.) i seleccionen frases que apareguin en posicions destacades en el text (al principi de cada paràgraf, al final a mode de conclusions).

- **Mòdul 3: Detecció d'anàfores**. Un cop té les frases seleccionades, pot ser que es doni el cas d'anàfores mal resoltes (una frase conté una anàfora que es trobava en la frase prèvia i que no ha estat seleccionada). El programa busca les anàfores (especialment els demostratius pronominals o pronoms personals, per exemple *aquest, aquell, la qual cosa*, *això*) i la seva posició en la frase: al principi, entre les sis primeres paraules, en altres posicions.

- **Mòdul 4: Selecció de frases**. Aquest mòdul computa tota la informació recollida en les fases anteriors: frases candidates, puntuacions, detecció d'anàfores. Selecciona les frases candidates de puntuació més alta fins arribar al tant per cent demanat per l'usuari. Si entre aquestes frases n'hi ha alguna que contingui una anàfora, se selecciona la frase anterior (que conté la paraula a la qual s'està fent referència) sempre que formi part de les frases candidates i no sobrepassi la longitud del resum.

- **Mòdul 5: Postprocessament de l'extracte**. La seva funció és detectar expressions que connecten parts del text, ja sigui per mostrar causalitat, contraposició, etc. Són expressions del tipus *per tant, en contra*, etc. Com en el cas de les anàfores, si formen part d'una frase seleccionada es procura incloure en el resum la frase amb la qual estan relacionades.

Finalment, cal recordar que alguns processadors de textos, com el Microsoft Word, també ofereixen aquesta opció (*autosummarize* o autoresum).

A tall de conclusió

- Els resums automàtics (*extracts*) són una de les solucions aportades pel PLN per a fer front al maneig de grans volums d'informació en línia.

- Els primers investigadors que van treballar en el camp de l'automatització dels resums van ser Hans Peter Luhn, l'any 1958, i Edmundson, el 1969.

- Les tècniques han evolucionat des dels primers còmputs sobre la freqüència de les paraules, o la posició d'una frase dins un document, fins a les tècniques basades en coneixement i recursos lingüístics o basades en mètodes estadístics i d'aprenentatge automàtic.

- La base de totes les tècniques és el còmput de la freqüència de les paraules. A partir d'aquesta base estadística s'incorporen altres tècniques per a dotar el programa de més coneixement i pal·liar l'escassa coherència del resultat, com per exemple la resolució de l'anàfora o l'aplicació d'eines que proporcionin relacions com les de sinonímia o hiperonímia o mecanismes per a detectar i eliminar la redundància.

- Els experts consideren que la tecnologia actual no té problemes per a detectar les frases amb més significat, però sí en té per a ordenar-les segons la seva importància.

3. La indexació

"Indexar és l'acció de descriure o identificar un document en relació al seu contingut."

Norma UNE 50-121-91.

Indexar és el resultat d'examinar el document, seleccionar els conceptes i emmagatzemar-los en una base de dades.

Aquesta definició implica tres accions, de les quals la més significativa és la de la selecció dels conceptes i la seva traducció al llenguatge documental.

Igual que en el resum, la indexació pot ser feta per una persona o per un programa.

Vegeu també

La indexació s'estudia en els mòduls "Sistemes de classificació documentals", "Llistes d'encapçalaments i llistes d'autoritats", "Els tesaurus" i "Llistat de descriptors lliures i llistat de paraules clau".

Si la indexació és intel·lectual, és a dir, si la duen a terme persones, aquestes persones poden ser:

- **Professionals** (documentalistes), que duen a terme la tasca d'indexació de manera individual o en equip. Al seu torn, els equips poden indexar de manera centralitzada o coordinada.

- **Amateurs** (usuaris d'Internet que indexen de manera social o *tagging* –per exemple, a Delicious).

Vegeu també

La forma d'indexar dels equips es tracta en l'apartat 5 del mòdul "La cadena documental" d'aquesta assignatura.

L'element humà permet una anàlisi més rica del document, captant conceptes i matisos que un programa no arribaria a detectar, però té l'inconvenient del temps que s'hi ha de dedicar i la coherència entre indexadors.

La indexació automàtica es fa a través d'un programa informàtic. El seu funcionament és molt senzill: extreuen del títol, resum o text complet les paraules més significatives. És un mètode econòmic i molt ràpid.

Vegeu també

La indexació automàtica s'estudia en els apartats dedicats al llistat de paraules clau del mòdul "Llistat de descriptors lliures i llistat de paraules clau".

3.1. Llenguatge natural i llenguatge documental

Per a indexar necessitem els llenguatges documentals. Quina diferència hi ha entre el llenguatge natural i el documental?

Per **llenguatge natural** entenem el llenguatge que usem quotidianament, català, castellà, basc, gallec, francès, etc.

Per **llenguatge documental** entenem la llista o el vocabulari de termes que usem per a indexar i que poden estar en format lliure o controlat.

I per què cal controlar els termes del llenguatge natural? Perquè el llenguatge natural és ambigu i els conceptes es poden representar de diverses maneres, fet que dóna lloc a problemes de recuperació. El llenguatge natural és ric en terminologia, en formes (plurals i singulars), en temps verbals, acrònims, sinònims, polisèmies, etc.

La riquesa del llenguatge natural

- Exemples de sinònims del mateix concepte: cosmos / univers / infinit / firmament / cel.
- Exemple del mateix concepte en formes diferents, sigles o frases, i en idiomes diferents: OTAN / NATO / Organització del Tractat de l'Atlàntic Nord / Organización del Tratado del Atlántico Norte / North Atlantic Treaty Organization.
- Exemple de polisèmia: banc / planta / carta / serra / estrella / llengua / capital.

La principal diferència entre el llenguatge natural i el documental controlat és precisament el control terminològic, que permet representar els conceptes de manera unívoca, això és, sense ambigüitats.

Per ser més concrets, les diferències es donen en: el nombre de termes del vocabulari, el control de les formes, el control del significat i les relacions de significat entre termes.

Univocitat

La univocitat consisteix a representar un concepte amb un únic terme.

3.1.1. Nombre de termes

Els llenguatges documentals són entròpics (Blanca Gil, 2004, pàg. 20), és a dir, tendeixen a la selecció, a la restricció del vocabulari. És el procés contrari del llenguatge natural, que tendeix a l'abundància, a la reiteració de conceptes, a la sinonímia en benefici d'una expressió més rica.

Els llenguatges documentals redueixen considerablement el número de termes del llenguatge natural, ja que només prenen en consideració els substantius i alguns sintagmes nominals, però no adjectius, preposicions, conjuncions, adverbis, verbs, etc. A més, entre tots els substantius, n'escullen un que representarà la resta quan el significat sigui el mateix. I entre diverses formes d'un mateix terme, només una serà l'acceptada, com és el cas de les sigles.

Els llenguatges documentals són en essència senzills; i la seva eficàcia augmenta a mesura que les reiteracions i la redundància són controlades en una única forma que aplega conceptes afins.

3.1.2. Control de les formes

Els llenguatges documentals controlen les formes plural/singular, l'ús d'acrònims i sigles i la construcció de les frases, i d'aquesta manera estableixen uns models.

Exemple

Model	Exemple
Substantiu	Pintura
Substantiu + adjectiu	Pintura medieval
Substantiu + preposició + substantiu	Pintors de vitralls

Aquestes regles gramaticals i sintàctiques unifiquen les paraules seleccionades i les frases.

Exemples en les llistes d'encapçalaments de matèria

- S'acostuma a usar el singular per a expressar conceptes abstractes. Així, per exemple, és *solidaritat* i no pas *solidaritats*.
- No es permet l'ús de sigles; es prefereix l'expressió sencera del concepte i en la llengua del servei d'informació i documentació (SID[3]). Per exemple, Organització del Tractat de l'Atlàntic Nord.
- És preferible l'expressió natural del concepte compost i no la forma inversa. És correcte *Objectes d'art*, i no, *Art, objectes de*.

Vegeu també

Els millors exemples es veuen en els mòduls "Llistes d'encapçalaments i llistes d'autoritats" i "Els tesaurus".

(3) *SID* és la sigla de *servei d'informació i documentació*.

3.1.3. Control del significat

Els problemes més importants quant al significat són la sinonímia i la polisèmia.

a) Sinonímia: diem que les paraules són sinònimes quan tenen el mateix significat. En un sistema documental, si no es controlen i s'usen indiscriminadament, comporten silenci documental. En el cas d'"aliment, nutrient, menjar, provisió", l'usuari pot estar cercant per "aliment" i no recuperar documents perquè es troben indexats amb altres formes, com ara "nutrient". La solució dels llenguatges controlats és recollir tots els termes sinònims i seleccionar-ne un per a representar tot el conjunt de termes que tenen el mateix significat perquè dos sinònims són substituïbles l'un per l'altre en qualsevol context.

Exemple

Una llista d'encapçalaments de matèria com la del Centre Superior d'Investigacions Científiques (CSIC) recull tots aquests sinònims:

- Hispanoamericanos.
- Iberoamericanos.
- Latinoamericanos.
- Sudamericanos.

Però només dóna com a terme acceptat "Latinoamericanos". Si al SID arribés un document titulat "Los sudamericanos del siglo XX", l'analista l'indexaria com a **Latinoamericanos**, ja que és el terme acceptat.

b) Polisèmia: diem que dues paraules són polisèmiques quan el mateix signe lingüístic, paraula o so té més d'un significat. Habitualment el context de la conversa o de la lectura en què està inserida la paraula desfà els problemes d'ambigüitat, però un mot polisèmic introduït en un sistema documental, sense el context, pot donar lloc a soroll documental.

Exemple

Un usuari pot estar buscant informació sobre columnes en arquitectura i recuperar dades sobre columnes tipogràfiques de diaris. Els llenguatges documentals controlen la polisèmia diferenciant cada significat amb parèntesis, usant el plural o el singular, adjectivant, etc.

Un tipus de polisèmia és la homonímia. La diferència entre aquests dos conceptes rau en l'etimologia de la paraula. Si l'etimologia de les dues paraules és la mateixa, parlem de polisèmia; si l'etimologia és diferent, parlem d'homonímia.

Exemples de polisèmia i homonímia

Mateixa etimologia = polisèmia

La polisèmia es dóna quan una paraula té un únic origen etimològic i acaba tenint significats diferents sense canviar la seva categoria gramatical: per exemple, no passa de substantiu a verb. És una paraula que amb el temps ha anat adquirint diferents significats, però tot i així, tots mantenen una relació de significat, per exemple, en català i castellà fulla/hoja que ve del llatí *folia*, té diversos significats: fulla d'una planta, fulla de metall d'una eina, pàgina d'un llibre, cada una de les parts d'una porta doble o finestra, etc. I en tots els significats té implícita la idea d'una làmina.

Si volem saber si una paraula és gramaticalment polisèmica només cal consultar un diccionari etimològic i veure si prové d'un mateix origen. Trobarem la paraula, un únic origen i una llista de diferents significats. En català podem consultar el *Diccionari de la llengua catalana* i en castellà, el *Diccionario de la Real Academia*.

Més exemples de polisèmia:

- *Servei*, del llatí *servitium*, que ha donat lloc a oficis religiosos, lavabos, missions militars, coberts per menjar i, en esports, posar la pilota en joc. I en tots ells roman la idea de ser útil.

- *Creuer*, del llatí *crux*, que significa 'creu', intersecció entre les dues naus d'una església, encarregat de dur la creu al davant d'una processó, viatge de plaer pel mar, etc. En aquests significants la idea és la de la forma de creu, creuar com anar d'un extrem a un altre.

- *Columna*, del llatí *columna*, que usem per a referir-nos als pilars arquitectònics, les parts verticals d'una pàgina impresa d'un diari, en física la forma que adopten alguns fluids, com "columnes de fum", en l'àmbit militar, la formació de vaixells o soldats. I la idea que roman és la de verticalitat.

Diferent etimologia = homonímia

L'homonímia es dóna quan dos conceptes han arribat a tenir el mateix nom, la mateixa forma, però tenen orígens diferents i, per tant, etimologies diferents.

Per exemple *metro* pot ser el transport urbà, una unitat de mesura o l'estri per a mesurar. Però l'origen etimològic entre el transport i els altres dos significats és evident, el primer és una abreviació de la paraula anglesa *metropolitan* i el segon cas ve del grec *μέτρον* i significa mesura.

Un altre exemple, la paraula castellana *botín* pot venir del llatí *bota* i significarà 'calçat fins al turmell', o pot venir del germànic *bytin* i significarà 'premi d'una conquesta'.

En castellà i català aquest fenomen és menys freqüent que en altres llengües, com l'anglès o el francès, en què abunden les paraules homònimes que donen molt de joc en els acudits.

Dins l'homonímia podem diferenciar els mots que tot i que s'escriuen igual tenen significats diferents, anomenades *homògrafs*, com els anteriors *metro* o *botín*, de les paraules que malgrat que sonen igual també tenen significats diferents, conegudes com *homòfones*: *vell*/*bell* en català, o *tubo*/*tuvo* en castellà.

També hi ha paraules on coincideixen les dues característiques i són polisèmiques i homònimes al mateix temps: en català la paraula *clau* pot venir del llatí *clavis* i significarà la clau per a obrir una porta, la clau d'un conflicte, la combinació d'una caixa forta, la clau d'un arc de volta, la clau de sol o les paraules clau, i en totes les accepcions sempre té el significat d'una peça essencial que obre i tanca. Però també pot venir del llatí *clavus* i significarà clau de ferreteria, instrument usat en cirurgia per unir fractures òssies, etc., i la idea és d'una unió. En castellà un bon exemple és *bota*, que si ve del llatí *bota* significa 'calçat', si ve del llatí *buttis* significa 'tina per a guardar el vi' o 'recipient de cuir tou amb brocal per a beure', i si ve del got *bauths* significa 'obtús', entre altres accepcions.

En resum, la sinonímia provoca silenci documental, i la polisèmia i les seves variants provoquen soroll documental. El control terminològic del vocabulari garanteix el criteri d'univocitat que han de tenir els llenguatges documentals controlats, segons el qual un concepte es representa amb un terme, i un terme només pot tenir un significat.

3.1.4. Relacions de significat dels termes

Per **relacions de significat** entenem la relació de genèric, específic o relacionat que pot tenir un terme respecte a un altre.

En llenguatge natural aquestes relacions són implícites. Per exemple, quan parlem de *pomes* tots entenem que es tracta d'una fruita fresca i que les Fuji i les Golden són varietats concretes. És a dir, situem el terme *poma* dins d'una jerarquia de termes conceptualment més genèrics (fruita) i més específics (Golden, Fuji). Fins i tot podem relacionar per associació d'idees la poma amb altres fruites, com la taronja o el plàtan. Però en un llenguatge documental cal definir aquestes relacions, agrupant i relacionant els termes afins.

L'estructura que relaciona els termes és implícita en el llenguatge natural, però en els llenguatges documentals cal fer-la explícita. Això es pot fer de dues maneres:

a) En una seqüència jeràrquica, on la mateixa posició del concepte ja defineix els seus termes genèrics i específics. També desfà problemes de significat.

Exemple de la pesca

Vegeu l'exemple de la pesca extret de la Classificació Decimal Universal (CDU). El concepte *pesca* pot ser l'activitat econòmica o la pesca com a esport. Si ens fixem en la cadena jeràrquica veiem que cada un penja d'una classe diferent:

```
6 Ciències aplicades. Medecina. Tecnologia
63 Agricultura i ciències relacionades
   639 Caça. Pesca

7 Belles arts. Jocs. Esports
79 Diversions. Espectacles. Jocs
   799 Caça esportiva. Pesca esportiva.
```

b) En una presentació alfabètica on cada terme s'acompanya de tots els seus termes relacionats, ja siguin equivalents, genèrics, específics o relacionats.

El tesaurus del CSIC

En el tesaurus de Psicologia del CSIC, consultem "somnis" i trobem:

Sueños

TG Dinámica de la personalidad

TE Contenido del sueño
TE Pesadilla

TR Déjà vu
TR Interpretación de los sueños
TR Sueño fisiológico
TR Sueño REM
TR Trastornos de conciencia

Les sigles informen del tipus de relació que estableixen: TG significa terme genèric (per sobre de "Somnis" el tesaurus té "Dinàmica de la personalitat"), TE són els termes específics (són termes específics de "Somnis": Contingut del somni, malsons) i els TR són els termes relacionats (es relacionen amb "Somni" "Déjà vu", la "Interpretació dels somnis", el "Son REM", etc.).

Finalment, els avantatges i els inconvenients principals del llenguatge natural i el documental controlat són:

Avantatges i inconvenients dels llenguatges documentals

	Avantatges	Inconvenients
Llenguatge natural	Amigable Actualitzat Econòmic	Dificulta la cerca Poc precís
Llenguatge documental controlat	Unívoc Facilita la cerca	Car Poc actualitzat

A tall de conclusió

Indexar és l'acció de descriure o d'identificar un document amb relació al seu contingut.

La indexació pot ser feta per una persona (de manera centralitzada o de manera coordinada) o per un programa.

Per *llenguatge natural* entenem el llenguatge que usem quotidianament (català, castellà, basc), i per *llenguatge documental* entenem la llista o el vocabulari de termes que usem per a indexar i que poden estar en format lliure o controlat. La principal diferència entre el llenguatge natural i el documental controlat és el control terminològic:

- El control del nombre de termes del vocabulari: els llenguatges documentals són entròpics, tendeixen a la selecció, és a dir, a la restricció del vocabulari.
- El control de les formes: els llenguatges controlats controlen les formes plural/singular, l'ús d'acrònims i sigles i la construcció de les frases.
- El control del significat: els llenguatges controlats controlen la sinonímia i la polisèmia. Diem que dues o més paraules són sinònimes quan tenen el mateix significat. Diem que dues paraules són polisèmiques quan el mateix signe lingüístic té més d'un significat. La sinonímia provoca silenci documental i la polisèmia, i les seves variants, provoca soroll documental. El control terminològic del vocabulari garanteix el criteri d'univocitat que han de tenir els llenguatges documentals controlats, segons el qual un concepte es representa amb un terme, i un terme només pot tenir un significat.
- Les relacions de significat entre els termes: són les relacions de genèric, específic o relacionat que pot tenir un terme respecte a un altre. En llenguatge natural aquestes relacions són implícites però en els llenguatges documentals cal fer-les explícites per mitjà d'una seqüència jeràrquica o una presentació alfabètica.

3.2. Com s'indexa?

Ara que ja hem vist la necessitat de comptar amb llenguatges documentals per a pal·liar l'ambigüitat del llenguatge natural, estem en condicions de preguntar-nos pel procés d'indexació que duu a terme un analista.

A continuació presentem les **fases** que proposen diversos autors abans d'arribar a la que ens servirà com a marc de referència en aquest subapartat:

- Dues fases: anàlisi del text i traducció (Chaumier, 1988; Fidel, 1994).
- Tres fases: anàlisi del text, identificació de conceptes i traducció (Amat, 1989; Norma UNE 50-121-91).
- Quatre fases: anàlisi del text, identificació de conceptes, traducció i establiment d'enllaços sintàctics entre descriptors (Slype, 1991).
- Cinc fases: registre de dades, anàlisi del text, identificació de conceptes, traducció i examen de la indexació.

En aquest mòdul seguirem la **norma UNE 50-121-91** i les seves tres etapes:

Norma UNE 50-121-91

UNE 50-121-91. *Métodos para el análisis de documentos, determinación de su contenido y selección de términos de indexación.*

1) Examinar el document per a identificar-ne el contingut.

2) Seleccionar els conceptes principals del contingut.

3) Traduir a un llenguatge documental.

Exemple

Examinem un llibre titulat *Mites d'antigues civilitzacions*. Llegim el títol, el resum, el sumari, etc.

En una segona etapa seleccionem com a conceptes principals: Mites, Grècia, Roma, Índia, Japó, Indis nord-americans.

En la tercera etapa indexem. Si indexem amb un llenguatge lliure podem escriure el terme com desitgem o com surti al text. Per exemple:

```
Mitologia índia americana.
```

En canvi, si indexem amb un llenguatge controlat haurem de traduir aquests conceptes a una forma controlada. Posem, per exemple, que pensàvem indexar "Mitologia índia americana". Veiem com quedaria en tres llenguatges documentals diferents:

```
CDU                259.2
LEMAC              Mitologia ameríndia
LEM del CSIC       Indios de Amèrica - Religión y mitología
```

A continuació es detalla cada part del procés.

1) Examen del document i identificació dels conceptes

L'analista ha d'examinar amb precisió el document. La lectura completa és, sovint, impracticable, però sí que ha de prestar atenció al títol, al resum, al sumari, a la introducció, a les il·lustracions i les paraules o frases destacades en una tipografia diferent.

No es recomana la indexació només a partir del títol, perquè hi ha títols que porten a error, i tampoc no és bo confiar en què el resum sigui un substitut del text, ja que no tots els resums estan ben elaborats.

Exemple de títols i resums que no aporten dades significatives per a la indexació

- CHESNEAUX, Jean. *¿Hacemos tabla rasa del pasado?* México: Siglo XXI Editores 1981. La seva matèria és *Història, historiadors, historiografia*. Al catàleg de la Biblioteca Nacional de España (BNE[4]) el trobem indexat com a "Historia".

- MALLOL, Tomas. *Si la memòria no em falla*. Girona: CCG Edicions 2005.
  La seva matèria és *Memòries, cinema, col·leccionisme*. A la Biblioteca de Catalunya (BC[5]) el trobem indexat com "Cinema amateur".

Si recordem el resum del llibre de Carl Sagan, *Cosmos*, ens adonarem que el resum no era suficient per a indexar el contingut d'aquesta obra. Per aquests motius es recomana una lectura àgil de la resta de parts significatives del document.

(4) *BNE* és la sigla de *Biblioteca Nacional d'Espanya*.

(5) *BC* és la sigla de *Biblioteca de Catalunya*.

Vegeu també

Recordeu que l'exemple del resum del llibre de Carl Sagan, *Cosmos*, sortia a l'apartat 2 d'aquest mòdul.

2) Selecció dels termes d'indexació

Tal com diu la norma UNE, l'analista ha d'identificar les nocions que són elements essencials de la descripció del contingut. Si la indexació és compartida, la institució que la patrocina ha d'establir clarament els factors que considera importants.

Per a seleccionar els conceptes del document, l'analista ha de ser conscient del nombre de conceptes (criteri d'exhaustivitat) i de la seva exactitud (criteri d'especificitat).

a) Exhaustivitat

A mesura que l'analista va llegint ha d'anar prenent nota dels conceptes interessants del document.

Una bona praxi és la que identifica els conceptes rellevants sobre:

- El tema.
- Els noms personals que puguin ser interessants d'indexar.
- Els noms geogràfics.
- Les dates cronològiques.
- La forma en què es presenta el document: article, estadística, formulari o divulgació, científic, etc.

L'exhaustivitat és un criteri relacionat amb el nombre de conceptes que es tenen en compte per a caracteritzar el contingut sencer d'un document. El criteri de selecció principal és el valor potencial del concepte per als usuaris del seu SID.

Podem distingir entre una exhaustivitat baixa, mitja i alta en funció del nombre de descriptors. És en aquest entorn on la norma UNE 50-121-91 dóna les seves recomanacions quant a l'exhaustivitat. Els criteris que l'indexador ha de tenir en compte són:

- El tipus de SID i el perfil d'usuari. No és el mateix indexar per a una base de dades genèrica que per a una d'específica.
- El tipus de document. No s'indexa amb el mateix nombre de descriptors una monografia que un article de revista o una tesi.

Tal com recomana la norma UNE no és convenient ser estrictes amb el nombre de termes, no s'ha de limitar el nombre de manera arbitrària dient, per exemple, "per a una monografia dos termes d'indexació", ja que aquesta idea pot conduir a una pèrdua d'objectivitat i a una deformació de la informació. És preferible suggerir un barem, entre tants i tants termes, per a cada tipus documental i SID i ser flexibles, ja que els criteris que han de regir són el contingut del document i la seva recuperació posterior.

Exemple

A partir del resum informatiu següent, elaborarem tres tipus d'indexacions suggerint un barem (per a aquesta assignatura i les seves pràctiques) i una finalitat:

Cuervo Herrero, C.; Fernández González, A.: "Objetos celestes erróneos". *Tribuna de Astronomía y Universo. Revista de Astronomía, Astrofísica y Ciencias del espacio*. 2000. II Época, n° 16 – octubre. p. 36-40.

"Anàlisi i descripció dels errors més freqüents que cometen els professionals i afeccionats a la fotografia astronòmica mentre intenten descobrir nous objectes celests encara no identificats.

Aquests errors són deguts a quatre causes: errors en el procés de positivat de la còpia com a conseqüència de la presència de partícules de pols en els negatius o a les lents de l'equip de laboratori; errors en el negatiu, deguts a defectes de rentat, deficiències en l'emulsió, ratlles i rascades o per l'ús de pel·lícules de color destinades a ser forçades, i errors en les lents dels objectius, deguts a efectes de distorsió i a alteracions en la refracció. Finalment, es descriuen altres causes: reflexos de la llum del sol sobre les antenes de satèl·lits artificials Iridum, els retocs digitals o de fotocopiadores i duplicadores, ús d'objectius senzills i poc potents per a captar imatges de cel profund i, en darrer terme, oscil·lacions del condensador de llum del microscopi.

Tots aquests errors poden donar lloc a imatges falsejades: objectes inèdits, diàmetres erronis, efectes d'arrodoniment, alineacions planetàries errònies, etc. L'article facilita imatges d'aquests errors fotogràfics.

Els autors conclouen que cal ser cautelós i fer les oportunes comprovacions abans de donar a conèixer el descobriment d'un nou objecte celeste a les societats astronòmiques."

Exemple dels tres graus d'exhaustivitat

Exhaustivitat baixa	**Exhaustivitat mitjana**	**Exhaustivitat alta**
Barem 1-3	**Barem 4-6**	**Barem 7-...**
Exemple d'us: catàleg d'una biblioteca pública	Exemple d'ús: bases de dades d'una biblioteca especialitzada en Astronomia	Exemple d'ús: bases de dades d'una biblioteca especialitzada en Astrofotografia
• Errors fotogràfics • Fotografia astronòmica	• Astrofotografia • Errors fotogràfics • Descobriments • Identificació d'objectes celestes • Objectes erronis	• Alineacions planetàries • Defectes de rentat • Deficiències de l'emulsió • Diàmetres erronis • Efectes d'arrodoniment • Errors en el negatiu • Errors en el positivat • Errors en les lents • Objectes inèdits • Objectius • Oscil·lacions del microscopi • Partícules de pols • Ratllades • Reflexos del sol • Retocs digitals

b) Especificitat

L'especificitat està relacionada amb l'exactitud en què un concepte particular que apareix en un document està representat per un terme d'indexació.

Exemple

Si en el text que estem indexant apareix el concepte *Diplomàcia*, i aquest terme apareix en el llenguatge documental controlat, hem d'indexar "Diplomàcia". Si indexem "Relacions internacionals" o "Ambaixadors" no estarem essent específics, com podeu veure a la taula següent:

Exemple d'especificitat

Matèria	Correcte i, per tant:	Incorrecte per	
	Específic	Genèric	Massa específic
Diplomàcia	**Diplomàcia**	Relacions internacionals	Ambaixadors

Els conceptes s'han d'identificar de la manera més específica possible, però en determinats casos es poden preferir nocions més genèriques:

- Quan l'indexador consideri que un excés d'especificitat pot ser negatiu en la recuperació; per exemple, pot decidir que un model molt específic d'una màquina s'indexi amb el nom més genèric d'aquest tipus de màquines.

- Quan la idea no estigui plenament desenvolupada en el document, o només s'hi fa al·lusió.

- Quan s'estigui a l'espera de validar el terme més específic.

3) Traducció a un llenguatge documental controlat

Per a traduir el concepte inicial escrit en llenguatge natural a un llenguatge documental l'indexador ha de consultar les llistes del llenguatge buscant la forma correcta d'introduir el concepte.

Exemples

Concepte tal com surt al text	Traducció	Llenguatge documental utilitzat
Tragicomèdia	791.221.28	Classificació Decimal Universal (CDU)
Eolític	Edat de la pedra	Llista d'encapçalaments de matèria en català
Matriz	Útero	Lista de encabezamientos del CSIC
Monarquía absoluta	Absolutismo	Tesaurus d'Història contemporània del CSIC

Quan l'analista procedeix a traduir el concepte del text es pot trobar en les situacions següents:

a) Troba el concepte, sol o repartit per les taules:

- Consulta el llenguatge i troba el concepte a la primera. Llavors indexa amb aquest terme d'indexació. Per exemple, buscava "Eolític" i troba que ha d'indexar "Edat de la pedra".

- Consulta el llenguatge i troba el concepte o les parts del concepte repartits pel llenguatge. Llavors ha de conèixer les regles de combinació de les parts integrants del terme d'indexació. Exemples:
  - Una notació amb CDU com 391.91(961.3) "Tatuatges de l'illa de Samoa" està formada per 2 elements: tatuatges + Samoa. Aquests elements van col·locats en un ordre determinat per les regles de precoordinació de la CDU (primer la classe principal + auxiliar).
  - Un encapçalament construït amb la LEM del CSIC com és Agua-Aspectos económicos està format per dues parts: Agua + Aspectos económicos, que són un encapçalament i un subencapçalament respectivament i van en aquest ordre.

Amb els llenguatges tesaurus i la llista d'autoritats no hi ha una sintaxi de combinació.

b) No troba el concepte:

- Consulta el llenguatge i no troba el concepte. Llavors l'indexador ha de conèixer les obres de referència que el seu SID considera com a autoritats reconegudes en la matèria. Aquestes obres de referència són diccionaris, enciclopèdies, altres llenguatges documentals (especialment els tesaurus construïts d'acord les normes ISO i UNE 50-106 i UNE 50-125), atles, etc.

- Hi ha llenguatges, com ara tesaurus, en què l'indexador ha de proposar el terme nou com a descriptor candidat i esperar que la direcció del tesaurus el validi com a descriptor mentre indexa amb un terme més genèric.

3.3. Qualitat i coherència de la indexació

La **qualitat** i la **coherència** de la indexació depenen de factors com la competència de l'indexador i la qualitat dels instruments o llenguatges documentals. La coherència és un factor important en el comportament d'un sistema d'indexació, especialment quan forma part d'una xarxa de centres i la informació s'ha d'intercanviar entre ells.

La coherència es calcula de la manera següent: dos analistes indexen el mateix document, amb un llenguatge de descriptors com un tesaurus. Es compten separadament el nombre de descriptors idèntics entre els dos analistes sobre el total de descriptors.

Exemple

Com exemplifica van Slype:

- El documentalista 1 ha assignat els descriptors A, B, C, D, E, F.
- El documentalista 2 ha assignat els descriptors A, C, D, F, G, H.
- Hi ha 4 descriptors idèntics A, C, D, F i un total de 8 descriptors diferents. Taxa de coherència = 4/8 = 50% (van Slype, 1991, p. 123).

La consistència en la indexació sol oscil·lar entre el 20% de mínima i el 60% de màxima (Isidoro Gil, 2001).

A tall de conclusió

La norma UNE 50-121-91. *Mètodes per a l'anàlisi de documents, determinació del seu contingut i selecció de termes d'indexació* estableix tres fases:

- Examinar el document per a identificar-ne el contingut: l'analista ha d'examinar amb precisió el document. La lectura completa és, sovint, impracticable, però sí que ha de prestar atenció al títol, al resum, al sumari, a la introducció, a les il·lustracions i les paraules o frases destacades en una tipografia diferent.

- Seleccionar els conceptes principals del contingut: l'analista ha d'identificar les nocions que són elements essencials de la descripció del contingut, ha de ser cons-

Lectures complementàries

Podeu ampliar la informació sobre la coherència en la indexació llegint les obres següents:

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

I. Gil Leiva (2001).

cient del número de conceptes (criteri d'exhaustivitat) i la seva exactitud (criteri d'especificitat).

- Traduir a un llenguatge documental: per a traduir el concepte inicial escrit en llenguatge natural a un llenguatge documental cal consultar la llista del llenguatge buscant la forma acceptada.

4. Els llenguatges documentals

Un **llenguatge documental** és un vocabulari de termes en llenguatge natural o un sistema artificial de signes normalitzats que faciliten la representació del contingut dels documents.

Les seves funcions principals són indexar el contingut dels documents i permetre'n la recuperació a partir del camp matèria.

Hi ha sis llenguatges documentals:

- Els sistemes de classificació.
- Les llistes d'encapçalaments de matèria.
- Les llistes d'autoritats.
- Els tesaurus.
- Els llistats de descriptors lliures.
- Els llistats de paraules clau.

En teoria tots els documents es poden indexar amb qualsevol d'aquests sis llenguatges, però a la pràctica la tipologia del SID (si és arxiu, biblioteca o centre de documentació) i el tipus d'usuari (general o especialitzat) condicionen que un SID indexi i recuperi amb un o altre llenguatge. En línies generals:

- les biblioteques indexen amb sistemes de classificació + llistes d'encapçalaments de matèria + llistes d'autoritats;
- els centres de documentació indexen amb tesaurus + llistes de paraules clau;
- els arxius amb sistemes de classificació i/o tesaurus.

Com podeu observar, els SID poden treballar amb un sol llenguatge o amb una combinació de llenguatges.

4.1. Els termes d'indexació

Anomenem *terme d'indexació* a la representació d'un concepte en llenguatge natural o un codi de classificació.

Els termes d'indexació poden estar formats per una o més d'una paraula.

Exemple

Exemples de termes d'indexació:

- D'una paraula: "Boscos".
- De més d'una paraula: "Font d'informació".

La part més petita amb significat d'un terme d'indexació es coneix com a *unitermе*.

La norma UNE 50-113-92/1 defineix els unitermes com:

"l'element significatiu més petit d'un llenguatge documental utilitzat per a representar un concepte específic en un sistema d'indexació coordinat, no s'ha de confondre amb paraula clau o descriptor."

UNE 50-113-92/1.

Exemple

El descriptor "Font d'informació" està format per dos unitermes: "Font" i "Informació". La preposició "de" no s'indexa.

Cada llenguatge documental dóna un nom diferent al seu terme d'indexació. Aquesta és la terminologia que usarem en aquesta assignatura:

Termes d'indexació

Llenguatge documental	El seu terme d'indexació es coneix com a
Sistemes de classificació	Notació o símbol de classe
Llistes d'encapçalaments de matèria	Encapçalament
Llistes d'autoritats	Autoritat, identificador o descriptor
Tesaurus	Descriptor
Llistes de descriptors lliures	Descriptor
Llistes de paraules clau	Paraula clau

La norma UNE 50-113-92/1 defineix aquests conceptes de la manera següent:

- "Notació/Símbol de classe: és la representació d'una classe mitjançant la notació d'un sistema de classificació.
- Identificador: nom utilitzat com a descriptor.
- Descriptor: termes d'indexació assignats per l'analista fruit d'alguna de les operacions intel·lectuals que implica el procés d'indexació.
- Paraula clau: una paraula o un grup de paraules seleccionades de manera automàtica del títol, resum o text d'un document que en representen el contingut i en permeten la recuperació."

Norma UNE 50-113-92/1. *Documentación e información. Vocabulario. Parte 1. Conceptos fundamentales*

Lectura recomanada

Per a qüestions de terminologia recomanem la consulta de la norma UNE 50-113-92/1. *Documentación e información. Vocabulario. Parte 1. Conceptos fundamentales*. A: *Documentación: Normas fundamentales*. Madrid: AENOR, 1994.

A tall de conclusió

Un llenguatge documental és un vocabulari de termes en llenguatge natural o un sistema artificial de signes normalitzats que faciliten la representació del contingut dels documents. Les seves funcions principals són indexar el contingut dels documents i permetre'n la recuperació a partir del camp matèria.

Hi ha sis llenguatges documentals:

- Els sistemes de classificació.
- Les llistes d'encapçalaments de matèria.
- Les llistes d'autoritats.
- Els tesaurus.
- Els llistats de descriptors lliures.
- Els llistats de paraules clau.

Anomenem *terme d'indexació* a la representació d'un concepte en llenguatge natural o un codi de classificació. Els termes d'indexació poden estar formats per una paraula o més d'una.

4.2. Evolució històrica dels llenguatges documentals

Els primers analistes mesopotàmics, egipcis o romans, llegien el document, copiaven les primeres línies del text o seleccionaven els conceptes que representaven millor el contingut i els escrivien en la tauleta, *pinake*, *cartela* o fitxa corresponent. Mica en mica, aquestes matèries van anar conformant una llista de temes. A l'edat mitjana sabem de l'existència de catàlegs d'algunes grans biblioteques, com la de Lorsh a Alemanya, que tenia 600 títols classificats en 63 matèries.

Edat contemporània

Ara bé, per a molts autors la història dels llenguatges documentals comença a les biblioteques del segle XIX amb els sistemes de classificació, ja que van ser el primer intent seriós de controlar les matèries dels documents.

Library of Congress

Els sistemes de classificació van començar a ser considerats pròpiament llenguatges al segle XIX amb les **classificacions bibliogràfiques** de Brunet, Harris, Dewey, Cutter o la de la Library of Congress. Eren quadres de classificació jeràrquics, de caire enciclopèdic, i les seves classes es combinaven d'una forma definida amb anterioritat, és a dir, precoordinada. Els conceptes es representaven amb codis, no paraules. Per exemple, el concepte "Fotografia" era el codi 77 (exemple extret de la CDU).

Charles Ammi Cutter

El següent pas en l'evolució dels llenguatges va ser formulat per Charles Ammi Cutter el 1876, que va crear una llista de matèries, escrites en llenguatge natural. Ja no s'usava un codi, sinó que s'expressava el concepte (com "Fotografia") amb totes les lletres. Aquestes llistes, anomenades **llistes d'encapçalaments de matèria**, eren alfabètiques i es basaven en els principis d'especificitat (cal indexar amb el terme específic, no el genèric) i el d'entrada directa (cal respectar l'ordre natural de les expressions i no optar per formes inverses del tipus "Electrònic, comerç").

Les col·leccions bibliotecàries estaven recollides amb aquests dos llenguatges documentals: sistemes de classificació + llistes d'encapçalaments de matèria. Les llistes d'autoritats controlaven la resta d'autoritats. A més, es combinaven en els registres bibliogràfics per tal de minimitzar l'inconvenient de la codificació, ja que no era de fàcil comprensió per als usuaris. La indexació era sintètica, sumària, dues o tres entrades per al camp matèria ja que hem de ser conscients que van néixer en sistemes no automatitzats.

A mesura que la producció científica anava generant cada cop més volum d'informació, sorgeix la necessitat d'indexar d'una manera més analítica, amb més conceptes. Es creen **centres de documentació** amb una vocació més espe-

cialitzada que les biblioteques. L'ús de tecnologia informàtica facilitava l'accés a un document per diversos punts d'accés. Neixen els llenguatges especialitzats per excel·lència, els **tesaurus**. S'apliquen als centres de documentació i a alguns arxius històrics i administratius.

Els tesaurus recullen el bo i millor dels seus antecessors: l'estructura arborescent dels sistemes de classificació, que aplica a la seva presentació jeràrquica, i l'estructura combinatòria de les llistes d'encapçalaments de matèria, que aplica a la seva presentació alfabètica. A més, inclou noves estructures de presentació, com la gràfica i la d'índexs permutats.

Els tesaurus s'automatitzen i, des de mitjans dels anys setanta del segle passat, el creixement de la indústria de les bases de dades possibilita la consulta en línia de moltes publicacions seriades. Neix el darrer llenguatge documental, la **llista de paraules clau o indexació automàtica**.

Internet

La darrera gran etapa la marca **internet**. La globalització de la xarxa a partir de la dècada dels anys noranta impulsa l'accés a la informació. No cal que els SID disposin del document en propietat, ja que la xarxa permet accedir a la informació hostatjada en qualsevol altre centre d'informació. La cooperació impulsa tots els llenguatges documentals a automatitzar-se i a formar part de projectes col·lectius (catàlegs col·lectius, consorcis, xarxes). En el mateix sentit es busquen passarel·les entre els diferents llenguatges per a solucionar problemes idiomàtics entre països.

Sorgeix la necessitat d'indexar l'abundant producció de recursos electrònics, com per exemple amb l'ús de metadades per a definir i intercanviar dades entre sistemes informàtics (etiquetes del tipus <subject>,<keywords>) i explotar la indexació automàtica en els potents robots dels cercadors. També els usuaris poden indexar els recursos gràcies a iniciatives d'indexació social o *tagging*.

Els experts opinen que en l'actualitat el problema principal no és tant indexar o recuperar, sinó presentar els resultats en algun ordre significatiu, la qual cosa implica l'ús d'algoritmes que valorin els resultats.

Lectura complementària

Podeu trobar aquesta cronologia a l'obra següent:

I. Gil Leiva (2008). *Manual de indización. Teoría y pràctica* (pàg. 110-114). Gijón: Ediciones Trea ("Biblioteconomía y Administración Cultural", 193).

A continuació reproduïm algunes de les dates més significatives, extretes de la cronologia d'Isidoro Gil (2008) sobre les llistes d'encapçalament de matèria, els tesaurus i la indexació automàtica:

Cronologia de l'evolució dels llenguatges documentals

Dates	Concepte	Breu explicació
30.000 aC	**Etiquetes de fang**	**Els antics escribes mesopotàmics guardaven les tauletes de fang (documents) en cistelles de mim. Per fora la cistella duia una altra tauleta de fang amb el contingut.**

Dates	Concepte	Breu explicació
Egipte	Les *carteles* d'Egipte	Els egipcis introdueixen el papir com a suport documental. El papir s'enrotllava al voltant d'una vareta de fusta o metall. Per a no desplegar completament el rotllo, posaven les primeres frases del document en una etiqueta o *cartela* en un extrem.
1876	**Charles A. Cutter Rules for a dictionary catalog**	
1895	List of subject headings for use in dictionary catalogs	Publicat per l'American Library Association (ALA) per a biblioteques mitjanes i petites, amb fons no especialitzats.
1909	**Library of Congress Subject headings**	**Neix a partir de la llista de l'ALA i les regles de Cutter. A partir d'aquí aquesta llista es converteix en el referent de totes les llistes d'encapçalaments de matèria del món.**
1923	List of subject headings for small libraries	Minnie Earl Sears és l'autora d'aquesta llista coneguda com SEARS. És una versió reduïda de la LCSH per a biblioteques petites.
1934	**Guia para los encabezamientos de materia**	**Juan Manrique Lara publica la primera llista d'encapçalaments en castellà a Mèxic. Era una traducció de la Library of Congress Subject Headings (LCSH), l'ALA i la SEARS.**
1946	Répertoire de vedettes-matière RVM	Primera llista d'encapçalaments en francès (Universitat de Laval Canadà).
1951	**Descriptor**	**Calvin Mooers encunya el terme.**
1952	Uniterme	Mortimer Taube encunya el terme.
1957	**Indexació automàtica**	**Hans Meter Luhn comença a treballar en indexació automàtica aplicant el mètode de la freqüència.**
1960	Compatibilitat	En la dècada dels anys seixanta del segle passat s'inicien els primers projectes per a fer compatibles els diferents llenguatges documentals mitjançant taules d'equivalència.
1961	**Sistema SMART**	**Gerald Staton desenvolupa el sistema SMART per a anàlisi automàtica de textos.**
1967	Guidelines for the development of information retrieval thesauri	Directrius per elaborar Tesaurus confeccionades pel US Federal Council for Science and tecnology de Washington.
1967	**Lista de encabezamientos de materia para bibliotecas**	**Llista compilada per Carmen Rovira i Jorge Aguayo en espanyol per a la Unió Panamericana.**
1974	Norma ISO 2788:1974 Guidelines for the establishment and development of monolingual thesauri	1a. edició de la norma ISO per a la confecció de tesaurus monolingües.
1980	**Répertoire d'autorité-matière encyclopédique et alphabétique unifié - Rameau**	**Primera llista d'encapçalaments de matèria de la Biblioteca Nacional de França. Es varen basar en la RVM i la LCSH.**
1983	Bilindex	Llista d'encapçalaments bilingüe en anglès i castellà. És equivalent a la LCSH. L'any 2007 editava la 15a. ed.
1985	**Norma ISO 5963:1985 Methods for examinating documents**	**Norma ISO que no va ser traduïda a norma UNE fins l'any 1991 amb el número UNE 50-121-91.**
1985	Norma ISO 5964:1985 Guidelines for the establishment and development of multilingual thesauri	1a. edició de la norma ISO per a la confecció de tesaurus multilingües.
1986	**Abandonament dels símbols tradicionals de les llistes d'encapçalaments pels propis dels tesaurus**	**La LCSH en la seva 10a. edició abandona els símbols de x, see, xx, v, a pels propis dels tesaurus Use, BT, NT, RT. La resta de llistes mundials també els adopten.**

Dates	Concepte	Breu explicació
1986	Unified medical language system	Sistema unificat de llenguatges en medicina és un projecte per a integrar els diferents vocabularis en ciències de la salut. És un projecte de la Biblioteca Nacional de Medicina dels EUA (actualment coordina el MESH Medical subject headings).
1995	**Universalització d'internet**	**Internet ha difós i popularitzat conceptes, tècniques i pràctiques pròpies de documentalistes.**
1995	Metadades	Ús de metadades per a definir i intercanviar dades entre sistemes informàtics. Els llenguatges de marcatge tenen etiquetes per al resultat de la indexació del tipus <subject> i <keywords>.
1997	**Projecte MACS**	**Iniciativa de la Conference of European National Libraries CENL per a fer compatibles tres llistes d'encapçalaments de matèria, l'alemanya SWD, el RAMEAU francès i la LCSH usada a Gran Bretanya i a Suïssa.**

A tall de conclusió

Per a molts autors la història dels llenguatges documentals comença a les biblioteques del segle XIX amb els sistemes de classificació, ja que van ser el primer intent seriós de controlar les matèries dels documents.

El següent pas en l'evolució dels llenguatges, el va formular Charles Ammi Cutter el 1876, creant una llista de matèries, escrites en llenguatge natural.

A mesura que la producció científica anava generant cada cop més volum d'informació, sorgeix la necessitat d'indexar d'una manera més analítica, amb més conceptes. Es creen centres de documentació amb una vocació més especialitzada que les biblioteques. Neixen els llenguatges especialitzats per excel·lència, els tesaurus.

Des de mitjans dels anys setanta el creixement de la indústria de les bases de dades possibilita la consulta en línia de moltes publicacions seriades. Neix el darrer llenguatge documental, el llistat de paraules clau o indexació automàtica.

La darrera gran etapa la marca Internet. La globalització de la xarxa a partir dels anys 1990 impulsa l'accés a la informació. La cooperació impulsa tots els llenguatges documentals a automatitzar-se i a formar part de projectes col·lectius (catàlegs col·lectius, consorcis, xarxes). En el mateix sentit es busquen passarel·les entre els diferents llenguatges per a solucionar problemes idiomàtics entre països.

Sorgeix la necessitat d'indexar l'abundant producció de recursos electrònics, com per exemple amb l'ús de metadades per a definir i intercanviar dades entre sistemes informàtics (etiquetes del tipus <subject> i<keywords>) i explotar la indexació automàtica en els potents robots dels cercadors. També els usuaris poden indexar els recursos gràcies a iniciatives d'indexació social o *tagging*.

4.3. Quan són necessaris els llenguatges documentals?

Els llenguatges documentals són necessaris en dos moments de la cadena documental:

- La fase d'anàlisi i tractament > Anàlisi documental > Anàlisi de contingut > Indexació.
- La fase de sortida > Instruments de recuperació.

Tant en la fase indexació com en la fase recuperació, el procés d'anàlisi-selecció-traducció de conceptes és el mateix. En el moment de la indexació l'analista llegeix el document, extreu conceptes i, si cal, els tradueix a un llenguatge controlat per a emmagatzemar-los en el sistema. En el moment de la recuperació, l'analista ha de treballar amb la consulta de l'usuari, extreure'n els conceptes i traduir-los. Si es tracta d'un llenguatge postcoordinat, a més, haurà de saber com convertir els descriptors a una equació de cerca.

Exemple de la fase de recuperació

- **Usuari**: "Necessito informació sobre les instal·lacions esportives d'hoquei herba que es varen construir a la ciutat de Terrassa amb motiu de la celebració dels Jocs Olímpics del 1992".

- **Analista**: selecciona els conceptes més rellevants per a la cerca, instal·lacions esportives, hoquei herba, Terrassa, Jocs Olímpics. El pròxim pas és traduir els conceptes a un llenguatge documental. En l'exemple, el "Tesaurus d'Història local de Catalunya". Com es pot apreciar entre l'expressió en llenguatge natural de l'usuari i els descriptors acceptats del tesaurus hi ha certes diferencies:

En l'expressió de l'usuari:	Traduït al tesaurus:
Instal·lacions esportives	Equipaments esportius
Hoquei herba	Hoquei
Terrassa	Terrassa
Olimpíades	Jocs Olímpics 1992

Traduït a una equació de cerca: Equipaments esportius AND Hoquei AND Terrassa AND Jocs Olímpics 1992.

Slype, van G. (1991, pàg. 161) considera que els llenguatges documentals poden intervenir, com a màxim, en sis moments diferents en la recuperació:

1) Selecció dels sistemes documentals que s'interrogaran: quins catàlegs, quines bases de dades, etc.

2) Selecció dels conceptes expressats per l'usuari en el seu enunciat.

3) Traducció a un llenguatge documental controlat.

4) Formulació de l'equació de cerca.

5) Extensió assistida per ordinador.

6) Avaluació final de la pertinença dels resultats obtinguts.

Hi ha una tercera funció dins la cadena documental, però només afecta un llenguatge documental concret, que són els sistemes de classificació:

Lectura complementària

Podeu ampliar la informació sobre els llenguatges documentals a l'obra següent:

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").

- La fase d'anàlisi i tractament > processament tècnic > ordenació.

Els codis numèrics dels sistemes de classificació jeràrquics, com la CDU, són l'eina per a ordenar els documents en les prestatgeries d'acord a un ordre seqüencial de les matèries (ordenació altament significativa).

En teoria, tot document es podria indexar amb qualsevol dels sis llenguatges. A la pràctica cada tipologia de SID tendeix a utilitzar un llenguatge o una combinació de llenguatges concreta.

Exemple: un document i sis indexacions

Vegeu com seria el resultat d'indexar el mateix document amb cadascun dels sis llenguatges documentals:

El mercado del tabaco en España durante el siglo XVIII: fiscalidad y consumo / Santiago de Luxán Meléndez, Sergio Solbes Ferri, Juan José Laforet (ed.). Las Palmas de Gran Canaria: Universidad de Las Palmas de Gran Canaria, Servicio de Publicaciones, 2000.

Resum:

"En este libro se ha querido poner el énfasis en un tema hasta ahora poco tratado como es el consumo de tabaco en España durante el siglo XVIII.

No obstante también se atienden otros aspectos como los fiscales. La obra se ha estructurado en tres partes: la primera se ocupa de la fiscalidad, la segunda atiende el área del monopolio y la tercera analiza los mercados regionales de Canarias y Navarra. El libro se cierra con un apartado dedicado al cultivo del tabaco."

Exemple d'un únic document i sis indexacions

Sistema de classificació: CDU	**Llistes d'encapçalaments de matèria: LEMAC**	**Llista d'autoritats: Gran Enciclopèdia Catalana**
336.226(460)"17":663.97	Indústria tabaquera- Espanya- Història – S. XVIII Tabac – Impostos – Espanya – Història – S. XVIII	Canàries Espanya Navarra
Tesaurus: Tesaurus d'Història local de Catalunya (UAB)	**Llistat de descriptors lliures: Consultors de l'assignatura**	**Llistat de paraules clau: programa Swesum**
Tabac Consum Història Impost de consums Conreus Monopolis Segle XVIII	Canàries Conreu Consum Espanya Fiscalitat Monopoli Navarra Segle XVIII Tabac	libro tabaco

Encara que en aquest moment l'estudiant no conegui el funcionament d'aquests llenguatges, si que és capaç d'observar alguns trets característics de cada un:

- El sistema de classificació ha indexat un codi, no són paraules. És un codi construït a base de números i símbols. Incomprensible a primera vista per a un profà.

- La llista d'encapçalaments de matèria ha indexat dos termes en llenguatge natural, que estan formats per diverses paraules separades amb guions.

- La llista d'autoritats ha indexat només noms geogràfics i ha prescindit de la resta de conceptes. També ha usat el llenguatge natural.

- El tesaurus ha indexat uns quants descriptors en llenguatge natural, posant un terme sota l'altre.

- El llistat de descriptors lliures no es diferencia a simple vista de la indexació amb tesaurus. En canvi, la diferència és fonamental ja que el tesaurus és controlat i els descriptors lliures no.

- El llistat de paraules clau ha indexat en castellà. Aquesta indexació no l'ha feta una persona, sinó un programa informàtic, que ha seleccionat les paraules *libro* i *tabaco* perquè surten dues vegades al text, i són les paraules més repetides.

Vegeu també

Tots aquests temes seran desenvolupats en els mòduls següents, dedicats a cada un dels llenguatges documentals.

A tall de conclusió

Els llenguatges documentals són necessaris en dos moments de la cadena documental:

- La fase d'anàlisi i tractament > anàlisi documental > anàlisi de contingut > indexació.
- La fase sortida > instruments de recuperació.

Els sistemes de classificació també són útils en:

- La fase d'anàlisi i tractament > processament tècnic > ordenació.

4.4. Complementarietat dels llenguatges documentals

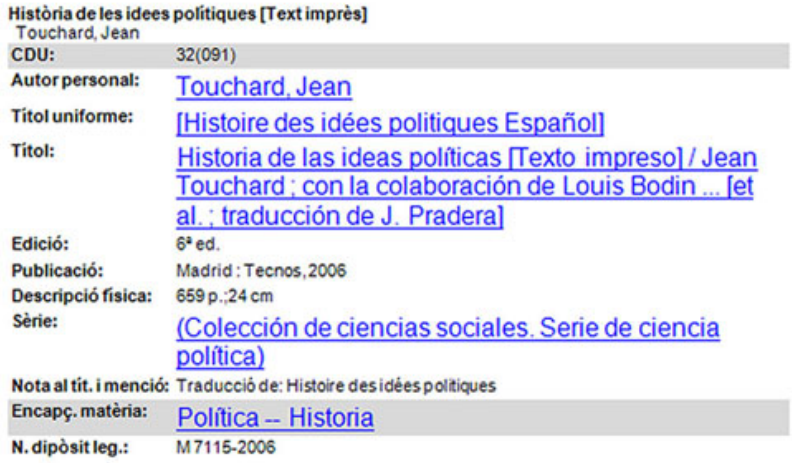
Indexar en més d'un llenguatge documental alhora és molt convenient, perquè així se sumen els avantatges i es minimitzen els inconvenients dels diferents sistemes. Significa un esforç afegit en el moment de la indexació, però permet recuperar la informació de manera més precisa. És a dir, combinem llenguatges per a recuperar millor.

Algunes de les combinacions possibles són les següents:

- Sistema de classificació + Llistes d'encapçalaments + Llistes d'autoritats.
- Sistema de classificació + Llistes d'encapçalaments + Llistes d'autoritats + Paraules clau.
- Sistemes de classificació + Tesaurus.
- Tesaurus + Llistes d'autoritats + Paraules clau.

Exemple de combinació de llenguatges

Exemple d'una captura d'un registre del catàleg de la Biblioteca Nacional d'Espanya on veiem un camp per a la notació amb CDU i un per a un encapçalament de matèria.



Història de les idees polítiques [Text imprès]
Touchard, Jean

CDU:	32(091)
Autor personal:	Touchard, Jean
Títol uniforme:	[Histoire des idées politiques Español]
Títol:	Historia de las ideas políticas [Texto impreso] / Jean Touchard ; con la colaboración de Louis Bodin ... [et al. ; traducción de J. Pradera]
Edició:	6ª ed.
Publicació:	Madrid : Tecnos, 2006
Descripció física:	659 p. ; 24 cm
Sèrie:	(Colección de ciencias sociales. Serie de ciencia política)
Nota al tít. i menció:	Traducció de: Histoire des idées politiques
Encapç. matèria:	Política -- Historia
N. dipòsit leg.:	M 7115-2006

A tall de conclusió

En teoria, tots els document es podrien indexar amb qualsevol dels sis llenguatges. A la pràctica cada tipologia de SID tendeix a utilitzar un llenguatge o una combinació de llenguatges concreta.

Indexar en més d'un llenguatge documental alhora és molt convenient, perquè així se sumen els avantatges i es minimitzen els inconvenients dels diferents sistemes. Significa un esforç afegit en el moment de la indexació, però permet recuperar la informació de manera més precisa. És a dir, es combinen llenguatges per a recuperar millor.

5. Tipologia dels llenguatges documentals

Podem classificar els sis llenguatges documentals a partir d'unes característiques o unes tipologies que els descriuen. Concretament, els llenguatges es tipifiquen segons la naturalesa dels seus termes, el nivell de control, el nivell de coordinació, l'estructura i el nivell d'anàlisi:

Tesaurus

Un llenguatge és la suma de diverses característiques. Així, per exemple, un tesaurus és natural, controlat, postcoordinat, jeràrquic i combinatori i indexa per conceptes.

Tipologia dels llenguatges documentals

		Sistemes de classificació	Llistes d'encapçalaments de matèria	Llistes d'autoritats	Tesaurus	Llistat de descriptors lliures	Llistat de paraules clau
Segons la naturalesa dels termes	**Codificat**	X					
	Natural		X	X	X	X	X
Segons el nivell de control sobre els termes	**Lliure**					X	X
	Controlat	X	X	X	X		
Segons el nivell de coordinació dels termes	**Precoordinat**	X	X				
	Postcoordinat			X	X	X	X
Segons la manera d'agrupar els termes o l'estructura	**Jeràrquic o sistemàtic**	X			X		
	Combinatori		X	X	X	X	X
Segons el nivell d'anàlisi	**Per matèries**	X	X				
	Per conceptes			X	X	X	
	Per paraules clau						X

A continuació veurem aquestes característiques.

5.1. Naturalesa del terme: codificat o natural

Els termes poden expressar-se en llenguatges codificats o naturals:

a) Llenguatges codificats. Entenem per *codificat* l'ús d'un codi artificial compost per nombres, lletres i símbols que tradueixen un concepte. Per exemple, el Sol, en un llenguatge com la CDU, seria 523.9.

Els llenguatges codificats són llenguatges sintètics, molt usats en biblioteques, ja que, a més de classificar el contingut del fons documental, són operatius en qualsevol idioma i permeten l'ordenació dels fons. D'altra banda, tenen l'inconvenient de ser poc comprensibles per part dels usuaris.

Només hi ha un tipus de llenguatge codificat: són els **sistemes de classificació**.

Alguns llenguatges codificats

Són exemples de llenguatges codificats la classificació decimal universal (CDU), la classificació Dewey (DDC), la classificació de la Library of Congress (LCC) o la classificació Colon (CC).

b) Llenguatges naturals. Entenem per *natural* l'ús de paraules del llenguatge usual, habitual, no codis. És molt més pròxim a l'usuari, més amigable. Hi ha cinc llenguatges documentals naturals:

- Les llistes d'encapçalaments de matèria.
- Les llistes d'autoritats.
- Els tesaurus.
- Els llistats de descriptors lliures.
- Els llistats de paraules clau.

Vegeu també

Els sistemes de classificació s'estudien amb més profunditat en el mòdul "Sistemes de classificació documentals" d'aquesta assignatura.

5.2. Nivell de control: lliure o controlat

Fa referència al control del vocabulari, és a dir, si les paraules seleccionades per a indexar corresponen al llenguatge natural o a un llenguatge artificial construït per a garantir la indexació i la recuperació:

a) Llenguatges lliures. Són llistes de termes extrets del llenguatge natural sense patir cap mena de control. Normalment els llenguatges lliures es fan servir en sistemes automatitzats on hi ha un fitxer invers o un diccionari de la base de dades. Tenen molts avantatges en la indexació, com ara la despesa mínima de construcció, l'actualització immediata, la coherència màxima i la riquesa terminològica. Però presenten inconvenients en la recuperació, ja que en treballar amb llenguatge natural, arrosseguen tots els problemes derivats de l'ambigüitat (sinonímia, polisèmia, homonímia).

Els llenguatges lliures són dos:

- Els llistats de descriptors lliures.
- Els llistats de paraules clau.

b) Llenguatges controlats. Considerem llenguatges controlats aquells que estan redactats prèviament en forma de llistes o llistats de termes que es consideren acceptats i unívocs per a la indexació. Només els termes de la llista es poden usar per a indexar.

Són termes seleccionats tant en la seva forma (plural, singular, sintagma nominal, adjectivat, sigles, etc.), com en el seu contingut (de tots els sinònims se n'escull un, els homònims es diferencien entre ells, etc.), com en les seves relacions de jerarquia i d'associació (termes conceptualment més genèrics o específics i termes que s'evoquen mútuament). Requereixen una feina de cons-

trucció elevada, tant en personal qualificat com en temps. Per a molts autors són els veritables llenguatges documentals. També es coneixen amb nom de *llenguatges artificials*.

La seva funció documental és la de representar un concepte amb un únic terme i que només hi hagi un terme per concepte, procés que es coneix com a *univocitat*.

Els llenguatges controlats són quatre:

- Els sistemes de classificació.
- Les llistes d'encapçalaments.
- Les llistes d'autoritats.
- Els tesaurus.

5.3. Nivell de coordinació: precoordinat o postcoordinat

a) Precoordinació. La precoordinació consisteix a determinar *a priori* com es combinen els termes, tant a l'hora de construir el llenguatge, com a l'hora d'indexar el document o a l'hora de recuperar-lo.

Llenguatges precoordinats

Un exemple de construcció amb un llenguatge precoordinat com la "Llista d'encapçalaments de matèria", preveu que la matèria *Construcció de trens* es representi com:

```
Ferrocarrils – Construcció
```

És a dir, per aquest ordre i separats amb un guió.

Un exemple d'indexació amb un llenguatge precoordinat, per exemple, d'una matèria composta per tres elements com *Enciclopèdia dels gossos d'atura europeus* es representa com:

```
Gossos d'atura – Europa – Enciclopèdies
```

L'encapçalament es fa en aquest ordre concret, i les regles sintàctiques del llenguatge eviten la possibilitat d'altres combinacions.

La precoordinació té dos gran avantatges:

- Agrupa en proximitat tots els documents que tenen una temàtica afí, de manera que si consultem el catàleg d'una biblioteca per *Ferrocarrils – Construcció*, també veurem altres documents com:
  ```
  Ferrocarrils – Construcció
  Ferrocarrils – Corbes i desviacions
  Ferrocarrils – Direcció i administració
  ```

- En un sol terme d'indexació reuneix els elements principals per a la cerca.

La precoordinació era una autèntica necessitat en l'entorn de les biblioteques manuals, ja que no es podia buscar per una combinació de dos o més termes.

Vegeu també

El tema de la precoordinació es tracta amb detall en els mòduls dedicats als dos llenguatges precoordinats: "Sistemes de classificació documentals" i "Llistes d'encapçalaments".

b) Postcoordinació. La postcoordinació consisteix a combinar els termes d'indexació en el moment de la recuperació. Permet combinar múltiples termes d'indexació seguint la lògica dels operadors booleans i d'aquesta manera aprofundir en l'anàlisi de contingut. No tenen sintaxi en el moment de la indexació. Cada terme indexat és un punt d'accés al document; com més termes indexem més possibilitat de recuperar-lo.

Llenguatges postcoordinats

Un llenguatge postcoordinat, com un tesaurus, representaria el document anterior de gossos d'atura com:

```
Gossos d'atura
Europa
Enciclopèdia
```

que serien recuperats seguint la lògica dels operadors booleans:

```
Gossos d'atura AND Europa
```

Els llenguatges postcoordinats només tenen sentit en sistemes documentals automatitzats que disposin d'un fitxer invers. El fitxer invers és on s'emmagatzemen tots els descriptors que l'analista va indexant, se situen un darrere l'altre de manera seqüencial i associats al document al qual fan referència.

Els llenguatges postcoordinats són quatre:

- Llistes d'autoritats.
- Tesaurus.
- Llistats de descriptors lliures.
- Llistats de paraules clau.

Exemple de fitxer invers

Document	Fitxer invers: concepte i núm. de document
Document 1	Alimentació (2) Enciclopèdia (1,3) Entrenament (2) Europa (1) Gossos d'atura (1,2) Química orgànica (3)
Gossos d'atura Europa Enciclopèdia	
Document 2	
Gossos d'atura Alimentació Entrenament	
Document 3	
Química orgànica Enciclopèdia	

5.4. Estructura: jeràrquica o combinatòria

El vocabulari dels llenguatges documentals s'organitza en dues estructures bàsiques, jeràrquica o combinatòria:

a) Jeràrquica: en l'estructura jeràrquica o arborescent el vocabulari es presenta en forma de cadena, amb termes genèrics que agrupen termes més específics. Tots els termes depenen d'un terme superior i de significat més genèric. Aquesta estructura permet agrupar els conceptes per temes. I també situar-los en context, ja que la seqüència jeràrquica ens informa del camp temàtic en què està adscrit el concepte.

Exemple

Posem un exemple extret de la CDU:

```
37 Educació
   371 Organització de l'educació
   372 Contingut. Matèries
   373 Tipus d'escoles
   374 Ensenyament extraescolar
   376 Escoles especials
   377 Formació professional
   378 Universitats
```

Així, el concepte "Universitats" depèn del concepte "37 Ensenyament", per tant fa referència a l'educació que s'imparteix a la universitat i no l'arquitectura de les universitats (que estaria dins "72 Arquitectura").

Els llenguatges jeràrquics són dos:

- Els sistemes de classificació.
- Els tesaurus (en la part de presentació sistemàtica o jeràrquica).

b) Combinatòria: en l'estructura combinatòria, els termes no formen cadena, i estan llistats per ordre alfabètic. Aquest tipus d'estructura va sorgir com a reacció a la rigidesa de l'estructura jeràrquica, que no era fàcil d'actualitzar.

Exemple extret de la Lista de encabezamientos del CSIC

Peajes
Pearcea
Pearl Harbor, Ataque a, 1941
Pecado
Pecado (Islam)
Pecado original
Pecados
Pecados capitales
Pecaris

L'estructura combinatòria permet la inclusió de nous termes i l'eliminació dels obsolets sense afectar la resta de l'estructura del llenguatge. La facilitat per a actualitzar el vocabulari els converteix en llenguatges adequats per a tota mena d'entorns: enciclopèdics, científics i tècnics.

Els llenguatges d'estructura combinatòria són cinc:

- Llistes d'encapçalaments de matèria.
- Llistes d'autoritats.
- Tesaurus.
- Llistats de descriptors lliures.
- Llistats de paraules clau.

Com es pot observar, els tesaurus participen de les dues estructures: tenen una presentació sistemàtica en forma jeràrquica i una presentació alfabètica en forma combinatòria.

El descriptor "Còmic"

Veiem el descriptor "Còmic" tant en una presentació com en l'altra (extret del tesaurus d'història local de Catalunya).

Presentació jeràrquica (esquerra) i alfabètica (dreta)

```
[Llengua i Literatura]

. Literatura
.. Crítica literària
.. Gèneres literaris
NA: No l'useu com a descriptor.
... Assaig
... Còmic
... Guions cinematogràfics
... Guions radiofònics
... Nadales
```

```
Comerços
USEU Botigues i comerços

Còmic
TC [Cultura i arts]
TC2 [Llengua i Literatura]
TG Gèneres literaris
TR Dibuixants Humoristes

Comissions de festes
TC [Societat]
TC2 [Vida pública i associativa]
TC3 [Recreatives i culturals]
TG Associacions
TR Associacions recreatives Festes

Comissions de propietaris
USEU Associacions de propietaris
```

5.5. Segons el nivell d'anàlisi: matèries, conceptes, paraules clau

Els llenguatges poden indexar més o menys conceptes, de manera que podem establir una darrera tipologia segons la quantitat d'informació que transmet cada un. En el punt més sintètic, amb un o dos termes d'indexació, tenim els llenguatges que indexen per matèries; en el punt intermedi, els llenguatges de conceptes, també anomenats *llenguatges de descriptors*, i en el punt més analític, els llenguatges de paraules clau.

Indexar per matèries, conceptes i paraules clau està en relació directa amb els dos paradigmes de cerca. La indexació per matèries és adequada per a sistemes de *browsing* (o de navegació o de directori). Per contra, les indexacions per conceptes i paraules clau s'adapten millor als sistemes d'interrogació en cercadors.

a) Per matèries: responen a la pregunta "quin és el tema d'aquest document?" Els llenguatges que indexen per matèries són dos:

- Els sistemes de classificació.
- Les llistes d'encapçalaments de matèria.

b) Per conceptes: indexar per conceptes significa indexar les idees i les nocions del text, sense reduir-lo a un tema principal. Responen a la pregunta: "quins són els conceptes d'aquest document?". Van lligats necessàriament a sistemes automatitzats, ja que no seria factible elaborar tantes fitxes de cartolina com conceptes s'indexessin.

Els llenguatges que indexen per conceptes són tres:

- Llistes d'autoritats.
- Tesaurus.

- Llistats de descriptors lliures.

c) Per paraules clau: indexar per paraules clau significa indexar totes i cadascuna de les paraules amb significat del text. És el procés més analític que existeix. No és una tasca d'indexació humana, sinó automàtica. Els programes que indexen per paraules clau seleccionen només les paraules que tenen significat (preferentment substantius).

Només hi ha un llenguatge per paraules clau, i és evidentment l'únic llenguatge automàtic: el llistat de paraules clau.

Exemple d'indexació amb els tres nivells d'anàlisi

Indexarem amb els tres nivells d'anàlisi el resum indicatiu següent:

MUÑOZ CRUZ, Valle. El papel del gestor de la información en las organizaciones a las puertas del siglo XXI. A. *Los sistemas de información al servicio de la sociedad: actas de las jornadas*. València: FESABID, 1998, vol. 2, p. 649-660.

"Article sobre el paper i funcions del gestor de la informació, un nou professional de la documentació, en les organitzacions del segle XXI. Descriu el panorama laboral espanyol, analitzant l'administració pública i l'empresa privada. Proposa desenvolupar una política nacional d'informació i una formació adaptada a les necessitats organitzatives de les institucions."

Per matèries	Per conceptes	Per paraules clau	
Gestor d'informació	Gestor d'informació Documentació Administració pública Empresa privada Política d'informació	Adaptada Administració Article Documentació Empresa Espanyol Formació Funcions Gestor Informació Institucions Laboral Nacional	Necessitats Nou Organitzacions Organitzatives Panorama Paper Política Privada Professional Pública Segle XXI

A tall de conclusió

Els llenguatges documentals es tipifiquen segons:

- **La naturalesa dels termes**: els termes poden expressar-se en llenguatge codificat o natural. Entenem per *codificat* l'ús d'un codi artificial compost per números, lletres i símbols que tradueixen un concepte. Entenem per *natural* l'ús de paraules del llenguatge usual, habitual, no codis.

- **El nivell de control del vocabulari**: els llenguatges poden ser lliures o controlats. Els llenguatges lliures són llistes de termes extrets del llenguatge natural. Considerem *llenguatges controlats*, aquells que estan redactats prèviament en forma de llistes o llistats de termes que es consideren acceptats i unívocs per a la indexació. Només els termes de la llista es poden usar per a indexar.

- **El nivell de coordinació**: precoordinat o postcoordinat. La precoordinació consisteix en determinar *a priori* com es combinen els termes, ja sigui a l'hora de construir el llenguatge, com a l'hora d'indexar el document, o a l'hora de recuperar-lo. La postcoordinació consisteix a no establir regles a l'hora de la indexació i combinar els termes d'indexació en el moment de la recuperació seguint la lògica dels operadors booleans.

- **L'estructura**: el vocabulari dels llenguatges documentals s'organitza en dues estructures: jeràrquica o combinatòria. En l'estructura jeràrquica o arborescent el vocabulari es presenta en forma de cadena, amb termes genèrics que agrupen termes més específics. En l'estructura combinatòria, els termes no formen cadena, estan llistats per ordre alfabètic.

- **El nivell d'anàlisi**: matèries, conceptes, paraules clau. Indexar per matèries consisteix a indexar la matèria principal del document. Indexar per conceptes significa indexar les idees i les nocions del text. Indexar per paraules clau significa indexar totes i cadascuna de les paraules amb significat del text. És el procés més analític que existeix. No és una tasca d'indexació humana, sinó automàtica.

5.6. Conclusions

L'estudi de les tipologies dels llenguatges documentals permet elaborar-ne una fitxa descriptiva individualment.

Sistemes de classificació	Llista d'encapçalament de matèries	Llista d'autoritats
• Sintètic per matèries • Símbols de classe o notacions • Humana • Codificat • Controlat • Precoordinat • Jeràrquic	• Sintètic per matèries • Encapçalaments • Humana • Natural • Controlat • Precoordinat • Combinatori	• Analític per conceptes • Identificadors i descriptors • Humana • Natural • Controlat • Postcoordinat • Combinatori
Tesaurus	**Llistat de descriptors lliures**	**Llistat de paraules clau**
• Analític per conceptes • Descriptors • Humana • Natural • Controlat • Postcoordinat • Jeràrquic • Combinatori	• Analític per conceptes • Descriptors • Humana • Natural • Lliure • Postcoordinat • Combinatori	• Analític per paraules clau • Paraules clau • Automàtica • Natural • Lliure • Postcoordinat • Combinatori

Activitats

1. A partir del següent article elabora un resum informatiu, un d'indicatiu, un de selectiu de conclusions i un d'automàtic que tingui una extensió semblant a l'informatiu.

VALLEZ, M; PEDRAZA-JIMÉNEZ, R. "El Procesamiento del Lenguaje Natural en la Recuperación de Información Textual y áreas afines" [en línia a http://www.hipertext.net/web/pag277.htm]. *Hipertext.net*, núm. 5, 2007. ISSN 1695-5498.

2. Indexeu el mateix article amb els tres nivells d'exhaustivitat. Argumenteu en quin tipus de base de dades i SID podria ser útil cadascuna.

3. Proposeu dos títols de documents, reals o inventats, en què la matèria s'expressi a través de dos sinònims.

4. Imagineu dos títols més, en què apareguin dos mots polisèmics, i proposa una manera de diferenciar-los. Busqueu l'origen etimològic de les paraules i digues si són polisèmiques o homònimes.

5. Responeu les preguntes següents justificant-ne la resposta:
a) Tot llenguatge controlat és codificat?
b) Tot llenguatge precoordinat és controlat?
c) Tot llenguatge lliure és natural?
d) El llenguatge que té la taxa de coherència més elevada és la llista de paraules clau?

6. El següent és un compendi d'errors i mitges veritats. Sabries localitzar-les i argumentar perquè no són correctes?

Usar llenguatges naturals en la indexació i la recuperació permet una bona comunicació documental. Els sistemes de classificació representen la matèria dels documents a través de notacions múltiples. Els llenguatges que indexen per matèries són els tesaurus i les llistes d'encapçalaments de matèria. Per a recuperar de manera precisa hem d'utilitzar sistemes de classificació i llistes d'encapçalaments de matèria. Els llenguatges controlats són molt amigables per a l'analista i per a l'usuari. Els llenguatges precoordinats permeten ordenar els documents a les prestatgeries.

Glossari

abstract *m* Terminologia anglosaxona per als resums redactats per persones.

anàfora *f* Relació de referència entre un element lingüístic i un d'anterior en el discurs.

anàlisi de contingut *f* Operacions d'anàlisi que identifiquen i representen de manera precisa la matèria dels documents, amb l'objectiu de permetre'n la recuperació. Les operacions són dues: el resum i la indexació. Aquesta part de l'anàlisi documental estableix el punts d'accés per matèries.

anàlisi morfosintàctica *f* Anàlisi que determina la categoria lèxica de cada paraula: substantiu, verb, adjectiu, article, preposició, etc. També en determina el lema. Aquestes operacions permeten destriar les paraules amb significat (substantius, adjectius, verbs) de les buides (articles, preposicions, pronoms, etc.). El lema permet agrupar totes les paraules que son flexions d'una altra (info/informar/informació/informador/informacional/etc.).

autoritat *f* Terme d'indexació propi del llenguatge documental llista d'autoritats. També es coneixen amb el nom d'*identificadors i descriptors*.

codificat -ada *adj* Dit del llenguatge documental que consisteix en l'ús d'un codi artificial compost per números, lletres i símbols que tradueixen un concepte.

combinatòria *f* Tipologia de llenguatge documental que consisteix a estructurar els termes d'indexació per ordre alfabètic. L'estructura combinatòria permet la inclusió de nous termes i l'eliminació dels obsolets sense afectar la resta de l'estructura del llenguatge. Els llenguatges d'estructura combinatòria són cinc: llistes d'encapçalaments de matèria, les llistes d'autoritats, els tesaurus, llistat de descriptors lliures i llistat de paraules clau.

controlat -ada *adj*. Tipologia de llenguatge documental que consisteix en llistes de termes seleccionats tant en la seva forma (plural, singular, sintagma nominal, adjectivat, sigles, etc.) com en el seu contingut (de tots els sinònims se n'escull un, els homònims es diferencien entre ells, etc.), com en les seves relacions de jerarquia i d'associació (termes conceptualment més genèrics o específics i termes que s'evoquen mútuament). Requereixen d'unes despeses de construcció elevades, tant en personal qualificat com en temps. Són els veritables llenguatges documentals. També es coneixen pel nom de llenguatges artificials. La seva funció documental és la de representar un concepte amb un únic terme i que només hi hagi un terme per concepte, el que es coneix com a *univocitat*. Els llenguatges controlats són quatre: els sistemes de classificació, les llistes d'encapçalaments, llistes d'autoritats i els tesaurus.

descripció característica *f* Vegeu **indexació**.

descriure el contingut *loc v* Vegeu **representar el contingut**.

descriptor *m* Terme d'indexació propi de tres llenguatges documentals llista d'autoritats, tesaurus, llistat de descriptors lliures.

encapçalament *m* Terme d'indexació propi del llenguatge documental llistes d'encapçalaments de matèria.

entropia *f* Qualitat aplicable als llenguatges documentals que tendeixen a la selecció, a la restricció del vocabulari. Es el procés contrari del llenguatge natural que tendeix a l'abundància, a la reiteració de conceptes, a la sinonímia en benefici d'una expressió més rica.

especificitat *f* Criteri relacionat amb l'exactitud en què un concepte particular que apareix en un document està representat per un terme d'indexació.

estructura *f* Tipologia dels llenguatges documentals que els classifica en jeràrquics o combinatoris.

examen del document *m* Primera fase en el procés d'indexació que consisteix en la lectura del títol, del resum, del sumari, de la introducció, de les il·lustracions i de les paraules o de les frases destacades en una tipografia diferent.

exhaustivitat *f* Criteri relacionat amb el nombre de conceptes que es tenen en compte per a caracteritzar el contingut sencer d'un document. El principal criteri de selecció és el valor potencial del concepte per als usuaris del seu SID. Podem distingir entre una exhaustivitat baixa, mitjana i alta en funció del nombre de descriptors.

extract *m* Terminologia anglosaxona per als resums automàtics. Els *extracts* són els resums formats a partir de l'extracció d'algunes frases del text prèviament seleccionades per un programa.

fitxer invers *m* Fitxer on s'emmagatzemen tots els termes d'indexació. Aquests se situen un darrere l'altre de manera seqüencial i associats al document al qual fan referència.

hiperònim *adj.* Paraula que té un camp significatiu que inclou un altre de menor extensió. Per exemple, *color* és un hiperònim respecte a *groc*, *taronja* i *verd*.

hipònim -a *adj.* Dit de la paraula que té un camp significatiu que queda inclòs en un altre de més extensió. Per exemple, *groc*, *taronja* i *verd* són hipònims, ja que pertanyen al terme *color*.

homonímia *f* Tipus de polisèmia que es dóna quan dos conceptes diferents han arribat a tenir el mateix nom, la mateixa forma, però tenen orígens diferents i, per tant, etimologies diferents.

identificador *m* Terme d'indexació propi del llenguatge documental llista d'autoritats. També es coneixen amb el nom d'*autoritat* i *descriptors*.

indexació *f* Descripció o identificació d'un document amb relació al seu contingut. Norma UNE 50-121-91. Indexar és el resultat d'examinar el document, seleccionar els conceptes i emmagatzemar-los en una base de dades. Aquesta definició implica tres accions, de les quals la més significativa és la de la selecció dels conceptes i la seva traducció al llenguatge documental.

indexar per conceptes *loc v* Indexar les idees i nocions del text, sense reduir-lo a un tema principal. Responen la pregunta "Quins són els conceptes d'aquest document?", van lligats necessàriament a sistemes automatitzats. Els llenguatges que indexen per conceptes són tres: llistes d'autoritats, tesaurus, llistes de descriptors lliures.

indexar per matèries *loc v* Indexar de manera sintètica. Responen la pregunta "Quin és el tema d'aquest document?". Els llenguatges que indexen per matèries són dos: els sistemes de classificació i les llistes d'encapçalaments de matèria.

indexar per paraules clau *loc v* Indexar totes i cadascuna de les paraules amb significat del text. És el procés més analític que hi ha. No és una tasca d'indexació humana, sinó automàtica. Els programes que indexen per paraules clau seleccionen només les paraules que tenen significat (preferentment substantius). Només hi ha un llenguatge per paraules clau, i és evidentment que és l'únic llenguatge automàtic: el llistat de paraules clau.

ISO 214:1976 *f* Norma internacional, traduïda per AENOR com a norma UNE 50-103-90 *Preparació de resums*.

jeràrquic –a *adj.* Dit del llenguatge documental que consisteix a estructurar els termes d'indexació de forma arborescent. El vocabulari es presenta en forma de cadena, amb termes genèrics que agrupen termes més específics. Tots els termes depenen d'un terme superior i de significat més genèric. Aquesta estructura permet agrupar els conceptes per temes.

llenguatge documental *m* Vocabulari de termes en llenguatge natural o un sistema artificial de signes normalitzats que faciliten la representació del contingut dels documents. Les seves funcions principals són indexar el contingut dels documents i permetre'n la recuperació a partir del camp matèria.

llenguatge natural *m* Llenguatge que usem quotidianament per a comunicar-nos.

llenguatge artificial *m* *Vegeu* controlat.

llista d'autoritats *m* Llenguatge documental. Analític per conceptes, natural, controlat, postcoordinat i combinatori. El seu terme d'indexació es coneix com a *identificador*, *autoritat* o *descriptor*.

llistat de descriptors lliures *m* Llenguatge documental. Analític per conceptes, natural, lliure, postcoordinat i combinatori. El seu terme d'indexació es coneix com a *descriptor*.

llista d'encapçalaments de matèria *m* Llenguatge documental. Sintètic per matèries, natural, controlat, precoordinat i combinatori. El seu terme d'indexació es coneix com a *encapçalament*.

llistat de paraules clau *m* Llenguatge documental. Analític per paraules clau, natural, lliure, postcoordinat i combinatori. El seu terme d'indexació es coneix com a *paraula clau*.

lliure *adj* Dit del llenguatge documental que consisteix en llistes de termes extrets del llenguatge natural sense formar part de cap llista establerta *a priori*, ni haver passat un procés de control del seu vocabulari.

natural *adj* Dit del llenguatge documental que consisteix en l'ús de paraules del llenguatge usual, habitual, no codis. Hi ha cinc llenguatges documentals naturals: les llistes d'encapçalaments de matèria, les llistes d'autoritats, els tesaurus, les llistes de descriptors lliures i les llistes de paraules clau.

naturalesa dels llenguatges *f* Tipologia dels llenguatges documentals que els classifica en codificats o naturals.

nivell d'anàlisi *m* Tipologia dels llenguatges documentals que els classifica en llenguatges de matèries, conceptes i paraules clau.

nivell de control *m* Tipologia dels llenguatges documentals que els classifica en lliures o controlats.

nivell de coordinació *m* Tipologia dels llenguatges documentals que els classifica en precoordinats o postcoordinats.

notació *f* Terme d'indexació propi del llenguatge documental dels sistemes de classificació.

paraula clau *f* Terme d'indexació propi del llenguatge documental de les paraules clau o indexació automàtica. Paraula o grup de paraules seleccionades de manera automàtica del títol, resum o text d'un document que en representen el contingut i en permeten la recuperació.

paraula buida *f* Paraula sense significat en les operacions d'indexació i resum. Per exemple, preposicions, articles, verbs, adverbis, etc.

PLN *f* Vegeu **processament del llenguatge natural**.

polisèmia *f* Propietat d'un signe lingüístic de tenir més d'un significat. Diem que dues paraules són polisèmiques quan el mateix signe lingüístic, paraula o so té més d'un significat. La paraula té un únic origen etimològic i acaba tenint significats diferents sense canviar-ne la categoria gramatical.

ponderació (de frases, de paraules) *f* Mètode que avalua les frases i les paraules d'un text en funció de paràmetres com la freqüència, la presència de paraules indicatives (busquen paraules com ara *important, essencial, conclusions*, etc.), l'aparició en llocs destacats com ara el títol, l'inici de cada paràgraf, el final a mode de conclusions, etc.

postcoordinació *f* Tipologia de llenguatge documental que consisteix a combinar els termes d'indexació en el moment de la recuperació. Els llenguatges postcoordinats només tenen sentit en sistemes documentals automatitzats que disposin d'un fitxer invers. Els llenguatges postcoordinats són quatre: llistes d'autoritats, tesaurus, llistes de descriptors lliures i llistes de paraules clau.

precoordinació *f* Tipologia de llenguatge documental que consisteix a determinar *a priori* com es combinen els termes, ja sigui a l'hora de construir el llenguatge, a l'hora d'indexar el document, o a l'hora de recuperar-lo. Els dos llenguatges precoordinats són els sistemes de classificació i les llistes d'encapçalaments de matèria.

processament en llenguatge natural *m* Branca de la intel·ligència artificial i de la lingüística computacional que estudia els llenguatges que usen els humans per a interactuar amb els ordinadors en contextos escrits i orals. EL PLN estudia com emular el coneixement humà, quant a la identificació dels conceptes i frases amb contingut rellevant.
sigla **PLN**

relació de significat *f* Vegeu **relació semàntica**.

relació semàntica *f* Relació de significat de les paraules. Les relacions poden ser de tipus genèric, específic o relacionat d'un terme respecte a un altre. En llenguatge natural aquestes relacions són implícites però en un llenguatge documental cal definir aquestes relacions, agrupant i relacionant els termes afins.

representar el contingut *loc v* Descriure el tema o els temes d'un document.

resum *m* Presentació abreujada i precisa d'un document, sense interpretació ni crítica i sense menció expressa de l'autor del resum. Norma UNE 50-103-90. *Preparació de resums*.

resum indicatiu *m* Resum que consigna només les idees centrals del document i la lectura del qual no pot substituir la de l'original.

resum informatiu *m* Resum que consigna el tema central, els temes addicionals, la naturalesa i l'objectiu del document, la metodologia, els resultats, les conclusions i els annexos. La idea de fons és que un resum informatiu pot substituir en algunes ocasions la lectura del document original.

resum selectiu *m* Resum que consigna només una part concreta del document. El més habitual és el resum de conclusions, però també hi ha altres tipus com la ressenya (*review*).

selecció dels termes d'indexació *f* Segona fase en el procés d'indexació que consisteix a identificar les nocions que són elements essencials de la descripció del contingut. Els criteris de selecció són el nombre de conceptes (criteri d'exhaustivitat) i la seva exactitud (criteri d'especificitat).

símbol de classe *m* Vegeu **notació**.

sinonímia *f* Paraules que tenen el mateix significat. Per exemple, *aliment*, *nutrient*, *menjar* i *provisió*. En un sistema documental si no es controlen i s'usen indiscriminadament, comporten silenci documental.

sistema de classificació *m* Llenguatge documental. Sintètic per matèries, Codificat, Controlat, Precoordinat i Jeràrquic. El seu terme d'indexació es coneix com a *notació* o *símbol de classe*.

terme d'indexació *m* Representació d'un concepte en llenguatge natural o un codi de classificació. Els termes d'indexació poden estar formats per una paraula o més d'una.

tesaurus *m* Llenguatge documental. Analític per conceptes, natural, controlat, postcoordinat i jeràrquic i combinatori. El seu terme d'indexació es coneix com a *descriptor*.

traducció a un llenguatge documental controlat *f* Cerca d'un concepte expressat en llenguatge natural en la llista de termes d'un llenguatge documental controlat i ús del terme controlat per a indexar i recuperar.

UNE 50-103-90. *Preparació de resums* *f* Norma espanyola que estableix les directrius que s'han de seguir per a presentar els resums en els documents. Posa un èmfasi especial en la preparació de resums per part dels autors dels documents primaris i en la mateixa publicació.

UNE 50-113-92/1 *f* Norma espanyola titulada "Documentación e información. Vocabulario. Parte 1. Conceptos fundamentales". A: *Documentación: Normas fundamentales*. Madrid: AENOR, 1994.

UNE 50-121-91 *f* Norma espanyola titulada *Mètodes per a l'anàlisi de documents, determinació del seu contingut i selecció de termes d'indexació*. Basa el procés d'indexació en tres fases: examinar el document per a identificar-ne el contingut, seleccionar els conceptes principals del contingut i traduir a un llenguatge documental.

uniterme *m* La part més petita amb significat d'un terme d'indexació. La norma UNE 50-113-92/1 defineix els unitermes com l'element significatiu més petit d'un llenguatge documental utilitzat per a representar un concepte específic en un sistema d'indexació coordinat. No s'ha de confondre amb paraula clau o descriptor.

univocitat *f* Representació d'un concepte amb un únic terme.

Bibliografia

Bibliografia sobre el resum

AENOR (1990). *Documentación. Preparación de resúmenes. UNE 50 103 90.* Madrid: AENOR.

Climent, S. (2001). "Sistemes de resum automàtic de documents". *Digit. Hum. Revista Digital D'humanitats.* ISSN 1575-2275.

Lloret, E.; Ferrández, O.; Muñoz, R.; Palomar, M. (2008). "Integración del reconocimiento de la implicación textual en tareas automaticas de resúmenes de textos". *Procesamiento del Lenguaje Natural* (núm. 41, pàg. 183-190).

Mateo, P. L.; González, J. C.; Villena, J; Martínez, J. L. (2003). "Un sistema para resumen automático de textos en castellano".

Pinto Molina, M. (1992). *El resumen documental: principios y métodos*. Madrid: Pirámide / Fundación Germán Sánchez Ruipérez ("Biblioteca del Libro", Y).

Bibliografia sobre la indexació

Abadal, E.; Codina, Ll. (2005). "Recuperación de Información". A: *Bases de datos documentales: características, funciones y método* (cap. 2. pàg. 29-92). Madrid: Síntesis.

AENOR (1997). *Métodos para el análisis de los documentos, determinación de su contenido y selección de los términos de indización. Norma UNE 50-121-91.* Madrid: AENOR.

AENOR (1997). "Documentación e información. Vocabulario. Parte 6: lenguajes documentales". A: *Revista Española de Documentación Científica*, Norma UNE-50-113/6 (ISO 5127/6), vol. 20, núm. 4, pàg. 417-436.

Cid, P.; Cuadrado, M.; Aguiriano, C. (1999). *Fonaments de llenguatges documentals*. [Document electrònic]. Barcelona: UOC.

Codina, Ll. (1994). "El papel del lenguaje natural en los sistemas multimedia: una reflexión sobre la tecno-simpleza y la ciber-ingenuidad". A: *Cuadernos de documentación multimedia*, núm. 3 (junio).

Gil Leiva, I. (2008). *Manual de indización. Teoría y práctica*. Gijón: Ediciones Trea ("Biblioteconomía y Adminitración Cultural", 193).

Gil, I.; Rodríguez Muñoz, J. V. (1996). "El Procesamiento del lenguaje natural aplicado al análisis del contenido de los documentos". *Revista General de Información y Documentación* (vol. 6, núm. 2, pàg. 205-218).

Gil Urdiciain, B. (1992). "Función de los lenguajes documentales en el tratamiento de la información en las organizaciones". *Revista General de Información y Documentación* (vol. 2, núm. 2, pàg. 195-200).

Gil Urdiciain, B. (2004). *Manual de lenguajes documentales*. Gijón: Ediciones Trea ("Biblioteconomía y Adminitración Cultural", 106).

Norma UNE 50-113-92/1. *Documentación e información. Vocabulario. Parte 1. Conceptos fundamentales* (1994). A: *Documentación: Normas fundamentales.* Madrid: AENOR.

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").