Llenguatges documentals: indexació, recuperació i avaluació

Manela Juncà Campdepadrós

PID_00193275

UOC
Universitat Oberta de Catalunya

Índex

Introducció

Aquest primer mòdul està estructurat en dos apartats dedicats respectivament a indexar i a indexar bé.

El primer apartat comença fent una repassada de les fases de la indexació i la recuperació, que són tres: l'examen del document o pregunta, la selecció de conceptes principals i la traducció a un llenguatge documental.

A continuació, tracta breument de la manera d'esbrinar quin llenguatge o combinació de llenguatges documentals indexen una font d'informació. Conèixer el llenguatge que hi ha al darrere d'una font ens permetrà fer cerques més precises, ja que usarem el nostre coneixement sobre termes equivalents i relacions semàntiques per a obrir o tancar el zoom de la nostra cerca.

Així mateix, tracta del paper de l'indexador, ja sigui un professional que indexa amb llenguatges controlats, un amateur que indexa amb etiquetes o, finalment, un algoritme que indexa automàticament.

L'apartat finalitza amb un resum dels aspectes terminològics més rellevants dels llenguatges documentals: els termes d'indexació que reben diversos noms segons cada llenguatge i les tipologies que ens permeten classificar-los en grups disjunts segons la naturalesa del terme, el grau de control, de coordinació, d'estructura i d'anàlisi.

El segon apartat està dedicat a indexar bé, és a dir, a mesurar la qualitat de la indexació i la recuperació. Tracta dels tres elements avaluables (els llenguatges, l'indexador i el resultat de la indexació), i també dels criteris que s'usen en cada cas. Es descriuen les característiques estructurals d'un bon llenguatge documental, els errors tècnics i ètics que pot cometre un indexador i es faciliten les fórmules per tal de calcular les taxes de coherència, silenci, soroll, exhaustivitat i precisió.

Objectius

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

1. Identificar les principals etapes en el procés d'indexació i recuperació: examen del document o pregunta, selecció de conceptes i traducció al llenguatge documental.

2. Reconèixer el llenguatge o la combinació de llenguatges que indexen una font d'informació.

3. Diferenciar els tres tipus d'indexadors: el professional, l'amateur i l'automàtic.

4. Interioritzar les tipologies dels llenguatges per poder-los comparar i combinar.

5. Avaluar la qualitat del llenguatge documental, el professional indexador i la indexació.

6. Ser conscients que la indexació comporta uns compromisos de qualitat i ètica professional.

7. Calcular les taxes de consistència, silenci, soroll, exhaustivitat i precisió.

8. Conèixer les implicacions de la manca d'especificitat d'un llenguatge en la recuperació.

1. Els llenguatges documentals en el procés d'indexació i recuperació

Aquest apartat té com a objectiu avivar la memòria dels estudiants sobre alguns conceptes fonamentals per a l'assignatura: en primer lloc, sobre les fases de la indexació, com se seleccionen els conceptes principals i en què consisteix la traducció a un llenguatge documental. En segon lloc, sobre la fase de recuperació i la implicació dels llenguatges en la fase de cerca d'informació. En tercer lloc, sobre el paper de l'indexador i els diversos tipus que hi ha. I, finalment, sobre els termes d'indexació i les diferents tipologies de llenguatges documentals.

1.1. La indexació

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

Norma UNE 50-121-91.

La indexació és el resultat d'aplicar aquestes tres fases:

1) Examinar el document per tal d'identificar-ne el contingut.
2) Seleccionar els conceptes principals del contingut.
3) Traduir a un llenguatge documental.

UNE 50-121-91

Seguim la norma UNE 50-121-91 i les seves tres etapes. UNE 50-121-91: Mètodes per a l'anàlisi de documents, determinació del seu contingut i selecció de termes d'indexació.

1.1.1. Examinar el document

Lectura recomanada

Alice Keefer (2007). "Los repositorios digitales universitarios y los autores" [en línia]. *Anales de Documentación* (núm. 10, pàg. 205-214). [Data de consulta: 23 de setembre de 2011].
<http://revistas.um.es/analesdoc/article/viewFile/1151/1201>

A partir de l'article d'Alice Keefer titulat "Los repositorios digitales universitarios y los autores" farem una explicació exemplificada.

Resum informatiu de l'article

Article científic sobre els repositoris digitals universitaris i els autors en el context del moviment de l'*open acces* (OA).

L'OA va començar com una resposta a la crisi de les revistes científiques. Els seus antecedents immediats són les iniciatives d'Arxiv, SPARC, Public Library of Science i ARL, però aviat es va transformar en un moviment que demanava l'accés gratuït al coneixement científic. El moviment es va perfilar entre el 2002 i 2003. La declaració de Budapest de 2002, que plantejava dues rutes –la ruta daurada (revistes en obert) i la ruta verda (repositori dels autors o autoarxiu)–, es considera el començament oficial de l'OA.

Els repositoris van representar una gram empenta a l'OA per la via verda. Entre 2003 i 2004 van créixer un 60% els de tipus temàtic i un 100% els de caràcter institucional. A més, també servien per a la gestió institucional i la preservació a llarg termini.

Els avantatges de l'autoarxiu per als autors són diversos: tenir més accés a treballs científics, més visibilitat i citacions. Així i tot, n'hi ha que es mostren reticents a col·laborar-hi. En l'article hi ha recollits una sèrie de motius.

Per a captar més autors, les estratègies passen per difondre l'experiència d'èxit de companys (augment de citacions), ajudar en els procediments d'autoarxiu per evitar despeses de temps i mitjans, concedir més finançament mitjançant una ordre institucional.

L'autora conclou que sense una promoció activa els repositoris tardaran molts anys a captar un percentatge important dels treballs dels seus investigadors, i la meta de l'accés lliure tardaria a assolir-se.

Com que es tracta d'un article curt, unes deu pàgines, en llegim el títol, el resum i el text sencer. En el cas de documents extensos que no es puguin llegir totalment, com les monografies o tesis, es recomana prestar atenció al títol, el resum, el sumari, la introducció, les il·lustracions i les paraules o frases destacades en una tipografia diferent.

1.1.2. Seleccionar els conceptes principals

Per a seleccionar el nombre de conceptes del document (**criteri d'exhaustivitat**), l'analista ha de ser conscient dels objectius del seu servei d'informació (SID) i les necessitats dels seus usuaris: Un SID generalista o enciclopèdic, com una biblioteca pública, escollirà una exhaustivitat baixa, mentre que un SID especialitzat, com un centre de documentació o una biblioteca especialitzada, escollirà una exhaustivitat mitjana o alta.

Quant al tipus de conceptes per seleccionar, una bona praxi identificaria els elements següents:

- el tema,
- els noms personals que puguin ser interessants d'indexar,
- els noms geogràfics,
- les dates cronològiques,

- la manera en què es presenta el document (article, estadística, formulari o divulgació, científic, etc.).

A continuació proposem tres indexacions sobre el mateix article per a SID diferents i adjuntem en la casella inferior quin o quins llenguatges documentals usaríem.

Graus d'exhaustivitat

Exhaustivitat baixa	**Exhaustivitat mitjana**	**Exhaustivitat alta**
Barem 1-3	**Barem 4-6**	**Barem 7-...**
Exemple d'ús: catàleg d'una biblioteca pública	Exemple d'ús: bases de dades d'una biblioteca especialitzada o centre de documentació en documentació	Exemple d'ús: bases de dades d'una biblioteca especialitzada o centre de documentació en accés obert
Repositoris digitals universitaris	• *Open acces* • Ruta daurada • Ruta verda • Repositoris temàtics • Repositoris institucionals • Autoarxiu	• Repositoris digitals universitaris • Coneixements científics • Declaració de Budapest 2002 • Ruta daurada • Ruta verda • Repositoris temàtics • Repositoris institucionals • Preservació de fons digitals • Autoarxiu • Citacions • Investigació • Divulgació
Sistemes de classificació Llistes d'encapçalaments de matèria	• Llistes d'autoritats • Tesaurus • Llistes de descriptors lliures • Llista de paraules clau	

Observació

Fixeu-vos que en la tercera columna no s'indexen termes com *open acces*, *accés gratuït* o *accés obert*, perquè tota la base de dades té aquesta temàtica i, per tant, seria redundant.

1.1.3. Traduir els conceptes a un llenguatge documental

Traduïm el concepte inicial (escrit en llenguatge natural) a un llenguatge documental. Per a fer-ho correctament cal conèixer el **criteri d'especificitat** i també la sintaxi del llenguatge documental que utilitzarem.

Observació

Fixeu-vos que, de tot el procés, és en la selecció i en la traducció on hem de tenir més cura. En la selecció, el criteri és l'exhaustivitat i, en la traducció, és l'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.

En l'article que estem indexant hi apareix el concepte *universitat*, si indexem *sistemes educatius* o *escola politècnica* no serem específics: en el primer cas, perquè serem massa genèrics i, en el segon, perquè serem massa específics. Si el nostre article parla d'universitats i el llenguatge documental ens ho permet, indexarem *universitats* i no les altres dues opcions.

Termes més genèrics d'indexació

Malgrat que la norma ens diu que hem d'identificar de la manera més específica possible, de vegades es prefereixen termes més genèrics.

- Quan l'indexador consideri que un excés d'especificitat pot ser negativa en la recuperació. Per exemple: pot decidir que una assignatura molt específica d'un grau s'indexi amb un nom més genèric d'aquest tipus de disciplines.
- Quan la idea no estigui plenament desenvolupada en el document, o només s'hi faci al·lusió.
- Quan s'estigui a l'espera de validar el terme més específic.

La **sintaxi del llenguatge** fa referència només als llenguatges precoordinats (sistemes de classificació i llistes d'encapçalaments de matèria) i consisteix en les normes pròpies de cadascun a l'hora de combinar i ordenar les parts del terme d'indexació.

Exemple

Per a la sintaxi de la CDU, les classes i auxiliars van en aquest ordre:

classe principal + aux. de lloc + aux. de temps + aux. de forma + aux. de llengua + aux. de raça

A continuació, traduïm els conceptes seleccionats a cadascun dels sis llenguatge documentals.

Traducció als sis llenguatges documentals

De manera controlada				De manera lliure	
CDU	LEMAC	Llista d'autoritats	Tesaurus	Llista de descriptors lliures	Llista de paraules clau
004.65:027.7	Biblioteques digitals Biblioteques universitàries	University of Southampton University of Edinburgh	Biblioteques universitàries Fonts d'informació Documents electrònics Universitats Documentació Bases de dades	Accés obert Autoarxiu Autors Declaració de Budapest Repositoris digitals Revistes científiques Ruta daurada Ruta verda Universitats	Autor Repositorio Trabajo Científico Institución

1.2. La recuperació

La recuperació és un procés paral·lel a la indexació.

Si se cerca una dada concreta, com un títol (Hamlet, web semàntica) o un autor (Shakespeare, Lluís Codina), la cerca no representa cap dificultat, ja que la demanda es fa amb unes dades objectives i la resposta només pot ser "tinc resultats o no tinc resultats"; en canvi, quan no se cerca per una dada concreta, sinó per un tema, llavors entren en joc les mateixes tres fases (examen, selecció i traducció[1]) que en la indexació, però amb la diferència que el que s'examina i selecciona és la demanda de l'usuari.

(1)La traducció és idèntica a la de la fase d'indexació.

1) Examinar la demanda de l'usuari per identificar-ne el contingut.

2) Seleccionar els conceptes principals de la demanda.

3) Traduir a un llenguatge documental.

En la recuperació una de les claus és conèixer bé el llenguatge documental que hem de consultar, perquè si el coneixem podrem cercar amb més precisió, sobretot en el cas de llenguatges controlats (per les relacions semàntiques que estableixen entre els termes). El primer pas serà, doncs, esbrinar quin tipus d'indexació hi ha al darrere de la caixeta de cerca.

Els llenguatges documentals que hi ha al darrere d'una font d'informació no són evidents, tendeixen a la invisibilitat. Els programes prefereixen pantalles de cerca molt simples (per exemple, Scirus), en què apareix una caixeta en blanc. Senzill i amigable per a l'usuari, però que als nostres ulls no pot amagar que al darrere hi ha un llenguatge documental. O més probablement una combinació de llenguatges.

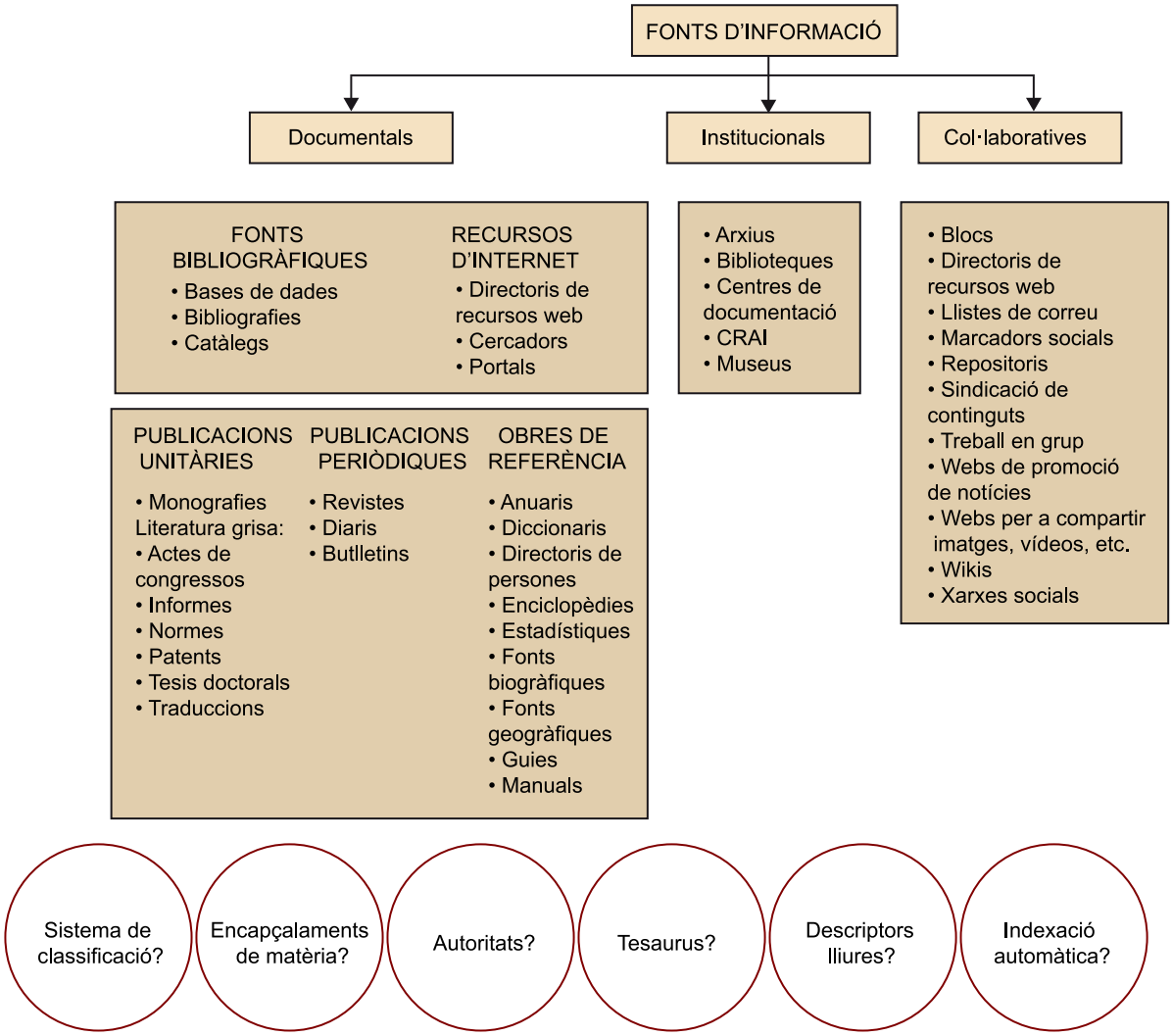
En el procés de cerca probablement passarem d'una font d'informació a una altra i, en conseqüència, d'un tipus d'indexació a un altre.

Mentre la cerca es faci en cercadors, la indexació serà **automàtica i lliure**, però quan entrem en intranets i bases de dades, la indexació canviarà, probablement, a una de **controlada**. Llavors, caldrà saber amb quin tipus de llenguatge.

Exemple

Usem un cercador general com Google (indexació automàtica) per a arribar al web de la Biblioteca de Catalunya i al seu catàleg, que està classificat amb CDU i LEMAC i LENOTI (tres llenguatges controlats).

Figura 1. Fonts d'informació i llenguatges documentals



Observació

No es pot dissenyar una taula en què relacionem el tipus de fonts d'informació i el llenguatge que utilitzen perquè, tot i que hi ha una certa tendència, no és sempre igual.

Les fonts d'informació més estàndard són els **catàlegs bibliotecaris** (que acostumen a estar indexats amb sistemes de classificació, llistes d'encapçalaments de matèria i llistes d'autoritats) i els **cercadors**, que no podrien existir sense la indexació automàtica. Ara bé, la resta és ben diversa i podem trobar bases de dades indexades per tesaurus (Unesco) o simplement per descriptors lliures (Delicious).

Per a saber quin llenguatge indexa la font, és útil observar si porta un menú d'opcions amb enllaços del tipus "Normalització", "per a professionals", o bé directament LEMAC o LCSH, és a dir, el nom del llenguatge que un profà no reconeixerà, però que els documentalistes sí que ho poden fer.

En segon terme podem reconèixer el llenguatge...

- per la forma del terme (un codi serà una classificació, dues paraules separades per guió serà un encapçalament de matèria);
- per un nombre de termes en plural (ens diu que es tracta de descriptors, caldrà esbrinar si són controlats (d'un tesaurus) o lliures (descriptors lliures o *tags*);
- pel tipus de font (un catàleg o un cercador usen sempre el mateix tipus de llenguatge);
- per la institució que hi ha al darrere;
- per l'experiència del documentalista.

1.3. L'indexador

La indexació es pot dur a terme de manera intel·lectual o de manera automàtica. En la indexació intel·lectual hem de diferenciar la feta per un professional (un documentalista) i la d'un amateur (usuari d'Internet que indexa de manera social o *tagging*).

- La **indexació intel·lectual feta per un professional** és unívoca i específica, capta els matisos i distingeix entre temes principals i temes col·laterals. És una indexació de qualitat, selectiva però també és lenta, cara.

- La **indexació intel·lectual feta per amateurs** és subjectiva, però pot indexar una gran quantitat de documents que fins ara quedava exclosa de la recuperació: les imatges fixes (fotografies) o en moviment (vídeo) que no duen text. No indexen amb criteris professionals d'exhaustivitat i especificitat però ofereixen quantitat, diversitat i un aspecte molt interessant: usen el vocabulari de l'usuari.

- La **indexació automàtica** és capaç d'assumir quantitats ingents de documentació i és ràpida i barata, però és exhaustiva en excés i es pot tornar inoperant amb els seus resultats desbordants. Aquí el problema no és com s'indexa sinó com es presenten el resultats obtinguts.

Tots tres mètodes s'usen conjuntament, ja que maximitzem els avantatges i minimitzem els inconvenients entre ells. Cada una d'aquestes opcions comporta l'ús d'un llenguatge documental o d'un altre:

Tipus d'indexació, indexació i llenguatge corresponent

Intel·lectual		**Automàtica**
Professional	Amateur	
• Sistemes de classificació • Llistes d'encapçalaments de matèria • Llistes d'autoritats • Tesaurus	Llista de descriptors lliures	Llista de paraules clau

1.4. El llenguatge documental

Per a indexar necessitem els llenguatges documentals, que són vocabularis de termes que faciliten la representació del contingut dels documents.

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

Tercera funció dels llenguatges documentals

Hi ha una tercera finalitat que només es dóna en els sistemes de classificació i que és l'ordenació altament significativa del fons documental del SID.

De llenguatges documentals n'hi ha sis:

1) Els sistemes de classificació
2) Les llistes d'encapçalaments de matèria
3) Les llistes d'autoritats
4) Els tesaurus
5) Les llistes de descriptors lliures
6) Les llistes de paraules clau o indexació automàtica

1.4.1. Els termes d'indexació

Cada llenguatge documental dóna un nom diferent al seu terme d'indexació i és convenient que, quan ens expressem, ho fem correctament.

Termes d'indexació

Llenguatge documental	El seu terme d'indexació es coneix com a	Exemple
Sistemes de classificació	Notació o símbol de classe	351.851:069 (Llei de Museus)
Llistes d'encapçalaments de matèria	Encapçalament	Francès – Argot
Llistes d'autoritats	Autoritat, identificador o descriptor	Bécquer, Gustavo Adolfo, 1836-1870
Tesaurus	Descriptor	Ramon Berenguer III el Gran NA: [1097-1131]
Llistes de descriptors lliures	Descriptor	Setmana_santa
Llistes de paraules clau	Paraula clau	Metro

Hi ha un altre terme, anomenat **uniterme**, que no fa referència a cap llenguatge documental concret, sinó al fet que el terme d'indexació sigui simple o compost.

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.

El descriptor *Setmana Santa* està format per dos unitermes: *Setmana* i *Santa*. I el descriptor *Nadal* està format per un únic uniterme.

Diferència entre descriptor i uniterme

Una paraula	Més d'una
Nadal	Setmana Santa

Cal fer atenció al terme **paraula clau** perquè el seu ús en la bibliografia científica té diverses aplicacions que ens poden confondre. És habitual trobar en els articles un apartat, sota el resum, anomenat *paraules clau*, en què l'autor ens dóna els termes que considera més representatius del text. Aquestes paraules clau són molt sovint descriptors de procedència desconeguda (desconeixem si són lliures o controlats). En canvi, en aquest material docent *paraula clau* s'entén com el terme d'indexació provinent de la indexació automàtica habitualment coincident amb un uniterme.

1.4.2. Les tipologies dels llenguatges documentals

Les tipologies dels llenguatge documentals són els criteris que ens permeten agrupar o classificar els sis llenguatge documentals en categories afins. Són les següents:

Reflexió

Si domineu les tipologies podreu respondre a preguntes del tipus: compareu llenguatges, busqueu avantatges i inconvenients, causes de la complementarietat, etc. Es recomana que les interioritzeu.

1) Naturalesa: codificat o natural

Per **codificat** entenem l'ús d'un codi artificial compost de nombres, lletres i símbols que tradueixen un concepte. Només hi ha un tipus de llenguatge codificat: els sistemes de classificació.

Exemples de termes d'indexació codificats

CDU	DDC	LCC
94	483	RE 1-994

Per **natural** entenem 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, les llistes de descriptors lliures i les llistes de paraules clau.

Seguint l'exemple anterior:

Exemples de termes d'indexació naturals

Història	Diccionaris de grec clàssic	Oftalmologia

2) Control: lliure o controlat

Un vocabulari lliure és una llista de termes extrets del llenguatge natural sense patir cap mena d'actuació sobre el nombre de termes, forma (singular, plural, masculí, femení), significat (sinònim, polisèmic) o relacions entre termes.

Normalment, els llenguatges lliures es fan servir en sistemes automatitzats en què hi ha un fitxer invers o 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. Ara bé, presenten inconvenients en la recuperació, ja que en treballar amb llenguatge natural arrossega tots els problemes derivats de l'ambigüitat (sinonímia, polisèmia, homonímia). Hi ha dos tipus de llenguatges lliures: les llistes de descriptors lliures i la llista de paraules clau.

Un **vocabulari controlat** és una llista prèviament redactada 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, adjectiu, sigles, etc.), com en el seu contingut (de tots els sinònims se n'escull un, els homònims es diferencien entre ells amb parèntesis o adjectius, etc.), com en les seves relacions de jerarquia i associació (termes conceptualment més genèrics o específics i termes que s'evoquen mútuament). Requereixen unes despeses de construcció elevades, tant en personal qualificat com en temps. Per a molts autors són els vertaders 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
- Les llistes d'autoritats
- Els tesaurus

Exemples de termes lliures i controlats

Concepte	Lliure	Controlat
Netedat	Higiene, Neteja, Profilaxi, Condícia, Sanitat, Desinfecció	CDU: 613 LEMAC: Higiene

3) Coordinació: precoordinació o postcoordinació

La **precoordinació** consisteix a determinar *a priori* com es combinen els termes, ja sigui en la construcció del llenguatge com a l'hora d'indexar o recuperar el document.

La precoordinació en les biblioteques manuals

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

Així mateix, es fa referència a la precoordinació com la sintaxi del llenguatge documental. Per exemple, en les llistes d'encapçalaments de matèria els epígrafs van en un ordre concret per tal d'evitar la dispersió d'encapçalaments.

Així, un document sobre congressos catalans sobre arqueologia submarina s'indexaria com a Arqueologia submarina – Catalunya – Congressos, i no amb cap altra de les **possibles combinacions**.

Possibles combinacions

Les combinacions errònies són les següents:

- Catalunya – Congressos – Arqueologia submarina
- Arqueologia submarina – Congressos – Catalunya
- Congressos – Arqueologia submarina – Catalunya
- Arqueologia submarina – Congressos – Catalunya

Recordem que l'ordre és determinat per les indicacions que acompanyen cada epígraf. Així veiem que *Arqueologia submarina* pot dur subdivisió geogràfica i que *Congressos* és una subdivisió que pot anar al darrere de noms propis de persona, famílies, entitats, classes de persones, grups ètnics, guerres i temes; per tant, l'únic ordre possible és el de la solució aportada.

Hi ha dos llenguatges precoordinats: els sistemes de classificació i les llistes d'encapçalaments de matèria.

La **postcoordinació** consisteix a indexar termes solts; no tenen sintaxi en el moment de la indexació. Només a l'hora de la recuperació es combinaran seguint la lògica dels operadors booleans.

Cada terme indexat és un punt d'accés al document: com més termes indexem més possibilitat de recuperar-lo tenim. Seguint el cas anterior, el formularíem posant els tres conceptes en qualsevol ordre, ja que no és rellevant, per exemple:

Congressos and Catalunya and Arqueologia submarina

Hi ha quatre llenguatges postcoordinats: les llistes d'autoritats, els tesaurus, les llistes de descriptors lliures i la indexació automàtica.

4) Estructura: jeràrquica o alfabètica (combinatòria)

En l'**estructura jeràrquica** o sistemàtica el vocabulari es presenta en forma d'arborescència, 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. També situar-los en context, ja que la seqüència jeràrquica ens informa del camp temàtic en què està adscrit el concepte.

Exemple

Posem, per exemple, el concepte *llibertat*, que té moltes accepcions. Solament veient on està inserit, ja deduïm si es tracta de la llibertat filosòfica, de drets humans o de la llibertat de moviments en màquines.

L'estructura jeràrquica informa del camp del coneixement

Classe 1	Classe 3	Classe 6
123 Libertad y necesidad 123.1 LIBERTAD. INDETERMINISMO 123.11 Casualidad 123.2 NECESIDAD 123.21 Fatalismo	342.7 DERECHOS FUNDAMENTALES. DERECHOS HUMANOS. DERECHOS Y DEBERES DE LOS CIUDADANOS. 342.71 Nacionalidad. Ciudadanía. 342.72/.73 Derechos de los ciudadanos. Derechos civiles. El Estado y el ciudadano. 342.721 Libertad individual. Habeas corpus.	62-23 ENGRANAJES. ELEMENTOS MECÁNICOS DE TRANSMISIÓN. DISPOSITIVOS TRANSPORTADORES Y DE SUJECIÓN 62-231 Estructuras de los mecanismos de transmisión 62-231.2 Sistemas lineales. Pares cinemáticos 62-231.21 Sistemas sin grados de libertad. Acoplamiento automático. Centrado automático 62-231.22 Sistemas con un grado de libertad. Cojinete. Barra de guía. Par de roscado (tornillo y tuerca)

De llenguatges jeràrquics n'hi ha dos: els **sistemes de classificació** i els **tesaurus** (en la part de presentació sistemàtica o jeràrquica).

En l'**estructura combinatòria**, els termes no formen cadena, estan organitzats en llistes per ordre alfabètic. Aquest tipus d'estructura va sorgir com a contrapunt a la rigidesa de l'estructura jeràrquica, que no era fàcil d'actualitzar.

Exemple extret de la *Lista de encabezamientos* del CSIC.

Árbol de la papaya

Árbol de la vida

Árbol del conocimiento

Árboles

Árboles – Crecimiento

Árboles – Cuidados

Árboles – Cultivo

Árboles – Culto

L'estructura combinatòria permet la inclusió de termes nous i l'eliminació dels obsolets sense afectar la resta de l'estructura del llenguatge.

En la seqüència anterior podríem incloure: Árboles – Adobo, sense alterar la resta.

La facilitat per a actualitzar el vocabulari els converteix en llenguatges adequats per a tota mena d'entorns: enciclopèdics, científics i tècnics. De llenguatges d'estructura combinatòria n'hi ha cinc:

- Les llistes d'encapçalaments de matèria
- Les llistes d'autoritats
- Els tesaurus
- La llista de descriptors lliures
- Les llistes de paraules clau

Tesaurus

Com es pot observar, el tesaurus participa de les dues estructures: té una presentació sistemàtica en forma jeràrquica i una presentació alfabètica en forma combinatòria.

5) Anàlisi: per matèries, per conceptes o per paraules clau

La diferència entre un i els altres rau a indexar un tema, diversos conceptes o totes les paraules amb significat del document.

Reflexió

Avui dia l'evolució i automatització dels sistemes d'informació possibiliten que aquests llenguatges, en origen sintètics, puguin indexar de manera més analítica, en especial els encapçalaments de matèria que poden indexar dos, tres o quatre encapçalaments. O les notacions amb sistemes de classificació que dupliquen el camp 080 del MARC.

a) Per matèries

És la indexació més sintètica: indexa un o dos termes d'indexació. Respon a la pregunta "quin és el tema d'aquest document?". De llenguatges que indexen per matèries n'hi ha dos: els sistemes de classificació i les llistes d'encapçalaments de matèria.

b) Per conceptes

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. De llenguatges que indexen per conceptes n'hi ha tres: les llistes d'autoritats, els tesaurus i les llistes de descriptors lliures.

c) Per paraules clau

Indexar per paraules clau representa 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. Només hi ha un llenguatge per paraules clau, i és evidentment l'únic llenguatge automàtic: és la llista de paraules clau.

Resum de les tipologies

		Sistemes de classificació	Llistes d'encapçalaments de matèria	Llistes d'autoritats	Tesaurus	Llista de descriptors lliures	Llista 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		

		Sistemes de classificació	Llistes d'encapçalaments de matèria	Llistes d'autoritats	Tesaurus	Llista de descriptors lliures	Llista de paraules clau
Segons el nivell de coordinació dels termes	Precoordinat	X	X				
	Postcoordinat			X	X	X	X
Segons la forma d'agrupar els termes o estructura	Jeràrquic	X			X		
	Alfabètic		X	X	X	X	X
Segons el nivell d'anàlisi	Per matèries	X	X				
	Per conceptes			X	X	X	
	Per paraules clau						X

Una bona praxi és estudiar els sis llenguatges segons la tipologia i recordar fórmules com ara:

1 codificat + 5 naturals = 6

4 controlats + 2 lliures = 6

2 precoordinats + 4 postcoordinats = 6

2 jeràrquics + 4 combinatoris = 6

2 per matèries + 3 per conceptes + 1 per paraules clau = 6

1.5. Activitats

1) Tot llenguatge controlat és codificat?
2) Tot llenguatge codificat és controlat?
3) Tot llenguatge lliure és natural?
4) Els llenguatges documentals naturals són ambigus?
5) Tots els llenguatges documentals donen el mateix nom al seu terme d'indexació?
6) Hi ha llenguatges controlats que no són codificats?

2. Avaluació de la indexació i la recuperació

Aquest apartat està dedicat a avaluar les tasques d'indexació i recuperació. Aprendrem a mesurar la qualitat dels analistes, els llenguatges, la indexació i la recuperació. Per a fer-ho, ens ajudaran unes fórmules molt senzilles i eficaces.

2.1. Mesurar la qualitat

La qualitat no és un concepte tan abstracte com pot semblar *a priori*. Es pot mesurar tal com indica la norma ISO 9000 en funció del grau en què es compleixen unes necessitats o expectatives prèviament establertes en normatives. Direm, doncs, que la indexació respectuosa amb les normes aplicades al seu SID serà una **indexació de qualitat**.

Web recomanat

Podeu consultar informació sobre la família d'estàndards ISO 9000 (International Standard Organization) a http://www.iso.org/iso/iso_catalogue/management_standards/iso_9000_iso_14000/iso_9000_essentials.htm

Quines normes regeixen la indexació? Les més representatives per a nosaltres són les següents:

- Normes internacionals d'indexació de la ISO traduïdes a norma UNE 50-121-91: Mètodes per a l'anàlisi de documents, determinació del seu contingut i selecció de termes d'indexació.

- Normes de creació de tesaurus monolingües NISO Z39.19 (2005) i multilingües.

- Normes de creació, introducció i intercanvi de dades referents a l'anàlisi de contingut (resum i indexació) de les Regles de catalogació AACr2 o MARC21.

- I, finalment, la Normativa de creació i manteniment de termes d'indexació de cada llenguatge documental controlat, és a dir, la norma que regeix el sistema de classificació que usem al nostre SID (pot ser la CDU), de la nostra llista d'encapçalaments de matèria (pot ser la LEMAC o la del CSIC), de les nostres autoritats (pot ser LENOTI o la de la BNE) o finalment la del nostre tesaurus (pot ser EUROVOC).

Observació

Fixeu-vos que falten dos llenguatges documentals: la llista de descriptors lliures i la llista de paraules clau, que en ser lliures no tenen normativa prèvia.

En aquestes normatives es dicten les convencions internacionals o nacionals per emplenar els camps del registre bibliogràfic, en quina forma, com s'ha de construir un llenguatge documental, o com cal mantenir el que ja tenim, com cal avaluar-lo amb tests i càlculs bibliomètrics, com s'ha d'indexar, com cal tractar l'ambigüitat del llenguatge natural, com s'han d'elaborar taxonomies o anells de sinònims, etc.

Bibliografia

AENOR (1990). *UNE-50-106 (ISO 2788-1986). Documentación: Directrices para el establecimiento y desarrollo de tesauros monolingües*.

AENOR (1997). *UNE-50-125 (ISO 5964-1985). Documentación: Directrices para la creación y desarrollo de tesaurus multilingües*.

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.

NISO Z39.19 (2005). *Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies*.

NISO Z39.19 (2003). *Guidelines for the Construction, Format, and Management of Monolingual Thesauri*.

És convenient que cada SID disposi d'un **sistema de gestió de qualitat**. Cada SID ha d'establir i mantenir un sistema de gestió de la qualitat d'acord amb la norma ISO 9001:2008, en el qual s'han de determinar els procediments per a analitzar el contingut del documents (classificació o indexació), els responsables, els procediments detallats pas a pas, les comprovacions i la implementació de mesures correctives quan calgui.

Cada SID també hauria de tenir un **manual de procediments**, en el qual s'expliqués pas a pas cada acció (com la selecció de termes que cal indexar), amb exemples, modificacions i autoritzacions de procediments nous. D'aquesta manera, si els procediments estan arbitrats, serà més fàcil saber com cal indexar i fer-ho amb correcció.

Web recomanat

A la Xarxa trobem diversos manuals de sistemes de gestió de qualitat en biblioteques, un dels quals és el de la Biblioteca universitària de Màlaga. http://www.rebiun.org/opencms/opencms/handle404?exporturi=/export/docReb/biblio_garciareche2.pdf&%5d

Per tal d'avaluar la indexació ens fixarem en dos elements: la **qualitat del llenguatge documental** i la **tasca de l'indexador**, que ens marcaran els criteris d'exhaustivitat, precisió i coherència. Les causes fonamentals de fallida en la recuperació són, a més de les anteriors, la cerca i la interrelació entre usuari i sistema.

2.2. Els llenguatges documentals

Per a indexar amb qualitat necessitem que els nostres llenguatges documentals estiguin actualitzats. Cal fer-ne un manteniment regular per a corregir els errors, les mancances detectades i seguir l'evolució dels dominis coberts pel sistema documental i les necessitats dels usuaris.

En més detall, el seguiment del llenguatge consistiria a fer:

- **Recomptes de freqüència d'ús dels termes d'indexació**. Així, observem cada terme i el nombre de documents que indexa. Les estadístiques ens mostren que hi ha gran diversitat, una gran quantitat de termes no s'usa, una petita quantitat s'usa molt. Com diu Van Slype, el 20% dels descriptors més usats representen el 80% de les indexacions. Per fer-ne el seguiment podem demanar al nostre SGBD que ens informi dels termes superiors o

inferiors al barem que considerem oportú; per exemple, saber quins termes han estat indexats per sobre de 100 casos o inferiors a 10.

- **Recomptes dels termes que els usuaris usen en les cerques**. Es tracta de comprovar la correlació entre els termes que indexem i els que són emprats pels usuaris. Potser nosaltres indexem *indústria tèxtil* i els nostres usuaris cerquen per *seda* o *cotó* (en aquest cas, estem indexant de manera genèrica) o, a la inversa, estem indexant *acetat de cel·lulosa* i no s'usa perquè els usuaris busquen per *indústria del plàstic*.

Els dos recomptes permeten llimar les mancances del llenguatge documental, ja sigui incorporant-hi els conceptes que hem detectat que no tenen correspondència al nostre sistema, com eliminant-ne els obsolets, com mantenint els que resultin útils.

Conceptes nous

Els termes afegits provenen de conceptes nous (per exemple, Whatsapp, Skysurfing) o termes que hem d'obrir en més termes específics perquè són molt buscats pels usuaris (per exemple, tenim ordinadors però obrim amb tauletes).

2.3. La qualitat de l'indexador

En aquest apartat analitzarem el paper que tenim nosaltres com a indexadors. Abans, però, fem una llista dels avantatges que ens faciliten la tasca:

1) Hi ha temes més fàcils d'indexar que altres pel coneixement que en tenim.

2) Hi ha llenguatges més fàcils que altres, com són els postcoordinats, que ens estalvien conèixer les regles de precoordinació. Els "fàcils" són les llistes d'autoritats, els tesaurus, els descriptors lliures i la llista de paraules clau.

3) És més fàcil indexar una dada que una matèria.

> És més fàcil indexar *Aristòtil* que les matèries d'algunes de les seves obres. Per tal d'indexar *Aristòtil* només cal consultar una Llista d'autoritats com Lenoti i un *vegeu* propi d'una relació d'equivalència ens diu que hem d'indexar *Aristòtil, 384-322 aC*. Fent dos clics tenim l'autoritat acceptada. En canvi, indexar la matèria d'una obra seva és més laboriós (decidir quina branca de la filosofia, conceptes, etc.).

4) Si el SID disposa de manuals i tutorials sobre el grau d'exhaustivitat i especificitat que volen, ens sentirem més guiats.

Així i tot, com a indexadors podem cometre dos tipus d'errors: els **tècnics** i els **ètics**.

2.3.1. Errors tècnics

Partim del supòsit que l'indexador no cometrà errors de coneixement del llenguatge que té a les mans, com ara no entendre les referències de *vegeu* d'un terme no acceptat a un d'acceptat. Amb tot, pot cometre els errors que indiquem a continuació.

1) Errors en la **selecció del tema del document:** l'indexador no ha captat la vertadera matèria del document. Causes: falta d'atenció o desconeixement de la matèria.

2) Errors en la **selecció numèrica dels termes:** es pot equivocar obviant temes interessants, és a dir, el document tracta, posem per cas, de quatre temes i n'escull només dos.

3) Errors en la **selecció del terme:** l'indexador escull un terme més genèric del que seria desitjable per una manca d'especificitat del llenguatge documental. L'absència del terme l'obliga a indexar amb un terme conceptualment més genèric. L'indexador comprèn la matèria, però el llenguatge documental no li permet expressar-se.

4) Errors per **omissió:** el document tracta d'un tema que no apareix al llenguatge documental i davant el dubte de què és, no l'indexa.

El document tracta sobre les aplicacions de l'Apple store i, com que no surt en el llenguatge, no indexa res quan seria millor indexar un terme genèric com *comerç electrònic*. Segons Lancaster, un 10% dels errors en l'exhaustivitat són deguts a omissions. Se solucionarien si el llenguatge disposés de referències de termes equivalents i de termes relacionats, tipus *Apple Store*, *Microsoft Store TR Comerç electrònic*.

5) Errors en la **formalització:** s'equivoca en la grafia del terme.

351.8w, per 351.82 (Administració pública de l'economia en la CDU), o Detns per Dents. Aquest error se soluciona no teclejant el terme sinó copiant-lo de fitxers o llistes d'autoritats.

6) Errors en la **coherència en equivocar-se amb la sintaxi del llenguatge precoordinat**, fet que impedeix reunir tots els documents que tracten del mateix tema.

La manca de consistència es pot donar en diversos nivells, que exemplificarem a partir d'un encapçalament compost, com és Dents – Cura i higiene – Estadístiques.

En el cas òptim que tots els indexadors coneixen la precoordinació, trobarem ordenats tots els documents indexats per la seqüència Dents, com, per exemple:

Dents – Cura i higiene – Estadístiques

En canvi, si un indexador altera l'ordre dels subencapçalaments, es produirà una barreja en què perdrem documents.

Cura i higiene – Dents – Estadístiques

Si un indexador indexa amb un terme genèric de *dents*, també perdrem la seqüència

Boca – Cura i higiene – Estadístiques.

7) Errors en l'**emmagatzematge en el catàleg:** són errors tècnics derivats del programa de gestió (manca d'espai en els camps, en la memòria, etc.).

Els dos primers errors no tenen relació amb el vocabulari del llenguatge. Els següents sí, i és en aquests darrers casos que un llenguatge ben construït pot ajudar a minimitzar-los: amb termes genèrics oberts en suficients termes específics, termes no usats que remeten amb *vegeu* als termes usats, amb notes d'aplicació i notes explicatives als descriptors, amb referències creuades i termes relacionats. Com més ric sigui el llenguatge, menys coneixements en la matèria ha de tenir l'indexador.

2.3.2. Errors ètics

Hi ha tres tipus d'errors ètics: els de discriminació (o ofensa), els de censura i els intencionats.

1) Errors per discriminació o ofensa

Cal evitar termes que puguin resultar ofensius o discriminatoris per qüestions de gènere, raça, religió, condició, etc.

El control del vocabulari és una eina de gran valor en aquesta missió, ja que els llenguatges controlats han passat per una tria de conceptes en què la majoria dels termes ofensius han estat rebutjats. I diem la majoria perquè alguns llenguatges arrosseguen concepcions antigues que costen de modificar. En la bibliografia científica sobre encapçalaments de matèria trobem molts articles que analitzen temes sensibles comparant encapçalaments en dues llistes i que demanen una revisió urgent dels epígrafs.

Lectura recomanada

Per tal d'ampliar aquest tema recomanem la lectura de Carmen Caro i R. San Segundo, *Lenguajes documentales y exclusión social* (http://dialnet.unirioja.es/servlet/articulo?codigo=1300420) en la qual s'analitzen encapçalaments que posen sota el mateix terme genèric les mares solteres amb els delinqüents dins el grup de marginats socials. O relacionen dos termes tan dispars com *anarquisme* amb *idiotesa*. Els sistemes de classificació també cometen errors ètics, per exemple, mantenint la rúbrica de la classe 159.922.76 per a nens amb defectes físics, mentals i superdotats.

En aquests casos, és recomanable no usar aquests termes per a indexar i proposar un acord intern del SID per a substituir-los. Si indexem amb un llenguatge en línia, accedirem a totes les actualitzacions, però en el cas que el nostre llenguatge estigui en paper, caldrà comprovar en les actualitzacions al web si el terme ofensiu ja ha estat modificat o no.

Exemple

Per exemple, el consorci de la CDU vetlla pel manteniment i actualització del Master Reference File, i en aquesta adreça http://www.udcc.org/major_changes.htm podem comprovar l'estat del terme que ens (pre)ocupa.

En entorns d'indexació lliure com cercadors generals o marcadors socials podem trobar etiquetes sobre temes sensibles expressats de manera vexatòria o sectària, ja que ningú més que el mateix autor del text o l'internauta pren la decisió d'indexar-los.

2) Errors per censura

Totes les fases de la indexació estan influïdes per un cert grau de subjectivitat de l'analista (per la seva formació, conviccions polítiques, creences religioses, etc.), però el documentalista, tal com recull el codi d'ètica de l'American Library Association, ha de distingir entre les seves conviccions personals i les seves responsabilitats professionals i no permetre que les creences personals interfereixin en la representació del contingut dels documents.

> "[...] We distinguish between our personal convictions and professional duties and do not allow our personal beliefs to interfere with fair representation of the aims of our institutions or the provision of access to their information resources [...]."
>
> Code of Ethics of the American Library Association: http://www.ala.org/advocacy/proethics/codeofethics/codeethics

3) Errors intencionats

Un tercer tipus d'error ètic és indexar intencionadament de manera equivocada per aconseguir un guany, com, per exemple, un millor posicionament web. És conegut com a *falsejament d'índexs* o *spamdexing*. Consisteix a indexar conceptes que ens asseguren més visibilitat a la Xarxa (exemple, *molt interessant*) augmentant les referències creuades i enriquint els enllaços cap al web. Per a evitar el falsejament d'índexs o per a comprovar que les etiquetes que hem assignat a un web no es considerin falsejades, val la pena consultar abans les polítiques dels cercadors.

Web recomanat

Eines per a administradors de webs (*webmasters*) de Google: http://support.google.com/webmasters/bin/answer.py?hl=es&answer=35769

2.3.3. Com es mesura la qualitat d'un indexador?

La qualitat d'un indexador es mesura comparant-lo amb un altre indexador. Aquesta operació es resol calculant la taxa de coherència.

Partirem d'un cas delimitat: 2 documentalistes, 10 documents i 3 descriptors. La fórmula de la taxa de consistència és:

c / a + b – c

Llegenda:

a equival a termes indexats a *Indexador a*

b equival a termes indexats a *Indexador b*

c equival a termes comuns en les dues indexacions

Descriptor	Documentalista	Doc 1	Doc 2	Doc 3	Doc 4	Doc 5	Doc 6	Doc 7	Doc 8	Doc 9	Doc 10
Cadaqués	A		x	x	x		x	x	x	x	x
	B		x		x		x			x	
Parc natural	A	x	x	x	x						
	B	x	x	x							

Descriptor Cadaqués: 4/8 = 50%.
Descriptor Parc natural: 3/4 = 75%
Descriptor Cala Culip : 0/5 = 0%

Descriptor	Documentalista	Doc 1	Doc 2	Doc 3	Doc 4	Doc 5	Doc 6	Doc 7	Doc 8	Doc 9	Doc 10
Cala Culip	A	x				x			x		
	B			x			x				

Descriptor Cadaqués: 4/8 = 50%.
Descriptor Parc natural: 3/4 = 75%
Descriptor Cala Culip : 0/5 = 0%

Nota

Algunes característiques dels llenguatges afavoreixen o dificulten la coherència. Respecte de la CDU, ha substituït l'ús del signe subdividir a favor del *colon* (:) i altres facetacions (com en la taula 9) perquè els indexadors interpretaven malament les instruccions i donaven lloc a taxes de coherència molt baixes.

2.4. Avaluació de la recuperació

En la recuperació s'avaluen conceptes de **microavaluació** (silenci, soroll) i de **macroavaluació** (exhaustivitat i precisió). I comparant-ho arribem al de consistència o coherència, que ja hem vist anteriorment.

Partim del quadre següent (Lancaster i Van Slype), en el qual veiem totes les possibilitats que es produeixen en la recuperació:

Llegenda dels elements de la recuperació

	Pertinents	No pertinents	Total
Extrets	A (encerts)	B (soroll)	A + B (recuperats)
No extrets	C (pèrdues)	D (correctament rebutjats)	C + D (no recuperats)
Total	A + C (total de documents rellevants)	B + D (total de documents no rellevants)	A + B + C + D (col·lecció sencera)

Com es calculen els documents pertinents i no pertinents? Cal que l'usuari valori com a pertinent o no pertinent el conjunt de documents que el sistema li ha donat. Dels quatre valors (A, B, C, D), podem saber A perquè són els que s'han recuperat i l'usuari considera rellevants i B perquè no els considera rellevants. En canvi, per a saber C i D necessitaríem un entorn ideal en què l'usuari mirés tota la col·lecció i decidís quins haurien estat pèrdues i quins no. Com que això no es pot fer pel volum de la col·lecció, s'agafa una secció i s'extrapola el resultat.

Aquest exemple servirà per a argumentar la resta del mòdul: imaginem que hem cercat per documents que continguin el terme *Cadaqués*:

	Pertinents	No pertinents	Total
Extrets	5	2	7
No extrets	3	30	33
Total	8	32	40

2.4.1. Microavaluació: silenci i soroll

Taxa de silenci: c / a + c.

En la cerca sobre Cadaqués observem que és 3 / 5 + 3 = 0,375, és a dir, el 37,5% dels documents pertinents no s'han recuperat. La taxa de silenci és del 37,5%.

Taxa de soroll: b / a + b

Sobre el mateix exemple, és 2 / 5 + 2 = 0,285. La taxa del soroll ha estat del 28,5%.

2.4.2. Macroavaluació: exhaustivitat i precisió

Taxa d'exhaustivitat: a / a + c

Nota

Exhaustivitat: *Recall* en anglès, *Rapell* en francès i *Llamada* en castellà.

L'exhaustivitat de la cerca sobre Cadaqués dóna 5 / 5 + 3 = 0,625. La taxa d'exhaustivitat ha estat del 62,5%. Els valors habituals són entre 0,6 i 0,8.

Aquesta taxa expressa la capacitat del sistema per a proporcionar el que es vol amb un grau satisfactori d'exhaustivitat. Ara bé, amb això sol no n'hi ha prou per a avaluar la qualitat, també necessitem que ens filtri el que no necessitem, i aquí entra la taxa de precisió.

Taxa de precisió: a / a + b

La precisió de la cerca sobre Cadaqués dóna 5 / 5 + 2 = 0,714.

Resum de les fórmules per a calcular silenci, soroll, exhaustivitat i precisió

Microavaluació		Macroavaluació	
Silenci	Soroll	Exhaustivitat	Precisió
c / a + c	b / a + b	a / a + c	a / a + b

Hem vist les taxes de silenci i soroll i les d'exhaustivitat i precisió, però una anàlisi completa comprèn l'examen dels documents, els registres d'indexació, els fulls de petició, les estratègies de cerca, els fulls de valoració de la rellevància

i qualsevol altra informació que pugui ser obtinguda dels usuaris que participin en l'estudi. A partir d'aquests registres es poden determinar les causes concretes dels errors del sistema en la recuperació.

2.5. El paper del vocabulari en la recuperació

Segons Lancaster, es produeixen tres errors en relació amb el vocabulari:

1) Manca d'especificitat del llenguatge documental
2) Relacions ambigües
3) Relacions falses entre termes

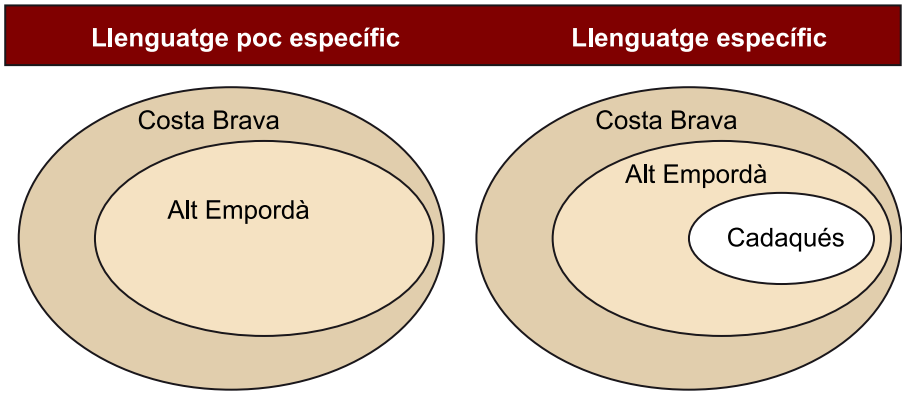
2.5.1. Manca d'especificitat del llenguatge documental

La manca d'especificitat del llenguatge documental és la causa principal de les mancances en la recuperació i es dóna principalment en l'àmbit dels **llenguatges controlats**.

Reflexió

Recordem que els llenguatges lliures no disposen *a priori* d'un vocabulari controlat; per tant, l'analista o l'algorisme del programa indexarien *Cadaqués* sense verificar si aquest terme existeix o no en una llista acotada. Els llenguatges lliures són tan específics com ho és el text.

Figura 2. Comparació entre dos llenguatges quant a la seva especificitat



Si el llenguatge no és específic, encara que l'analista vulgui indexar Cadaqués, no el podrà indexar i haurà de recórrer al seu terme genèric (TG), com, per exemple, Alt Empordà o Costa Brava. En aquest punt podem trobar problemes tant en la indexació com en la recuperació:

• En la **indexació**: si no hi ha remissions entre termes un analista podria indexar Costa Brava i un altre analista indexar Alt Empordà. En canvi, si el llenguatge té notes d'aplicació o equivalències que remetin al terme designat (tipus Cadaqués empreu Costa Brava), tots els analistes indexaran amb el mateix terme, posem Costa Brava, i no hi haurà problemes de consistència entre ells.

• En la **recuperació**: l'usuari, que no ha de conèixer el llenguatge amb antelació, busca Cadaqués i el sistema li respon 0 resultats, perquè no sap que els documents s'han indexat amb altres TG.

Si el llenguatge fos específic i tingués un terme per a Cadaqués, el llenguatge tendiria a:

• Augmentar la precisió: quan cerquem Cadaqués recuperem Cadaqués i no altres poblacions de la Costa Brava o de l'Alt Empordà.

• Disminuir l'exhaustivitat: només recuperem els documents que tracten de Cadaqués i no recuperem els de Llançà o el Port de la Selva.

Un vocabulari específic incrementa la precisió i disminueix l'exhaustivitat; per contra, un vocabulari poc específic facilita l'exhaustivitat, però baixa la precisió i també augmenta la consistència, ja que hi ha menys termes entre els quals triar.

Malgrat tot, és millor que el llenguatge documental sigui específic, és a dir, **és preferible la precisió abans que l'exhaustivitat**, ja que aquesta es pot aconseguir cercant pel TG.

En resum:

- Un **vocabulari específic** permet una precisió alta, però complica el fet d'aconseguir una exhaustivitat alta. També influeix en la consistència, ja que si els termes són molt pròxims, es pot dubtar entre un terme o un altre.
- Un **vocabulari poc específic** facilita la cerca genèrica i minimitza les incorreccions de la indexació i en conseqüència augmenta l'exhaustivitat, però dificulta una precisió alta.
- Amb tot, segons Lancaster, és més favorable d'un excés d'especificitat que el contrari, ja que si volem augmentar l'exhaustivitat només cal recórrer als TG. En canvi, la manca d'especificitat fa que no es pugui augmentar la precisió.

2.5.2. Coordinacions falses

Entre els termes hi ha dos tipus de relacions ambigües o falses: les **coordinacions falses** i les **relacions incorrectes entre termes**. Totes dues es produeixen perquè les paraules en si mateixes no tenen sintaxi, especialment si són unitermes o termes simples.

Diguem que una coordinació és falsa quan recuperem documents no pertinents, però que contenen els termes de cerca que hem demanat. La coordinació és falsa perquè en el document original els dos termes hi són però no estan relacionats.

En un sistema sense sintaxi, com més termes d'indexació hi hagi més alta és la probabilitat que es recuperin coordinacions falses. En canvi, és menys freqüent en sistemes precoordinats, en els quals tenen un control més estricte. Les coordinacions falses es podrien solucionar si a l'hora d'indexar el document es fessin evidents les relacions entre els termes, almenys en termes relacionats i no relacionats.

Exemple

Fem una consulta sobre finançament dels arxius a Barcelona, i recuperem el document amb el qual hem iniciat aquest mòdul, l'article d'Alice Keefer sobre repositoris digitals universitaris, pel fet que els termes *finançament*, *arxius* i *Barcelona* hi són presents, encara que l'article no parli del finançament dels arxius barcelonins –*finançament* hi apareix en relació amb l'*open acces*, *arxius* amb referència a l'autoarxiu del professorat i *Barcelona* surt en les dades formals de l'article.

2.5.3. Relacions incorrectes entre termes

Les relacions incorrectes entre termes es donen quan l'usuari cerca dos termes amb un tipus de relació que no és ben bé la que té el document, tot i que hi surt.

Exemple extret de Lancaster:

L'usuari cerca per *disseny d'ordinadors* i recupera documents sobre el disseny d'avions amb ordinador. Com es pot veure, els termes *disseny* i *ordinadors* són presents en el document, encara que no en el sentit de la demanda.

La solució no és posar sigles de termes relacionats (TR), perquè el problema és que no sabem quin tipus de relació tenen.

La manera de solucionar aquest problema en un entorn postcoordinat és **assignant rols o indicadors als descriptors**, que són codis o xifres, vertaders recursos (agents) sintàctics que en marquen el rol dins el document.

Per exemple, (2) podria indicar instrument mitjà i (4) objecte, subjecte. El resultat de les indexacions seria el següent:

Exemple de rol

Document del disseny d'avions amb ordinadors	Document del disseny d'ordinadors
Disseny Avions (4) Ordinadors (2)	Disseny Ordinadors (4)

Per a recuperar el document inicial que volia l'usuari, la cerca seria: Ordinadors (4) and Disseny.

Aquest sistema de rols és propi de les ontologies. Els llenguatges documentals actuals no arriben a especificar el rol de cada concepte, només marquen si són termes relacionats sense especificar de quin tipus.

Altres exemples:

Fem una consulta sobre *Pintura i guerra* (en el sentit de la guerra representada en la pintura, com el *Guernica* de Picasso) i recuperem documents sobre pintura de guerra (maquillatge durant la guerra).

Fem una consulta sobre pintura catalana, en el sentit de pintors catalans com Fortuny, Casas, Dalí, i recuperem, a més dels documents interessants, aquests altres:

- Catalunya en la pintura (p. ex. la visió de J. Sorolla sobre el litoral català)
- Pintura a Catalunya (tots aquells pintors que han pintat a Catalunya)
- Industrials de la pintura catalans (pintors de parets).

En resum:

Si augmentem l'especificitat del vocabulari, ens permet representar amb més matisos el significat; per tant, disminueix la consistència en la indexació, augmenta la precisió i baixa l'exhaustivitat.

Resum de l'augment de l'especificitat

Augment de l'especificitat	Augmenta la precisió
	Disminueix la consistència
	Disminueix l'exhaustivitat

Quant a la recuperació: probablement l'estructura del llenguatge condiciona la cerca de manera important. Com més estructurat sigui i més relacions tingui cada terme, més útil serà per a construir estratègies de cerca, malgrat ser costoses.

Les coordinacions falses: la causa d'aquest error és que els termes d'indexació es troben en el mateix document però en un context diferent del que cerca l'usuari.

Les relacions incorrectes: la causa d'aquest error és que el llenguatge no especifica el tipus de relació que tenen els termes entre ells.

Activitats

Seguidament, us proposem un seguit d'activitats perquè pugueu posar en pràctica les explicacions teòriques sobre indexació, recuperació i avaluació que hem plantejat en aquest mòdul.

Activitat 1

Observeu atentament aquests conjunts de dos llenguatges comparats. Com a bons indexadors i constructors de llenguatges documentals, fent tan sols un cop d'ull a l'estructura del llenguatge podeu deduir les qüestions següents:

A	B
. Administració de justícia .. Notaries ... Procediment legal .... Denúncies .... Procés judicial ..... Arbitratge ..... Decisions judicials ..... Actes resolutoris ..... Requeriments judicials ..... Sentències	. Administració de justícia .. Notaries ... Procediment legal .... Procés judicial ..... Decisions judicials

a) Quin dels dos llenguatges pot oferir *a priori* més precisió en la recuperació?

b) Quin dels dos llenguatges oferirà una taxa d'exhaustivitat més elevada?

c) Amb quin dels dos obtindran els indexadors una taxa de consistència més elevada?

C	D
LEMAC LENOTI AUTORITATS DEL CSIC	Tesaurus

d) Quin dels dos tipus de llenguatges és susceptible de crear més coordinacions falses?

E	F
Absolutisme UP Autocràcia Despotisme Monarquia absoluta TC [Política] TC2 [Teoria política] TG Actituds i moviments polítics TR Antic Règim Constitucionalisme Crisi de l'Antic Règim Dècada Ominosa Liberalisme Monarquia Monarquisme Nova Planta Voluntaris reialistes	**Absolutisme** UP Autocràcia Despotisme Monarquia absoluta TC [Política] TC2 [Teoria política] TG Actituds i moviments polítics TR Antic Règim

e) Amb quin dels dos llenguatges s'haurà d'esforçar més l'usuari o l'analista que faci la cerca?

G	H
Voldria saber el nombre de naixements per cesària de la comunitat de Galícia actualment?	Documents sobre naixements per cesària

f) Quina pregunta és més difícil de satisfer de manera precisa i exhaustiva per un analista?

Activitat 2

Consulteu les fonts d'informació següents i intenteu esbrinar quin o quins llenguatges documentals indexen el fons documental.

a) Bases de dades del CSIC

b) Biblioteca de Catalunya

c) Biblioteca Nacional de España

d) Biblioteca UOC

e) Bing

f) Delicious

g) Flickr

h) Google

i) Bases de dades UNESCO

j) Wikipedia/Viquipèdia

k) Yahoo

l) Youtube

Bibliografia

Manuals i normatives

AENOR (1990). *UNE-50-106 (ISO 2788-1986). Documentación: Directrices para el establecimiento y desarrollo de tesauros monolingües*.

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

AENOR (1997). *UNE-50-125 (ISO 5964-1985). Documentación: Directrices para la creación y desarrollo de tesaurus multilingües*.

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. Norma UNE-50-113/6 (ISO 5127/6)". *Revista Española de Documentación Científica* (vol. 20, núm. 4, pàg. 417-436).

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

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

Lambe, Patrick (2007). *Organising knowledge: taxonomies, knowledge and organisational effectiveness*. Oxford, Regne Unit: Chandos, cop.

Lancaster, F. Wilfrid (1995). *Indización y resumen: teoría y práctica*. Buenos Aires: EB Publicaciones.

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

Maniez, J. (1992). *Los lenguajes documentales y de clasificación: concepción, construcción y utilización en los sistemas documentales*. Madrid: Pirámide / Fundación Germán Sánchez Ruipérez.

NISO Z39.19 (2003). *Guidelines for the Construction, Format, and Management of Monolingual Thesauri*.

NISO Z39.19 (2005). *Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies*.

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").