Limbajele documentare: tipologie şi performanţe

Lector univ. dr. Rodica Mandeal

Abstract: *The work highlights the centrality of language in the retrieval process. The major functions of the controlled vocabulary and both general basic characteristics and performancies of precoordinate and postcoordinate vocabularies are reviewed.*

Key words: *Language in Retrieval; Controlled Vocabular; Natural Language; Characteristics; Performancies.*

Limbajele documentare permit reprezentarea conţinutului documentelor în scopul regăsirii documentelor pertinente ca răspuns la întrebări (cereri de informare) care se referă la acest conţinut. Deci, un limbaj documentar nu poate fi utilizat la regăsirea documentelor după alte criterii utilizate în cercetarea documentară – autorul documentului, limba în care este editat textul, data publicării etc.

Există două mari tipuri de limbaje documentare:

- limbajele sistematice, clasificatorii, utilizate în general pentru a reprezenta conţinutul documentului într-o manieră sintetică;
- limbajele analitice sau de indexare, denumite şi limbaje combinatorii, care permit reprezentarea conţinutului documentelor şi a cererilor de o manieră analitică.

Modelul clasificărilor ierarhice

În secolul al XIX-lea bibliotecile constituiau principalele resurse ale cunoaşterii. Documentul – carte este în acelaşi timp o unitate fizică, dar şi o unitate intelectuală. Această caracteristică a permis regruparea fizică a cărţilor în funcţie de conţinutul lor. Pentru a sluji acest obiectiv au fost concepute clasificările documentare.

Toate marile sisteme de clasificare universale (Dewey, CZU- Clasificarea Zecimală Universală, LC- Clasificarea Library of Congres) sunt limbaje documentare artificiale, care conţin strâns legate 2 subsisteme: o listă ierarhizată a tuturor subiectelor previzibile şi o codificare a fiecărui subiect.

Acest model organizează subiectele previzibile plecând de la general la particular. Codurile alocate permit cititorilor să aibă acces direct la rafturile cu cărți corespunzând centrului lor de interes (funcție de armonizare) şi de asemenea să lărgească cadrul investigației(funcție zoom).

Modelul clasificărilor ierarhice impune multe constrângeri, datorită:

- rigidității în structură (termenii prestabiliți);
- sărăciei indexării (indexare sumară);
- insuficienței relațiilor semantice;
- dificultății în asimilarea conceptelor noi apărute în diferite domenii.

Totuşi ele sunt solid implantate de un secol în bibliotecile din lumea întreaga (de ex. Clasificarea Dewey este prezentă în 200.000 de biblioteci din 135 de țări). Este păcat că aceste mari sisteme de clasificare nu s-au contopit, pentru a rezulta o normă universal recunoscută, cum a fost ambiția inițială[1]. Evoluția lor în ultima jumătate a secolului a depins mult de factorii locali şi economici; actualizarea clasificărilor cere reeditări frecvente, deci în fiecare din limbile internaționale şi mijloace financiare importante.

Modelul limbajele documentare cu structură analitică.

Spre mijlocul secolului XX, diseminarea cunoaşterii ştiințifice şi tehnice a cunoscut o mutație numită adesea „explozia informației": cartea nu dispare, dar nu mai este sursa, instrumentul, privilegiat de informații, ci articolul din periodic sau raportul de cercetare. Organizarea *documentelor fizice* (revistele) nu mai putea coincide cu ce al *documentelor intelectuale* (articolele).

Acestei situații nou create[2] îi corespunde un nou model de indexare, numită „indexare coordonată": un subiect nu mai este formulat global (ca în cazul clasificărilor ierarhice), ci compus printr-o suită de concepte elementare.

Deci, elementul de bază nu mai este subiectul, ci conceptul. Distincția dintre cei doi termen este mai dificilă fiindcă noțiunile sunt difuze:

- un subiect poate fi considerat ca o reprezentare mentală compusă din mai multe concepte.
- conceptul este o noțiune elementară existentă în universul mintal şi în vocabularul unui grup social la un moment dat. Deci, fiecare concept exprimă un aspect anumit, particular.

Limbajul documentar ideal devine în această situație un repertoriu de concepte aplicat unui domeniu al cunoaşterii şi prevăzut, în unele cazuri, cu reguli de sintaxă.

---

[1] motivația demersului lui Paul Otlet şi Henry La Fontaine de creare a CZU.

[2] …care a dus la diminuarea interesului pentru limbajele de clasificare…

Avantajele acestui model sunt considerabile:

- plecând de la un număr limitat de concepte se pot combina un număr nelimitat de subiecte;
- formula de indexare[3] poate comporta un număr variabil de concepte în funcție de politicile de indexare aplicate şi de documentul prelucrat;
- nu mai este necesară alocarea unor coduri artificial; limbajul, apropiat de cel natural este utilizabil atât de indexator, cât şi de cel care face cercetarea[4];
- în cazul regăsirii, compararea formulei de indexare cu ecuația de căutare devine mult mai simplă: nu mai este necesar ca ele să fie riguros identice[5]. în sfârşit acest tip de căutare este bine adaptat logicii booleene pe principiul căreia funcționează calculatoarele.

Dezavantajul acestor limbaje cvasinaturale în raport cu clasificările ierarhice derivă din faptul că ele sunt tributare particularităților lingvistice.

Apariția limbajelor documentare analitice este rezultatul evoluției rapide a metodelor de înmagazinare (stocare) şi regăsire din domeniile informării documentare.

A trebuit să treacă câțiva ani, de la elaborarea primelor limbaje analitice, la începutul anilor ’60, pentru a se înțelege că cele două mari tipuri de limbaje controlate aveau fiecare misiunea lor:

- limbajele clasificatorii în bibliotecile enciclopedice pentru clasificarea monografiilor, respectiv pentru reprezentarea sintetică a subiectului acestora în cataloage; în bibliotecile specializate şi centrele de documentare pentru organizarea documentelor (articole revistă, rapoarte de cercetare, comunicări la congrese) în rubricile buletinelor analitice şi de semnalare;
- limbajele analitice, tezaurele în special, în centrele de documentare şi ulterior la producătorii bazelor de date bibliografice, pentru indexarea documentelor, adică pentru reprezentarea analitică a conceptelor acestora, în vederea înmagazinării şi regăsirii informației.

Apariția limbajelor analitica, a tezaurelor în mod special, a fost deci justificată de circumstanțele în care acestea erau şi sunt utilizate: indexarea, mai analitică impune utilizarea unor limbaje specializate, în timp ce clasificarea, mai sintetică, poate exploata limbaje universale, ca CZU, de exemplu. Din acelaşi motiv, limbajele analitice au fost obligate să evolueze

[3] termenii de indexare atribuiți documentului

[4] indiferent dacă este un utilizator intermediar – documentaristul- sau unul final.

[5] Este suficient ca un termen din ecuația de căutare să fie prezent în formula de indexare pentru ca documentul să fie considerat pertinent.

rapid pentru a se adapta dezvoltării terminologiei ştiinţifice şi tehnologice, dar şi progreselor telematicii care au condus la facilitarea accesului la sistemele documentare specializate nu numai documentariştilor, ci şi utilizatorilor finali.

Spre sfârşitul anilor ’60 o nouă controversă se semnalează în lumea documentării[6]: de ce să se aloce resurse considerabile pentru construirea unui tezaur şi pentru indexarea documentelor, când este suficient să se înregistreze în memoria calculatorului titlurile şi rezumatele documentelor (mai târziu textul complet), iar regăsirea acestora să se facă după cuvintele cheie pe care le conţin (exprimate în limbaj natural).

Şi în acest caz au fost necesari câţiva ani pentru a se înţelege că limbajul controlat (tezaurul) şi limbajul liber (listele de cuvinte-cheie) aparţin de fapt, aceleaşi clase – limbajele de indexare – şi joacă un rol complementar şi nu antagonic:

- tezaurele, graţie conciziei şi absenţei ambiguităţii termenilor, permit desfăşurarea unor cercetări documentare cu un maximum de precizie, uneori în detrimentul exhaustivităţii;
- limbajul liber din titluri, rezumate sau texte, de o mare bogăţie semantică asigură o mai bună exhaustivitate cercetării, în detrimentul preciziei.

Începând cu anii ’80 asistăm la o evoluţie aparent paradoxală.

Pe de o parte acordarea unei importanţe mai mici preciziei cercetărilor documentare, datorate dezvoltării considerabile a centrelor care vând servicii on-line. Principala preocupare a acestor centre este rentabilizarea investiţiilor făcute în bazele de date şi în echipamentele informatice şi de comunicaţii prin vânzarea unui număr maxim de ore de cercetare şi de referinţe furnizate ca răspuns la cererile de informare. Deci principalul imperativ este exhaustivitatea cercetării, respectiv cercetarea în limbaj liber, neglijându-se într-un fel precizia (respectiv consultarea on-line a tezaurelor cu care au fost indexate respectivele baze de date).

Pe de altă parte se construiesc mai mult decât oricând tezaure în toate organizaţiile (întreprinderi sau administraţie) care dezvoltă, în această perioadă, un număr considerabil de baze de date documentare interne, aproape toate indexate cu limbaje controlate.

În sfârşit, în prezent, una dintre tendinţele care marchează sistemele de regăsire a informaţiei este utilizarea sistemelor expert.

În privinţa limbajelor de indexare, se pare că sistemele expert aduc mai degrabă o evoluţie decât o mutaţie: tezaurele constituie una din componentele sistemului de stocare şi cercetare documentară şi anume baza de cunoştinţe. Aceasta conţine lista conceptelor prezente în documente şi în

[6] În discuţie intră centrele de documentare şi producătorii bazelor de date.

cereri, sub o formă standardizată şi ansamblul relaţiilor semantice dintre aceste concepte.

O a doua componentă, aceea a motoarelor de cercetare, exploatează tezaurul, pentru a transforma cererile de informare, exprimate în limbaj natural în ecuaţii de cercetare, exprimate în limbaj controlat, ducând în final la regăsirea documentelor pertinente.

1.Tipologia limbajelor documentare analitice.

Limbajele de indexare, numite şi limbaje combinatorii sau analitice, permit reprezentarea conţinutului documentelor şi al cererilor de informare fie la nivelul conceptelor acestora, fie al cuvintelor conţinute în titlul, rezumatul şi eventual textul documentelor sau enunţate în cererile de informare.

Indexarea care utilizează un limbaj combinatoriu se numeşte indexare coordonată, deoarece conceptele şi/sau cuvintele care reprezintă conţinutul documentelor pot fi liber combinate între ele în timpul cercetării documentare pentru a reprezenta conţinutul cererilor şi deci regăsirea acestor documente.

În funcţie de nivelul de standardizare a terminologiei folosite, limbajele de indexare se clasifică în:

- limbaje libere, constituite ca urmare a indexării documentelor în limbaj natural;
- limbaje controlate, construite înainte de indexarea documentelor, reprezentate de listele de autoritate (de vedete de subiect) şi de tezaure de descriptori.

1.1. Lista cuvintelor-cheie

Este constituită dintr-o colecţie neordonată de cuvinte-cheie: cuvinte semnificative (non-vide) extrase automat, cu ajutorul calculatorului, din titlul, rezumatul sau textul complet al documentelor.

Cuvintele-cheie sunt cuvinte simple (*uniterm*), acceptate în toate formele gramaticale (substantiv, verb, adjectiv, plural, singular, masculin, feminin) şi ortografice, care definesc cu semnificaţie precisă şi sunt exprimate în limbile în care au fost editate documentele.

În funcţie de mărimea domeniului acoperit, o listă de cuvinte-cheie poate conţine de la câteva zeci de mii la sute de mii de cuvinte.

1.2. Lista descriptorilor liberi

Este constituită dintr-o colecţie neordonată de concepte conţinute în documente şi evidenţiate prin analiză intelectuală, exprimate prin cuvinte sau expresii preluate din documente sau propuse de documentarişti, fără a li se verifica existenţa într-o listă prestabilită.

Lista descriptorilor liberi este prima formă de control al limbajului de indexare, deoarece acceptă numai substantivele (nu formele verbale sau adjectivale), la singular, cuvintele sunt exprimate în aceeaşi limbă, indiferent de cea a documentului. În afara acestor reguli însă, sinonimiile sunt prezente, ca de altfel şi variantele ortografice.

În ceea ce priveşte volumul, o listă de descriptori liberi poate conține câteva zeci de mii de cuvinte.

1.3 Lista de vedete de subiect

Este constituită dintr-o colecție neordonată de concepte, utilizate pentru a reprezenta în mod univoc conținutul documentelor şi al cererilor de informare şi exprimate prin cuvinte sau expresii preluate din limbajul natural într-o formă canonică: substantiv la singular.

Lista de vedete de subiect este un limbaj controlat. Numărul de cuvinte sau expresii este limitat şi numai termenii figurând pe această listă pot fi utilizați la indexarea documentelor sau la formularea strategiei de căutare[7]. Dar între aceşti termeni nu există relații semantice.

În ceea ce priveşte volumul, o listă de vedete de subiect poate conține de la câteva sute la câteva mii de vedete de subiect.

1.4.Tezaurul de descriptori

Este o listă structurată de concepte numite *descriptori*, utilizate pentru a reprezenta în mod univoc conținutul documentelor şi al cererilor.

Ca şi lista de vedete de subiect, tezaurul de descriptori este un limbaj controlat: conceptele sunt exprimate prin cuvinte într-o formă gramaticală standardizată iar numărul de termeni este limitat. În plus, termenii sunt legați prin relații de echivalență semantică[8], de ierarhie şi de asociere.

Un tezaur monolingv conține în general câteva mii de descriptori şi de la câteva sute la câteva mii de non-descriptori.

2. Influența limbajelor documentare asupra eficienței sistemului de regăsire

În funcție de modul în care sunt reprezentate conceptele (prin termeni compuşi sau cuvinte simple), se disting două tipuri de limbaje controlate:

- limbaj precoordonat (de exemplu listele de vedete de subiecte);
- limbajul postcoordonat (cel mai reprezentativ fiind tezaurul).

---

[7] De fapt, ecuația de căutare

[8] de echivalență intralingvistică în cazul tezaurelor monolingve, la care se adaugă relațiile de echivalență interlingvistică, în cazul tezaurelor multilingve

În primul caz, conceptele sunt reprezentate prin termeni mai mult sau mai puțin complecşi. Avantajul utilizării unui astfel de limbaj constă în faptul că încă din etapa indexării conceptele sunt combinate logic: termenii sunt coordonați (combinați) într-o manieră explicită, ceea ce reduce posibilitatea apariției relațiilor ambigue între aceştia.

Dezavantajul limbajului precoordonat constă în inflexibilitatea termenilor, datorată structurii lor liniare.

Într-un sistem postcoordonat, la indexare, documentului îi sunt alocați termeni simpli, fiind obligatorie combinarea lor logică la stabilirea strategiei, în faza de cercetare.

Avantajul limbajului postcoordonat constă în flexibilitatea termenilor care permit o mare profunzime de indexare, precum şi desfăşurarea unor cercetări generice şi multidimensionale.

În funcție de tipul de limbaj controlat folosit în procesul de indexare a documentelor, indexarea este denumită indexare precoordonată şi, respectiv, indexare postcoordonată.

Câțiva indicatori pot sprijini o comparare a performanțelor limbajelor de indexare utilizate în sistemele documentare actuale.

2.1.Univocitatea semantică

Este un indicator legat de prezența sau absența *sinonimiei* şi *polisemiei*.

Un concept poate fi exprimat, în limbaj natural, printr-o serie de sinonime. Pentru regăsirea a maximum de documente pertinente, strategia de căutare trebuie să regrupeze toate posibilele sinonime care se referă la acest concept.

În cazul polisemiei un cuvânt poate exprima, în limbaj natural, mai multe concepte. Atunci când strategia de cercetare se formulează cu astfel de cuvinte, purtătoare de mai multe semnificații, unele documente regăsite nu vor fi pertinente.

În plus, pentru acelaşi domeniu, conținutul semantic al limbajului natural poate fi foarte diferit, de la o limbă la alta. Engleza şi franceza, de exemplu, numără în majoritate, cuvinte simple, care pot fi ambigue; germana, care are în majoritate cuvinte compuse, este mai univocă.

Analizând din punctul de vedere al univocității semantice cele patru tipuri de limbaje de indexare este evident că *lista cuvintelor-cheie* are cel mai scăzut nivel al acestui indicator, ea caracterizându-se printr-o foarte mare ambiguitate semantică.

Lista de descriptori liberi înregistrează mai puține sinonimii grație eliminării diversităților de forme gramaticale. Si polisemia este considerabil redusă datorită utilizării expresiilor, mai semnificative decât cuvintele *uniterm*.

În cazul *listei de vedete de subiect*, în principiu se asigură univocitatea semantică: fiecare concept este exprimat printr-o vedetă de subiect. În practică, absența structurii semantice face dificilă eliminarea completă a sinonimiilor şi polisemiilor.

Prin concepția sa – o listă controlată de termeni şi prezența unei structuri semantice – *tezaurul* este limbajul de indexare cel mai precis. Cu toate acestea se întâlnesc uneori polisemii, acceptate pentru descriptori care se situează în general la periferia domeniului acoperit de tezaur şi cvasisinonimii, adică includerea ca descriptori a unor termeni care ar fi trebuit să nu fie acceptați şi care poate fi explicată printr-o rigoare mai scăzută la construcția tezaurului.

2.2. Actualitatea terminologiei

Comparând acest indicator, limbajele libere sunt mult mai performante decât cele controlate: ele sunt actualizate o dată cu terminologia utilizată în documente. Limbajele controlate, constituite ca liste limitate de termeni, stabilite *a priori*, sunt actualizate cu o anumită întârziere.

2.3. Facilități oferite la indexarea documentelor

Limbajele libere permit o economie importantă de resurse umane şi înmagazinarea mai rapidă a descrierilor documentelor sau a documentelor integrale în bazele de date, deoarece utilizarea lor elimină faza de analiză conceptuală a documentelor şi cea de traducere a conceptelor din limbajul natural, etape obligatorii pentru limbajele controlate.

2.4. Facilități oferite la formularea strategiei de cercetare

Formularea strategiei de cercetare în limbajul liber impune identificarea tuturor cuvintelor sau expresiilor sinonime pentru reprezentarea conceptelor din cererea de informare. Absența unei structuri a limbajului nu permit sistemului să asiste utilizatorul în această etapă. De asemenea, prezența polisemiei şi a falselor coordonări poate duce la regăsirea unor documente nepertinente cererii.

Limbajele controlate elimină majoritatea acestor inconveniente. Tezaurele, sunt din acest punct de vedere cele mai performante, deoarece permit pe de o parte identificarea uşoară a descriptorilor ce exprimă conceptele cererii şi pe de altă parte, extinderea cercetării la alte concepte, mai specifice sau asociate celor din cererea de informare[9].

[9] Această „îmbogățire'' a conținutului cererii permite utilizatorului să cerceteze literatura din diferitele puncte de vedere ale autorilor care au tratat subiectul cererii de informare.

BIBLIOGRAFIE SELECTIVĂ

BANCIU, Doina; MANDEAL, Rodica Sisteme de informare – Sisteme de regăsire a informației. În: *Studii de bibliologie şi ştiința informării,* vol. 3, 1997, pag. 17-24

BATES, M.J. How to Use Controlled Vocabularies More Effectively in Online Searching. În: *Online*, 11, 1988, p.45-56

BECKER, J.; HAYES, R.M. *Information Storage and Retrieval*. New York: Elsevier Science Publisher, 1990

BLAIR, D. *Language and Representation in Information Retrieval*. New York: Elsevier Science Publisher, 1990

BUCKLAND, M. Relatedness, Relevance and Responsiveness in Retrieval Systems. În: *Information Processing and Management*, 19(3), 1983, p. 237-241

GUINCHAT, C.; MENOU, M. *Introduction generale aux sciences et techniques de l'information et de la documentation*. Paris: UNESCO, 1990

LANCASTER, F.W. ELLIKER, S.; CONNELL, T.M. Subject Analysis. În: *Annual Review of Information Science and Technologies*, 24, 1989, p.35-84

MANDEAL, Rodica Utilizatorul şi căutarea informației. În: *Buletinul ABIR*, vol. 13, 2002, nr. 2, p. 8-15