

## Entități bazate

Prof. univ. dr. Ion IVAN  
Daniel MILODIN  
Marius POPA  
Cosmin LUGOJI

### Rezumat

Articolul prezintă conceptul de entitate bazată și se evidențiază operații pe acest tip de entitate. Se definesc indicatori de apartenență și se identifică metode și căi de căutare. Se exemplifică evaluarea entităților bazate folosind software EBASOF.

### Abstract

*It is presented the based entity concept and operations on this kind of entity. There are defined appurtenance indicators and there are identified methods and ways for searching. The based entity evaluation is exemplified using the EBASOF software.*

**I**n managementul organizațiilor se definesc strategii de management și planuri de acțiune. Scopul este de a acționa și de a obține rezultatele conform obiectivelor stabilite.

Pentru situații și contexte similare este eficientă utilizarea de șabloane prin a căror instanțiere se obțin strategii și planuri care sunt adaptate la situația de fapt. Instanțele șabloanelor au și o formă fizică de reprezentare, constituindu-se în entități text. Dacă entitățile rezultate se încadrează în direcția dată de șabloane, atunci ele se numesc entități bazate.

Analiza entităților bazate conduce la concluzii privind diferențele dintre ele și modalitățile de alegere și aplicare a celor mai eficiente în managementul organizațiilor.

## Vocabular și subvocabular

Se consideră alfabetul  $A = \{a_1, a_2, \dots, a_n\}$ , unde  $n$  reprezintă numărul de simboluri, iar  $a_i$  este simbolul de pe poziția  $i$  din alfabet. Cu simbolurile alfabetului se construiesc cuvintele vocabularului  $V = \{c_1, c_2, \dots, c_m\}$ , unde  $m$  reprezintă numărul de cuvinte din vocabular, iar  $c_j$  reprezintă cuvântul de pe poziția  $j$  din vocabular.

Dintre cuvintele vocabularului  $V$ , numai unele sunt reprezentative pentru un domeniu al științei, tehnicii, artei, economiei sau practicii de zi cu zi.

Se consideră domeniile  $D_1, D_2, \dots, D_{nD}$ , unde  $nD$  reprezintă numărul de domenii. Prin domeniu se înțelege un ansamblu de cunoștințe pe baza cărora sunt derulate activități de o anumită natură, având caracteristici specifice. Caracteristicile domeniilor sunt, [IVAN05c]:

- termeni specifici; sunt descrise procese, instrumente, utilaje, activități, caracteristici de calitate a proceselor, produselor, serviciilor, stadii, etape, calificative;
- modalități de asociere a cuvintelor; se descriu procesele și fenomenele care se manifestă într-un domeniu;
- termeni comuni; domeniile nu sunt complet separate unele de celelalte.

Prin reprezentativitatea vocabularului se evidențiază măsura în care cuvintele descriu, ilustrează, sugerează un obiect, fenomen, proces dintr-un domeniu.

Pentru domeniul  $D$ , se extrage un subvocabular  $SV_i = \{c_{i1}, c_{i2}, \dots, c_{ik}\}$ , unde  $k$  reprezintă numărul de termeni specifici domeniului  $D$ , iar  $c_{ij}$  reprezintă cuvântul de pe poziția  $j$  din subvocabularul  $i$ .

După cum este alcătuit subvocabularul  $SV_i$ , rezultă  $SV_i \subset V$ .

Se consideră un text  $T$  constituit din cuvintele subvocabularului  $SV_x$ :

$$T = \langle c_{x1} c_{x2} \dots c_{xH} \rangle$$

Lungimea textului  $T$  măsurată în cuvinte este de  $H$  cuvinte.

Un text construit în sfera de cuprindere a domeniului  $D$ , trebuie să conțină un număr minim de termeni care aparțin domeniului pentru a fi considerat reprezentativ.

Textul T reprezintă un text de referință și e utilizat la construirea de texte  $TT_1, TT_2, \dots, TT_{nb}$ , unde  $nb$  reprezintă numărul de texte bazate.

În continuare, textele  $TT_i, i=1,2, \dots, nb$ , se numesc texte bazate deoarece:

- └ se construiesc cu cuvinte din vocabularul V;
- └ este obligatorie folosirea unei ponderi însemnate din cuvintele subvocabularului  $SV_x$ .

Dacă se consideră vocabularul V, subvocabularul  $SV_x$  și textul bazat  $TT_i$ , obiectivul de bază al unei analize este de a stabili măsura în care la generarea textului  $TT_i$  au fost întrebuințate cuvintele subvocabularului  $SV_x$ .

În cazul în care cu cuvinte din subvocabularul SV se construiește un text T, analiza trebuie adâncită la a stabili măsura în care textul bazat  $TT_i$  utilizează cuvinte din textul T construit pe subvocabularul SV.

### Gradul de includere

Se consideră vocabularul:

$V = \langle \text{ortogonalitate metrici alfabet mulțime matrice indicatori software} \rangle$

și subvocabularul:

$SV = \langle \text{ortogonalitate metrici alfabet mulțime software} \rangle$

Se consideră textele:

$T = \langle \text{se definește ortogonalitate mulțime precum și ortogonalitate alfabet și software, pe baza cărora rezultă gradul de asemănare între alfabet și software} \rangle$

și:

$TT_i = \langle \text{rezultă indicatori asemănare alfabet și software, precum și matrice ortogonalitate alfabet și software} \rangle$

Se construiește tabelul frecvențelor de utilizare a cuvintelor din vocabularul V și subvocabularul SV în textul T, respectiv în textul bazat  $TT_i$ , tabelul 1.

Frecvențele de utilizare a cuvintelor în T și TT<sub>i</sub>

Tabel 1

Vocabular	Subvocabular	T	TT <sub>i</sub>
ortogonalitate	ortogonalitate	2	1
metrici	metrici	0	0
alfabet	alfabet	2	2
mulțime	mulțime	1	0
software	software	2	2
matrice	-	0	1
indicatori	-	0	1

Se calculează indicatorul grad de încredere G, conform expresiei analitice:

$$G = \frac{\sum_{j=1}^{nsv} \alpha_j}{\sum_{j=1}^{nsv} \beta_j}$$

unde:

$\alpha_j$  – variabilă booleană care ia una din valorile:

1, pentru utilizarea cuvântului  $c_{ij}$  în textul T și în textul bazat TT<sub>i</sub>;

0, când cuvântul  $c_{ij}$  lipsește cel puțin dintr-un text, T sau TT<sub>i</sub>;

$\beta_j$  – variabilă booleană egală cu 1 atunci când cuvântul  $c_{ij}$  din subvocabularul SV este prezent în textul T;

nsv – numărul de cuvinte din subvocabularul SV.

Folosind datele din tabelul 1, se construiește matricea variabilelor booleene  $\alpha_j$  și  $\beta_j$ , tabelul 2.

Modul de utilizare a cuvintelor din  $TT_i$ 

Tabel 2

Vocabular	Subvocabular	T	$TT_i$	$\alpha_j$	$\beta_j$
ortogonalitate	ortogonalitate	2	1	1	1
metrici	metrici	0	0	0	0
alfabet	alfabet	2	2	1	1
mulțime	mulțime	1	0	0	1
software	software	2	2	1	1
matrice	-	0	1	0	0
indicatori	-	0	1	0	0
<b>Total</b>	-	<b>8</b>	<b>7</b>	<b>3</b>	<b>4</b>

Gradul de încredere pentru datele din tabelul 2 este  $G = 0,75$ . Aceasta înseamnă că textul bazat  $TT_i$  acoperă într-o proporție de 75% domeniul specificat în entitatea T.

Se consideră textul:

$TT_j = \langle \text{rezultă indicatori de asemănare a simbolurilor, precum și matrice de asemănare pentru simboluri} \rangle$

Analiza textului  $TT_j$  conduce la frecvențele de apariție înregistrate în tabelul 3.

Frecvențele de utilizare a cuvintelor în T și  $TT_j$ 

Tabel 3

Vocabular	Subvocabular	T	$TT_j$
ortogonalitate	ortogonalitate	2	0
metrici	metrici	0	0
alfabet	alfabet	2	0
mulțime	mulțime	1	0
software	software	2	0
matrice	-	0	1
indicatori	-	0	1

Se construiește tabelul variabilelor booleene  $\alpha_j$  și  $\beta_j$ .

#### Modul de utilizare a cuvintelor din $TT_j$

Tabel 4

Vocabular	Subvocabular	T	$TT_j$	$\alpha_j$	$\beta_j$
ortogonalitate	ortogonalitate	2	0	0	1
metrici	metrici	0	0	0	0
alfabet	alfabet	2	0	0	1
mulțime	mulțime	1	0	0	1
software	software	2	0	0	1
matrice	-	0	1	0	0
indicatori	-	0	1	0	0
<b>Total</b>	-	<b>8</b>	<b>2</b>	<b>0</b>	<b>4</b>

Pentru valorile din tabelul 4, se calculează  $G = 0$ . Acesta înseamnă că textul  $TT_j$  nu are nimic în comun cu T, nefiind utilizate cuvinte din subvocabularul SV.

Pentru textul:

*$TT_h = < \text{rezultă o mulțime de indicatori de asemănare alfabet și software, precum și matrice ortogonalitate alfabet și software} >$*

se construiește matricea:

#### Modul de utilizare a cuvintelor din $TT_h$

Tabel 5

Vocabular	Subvocabular	T	$TT_h$	$\alpha_j$	$\beta_j$
ortogonalitate	ortogonalitate	2	1	1	1
metrici	metrici	0	0	0	0
alfabet	alfabet	2	2	1	1
mulțime	mulțime	1	1	1	1
software	software	2	2	1	1
matrice	-	0	1	0	0
indicatori	-	0	1	0	0
<b>Total</b>	-	<b>8</b>	<b>8</b>	<b>4</b>	<b>4</b>

Gradul de încredere pentru calculat pentru  $TT_h$  este  $G = 1$ . Textul bazat  $TT_h$  acoperă într-o proporție de 100% domeniul specificat în entitatea  $T$ .

Gradul de externalizare evidențiază măsura în care în textul  $TT_i$  sunt utilizate cuvinte din vocabularul  $V$  care nu aparțin subvocabularului,  $V \setminus SV$ .

Se numără cuvintele distincte care aparțin mulțimii  $V \setminus SV$ . Se consideră variabila booleană  $\theta_j$  care ia una din valorile:

- 1, când cuvântul  $c_{ij}$  aparține mulțimii  $V \setminus SV$ ;
- 0, în caz contrar.

Gradul de externalizare  $G_e$  se determină conform relației:

$$G_e = \frac{\sum_{i=1}^{nv} \theta_i}{\sum_{j=1}^{nsv} \alpha_j}$$

unde:

- $\theta_j$  – variabilă booleană pentru apartenența cuvântului  $c_{ij}$  la mulțimea  $V \setminus SV$ ;
- $\alpha_j$  – variabilă booleană pentru apartenența cuvântului  $c_{ij}$  la textele  $T$  și  $TT_i$ ;
- $nsv$  – numărul de cuvinte al subvocabularului  $SV$ ;
- $nv$  – numărul de cuvinte al vocabularului  $V$ .

Pentru valorile din tabelul 2,  $G_e = 0,66$ . În cea mai mare parte, cuvintele utilizate sunt din subvocabularul  $SV$ . Astfel, entitatea  $TT_i$  este una bazată.

Dacă  $G_e$  tinde către zero, atunci textul  $TT_i$  este construit folosind restricțiile impuse prin definirea subvocabularului  $SV$ .

Dacă  $G_e$  tinde către 1 rezultă că textul  $TT_i$  operează cu cuvinte care aparțin în cea mai mare parte vocabularului  $V$ . Subvocabularul nu este bază în generarea de text, condiție necesară ca textul  $TT_i$  să se numească text bazat.

### **Omogenitatea și ortogonalitatea textelor bazate**

Textele construite pe subvocabularul  $SV$  au ca obiectiv devenirea de repere în elaborarea de texte bazate.

Pentru elaborarea de referate asupra unei lucrări, se alcătuieste un subvocabular din textul lucrării literare. Indicatorii  $G$  și  $G_e$  cuantifică măsura în care a fost abordată opera literară, în raport cu modul în care autorii referatelor utilizează cuvintele subvocabularului.

Textele bazate sunt actualizate pentru modul în care utilizează cuvintele textului  $T$ , respectiv, cuvintele din subvocabularul  $SV$ .

Culegerile tipărite de analize literare presupun o mulțime de opere literare,  $T_1, T_2, \dots, T_{nl}$  și un subvocabular, unde  $nl$  reprezintă numărul de opere literare. Analizele literare propuse de autorii culegerilor tipărite sunt modele date sub forma textelor  $TT_1, TT_2, \dots, TT_{nl}$ ,  $nl = nb$ . Se formează perechile  $(T_1, TT_1), (T_2, TT_2), \dots, (T_{nl}, TT_{nl})$ . Fiecărei opere îi corespunde un model de analiză literară. Autorii culegerii tipărite de analize literare urmăresc maximizarea nivelului indicatorului  $G$  și minimizarea indicatorului  $G_e$ .

Cei care elaborează referate, lucrări de clasă, teze folosind un text de bază  $T_i$  produc textele  $TT_{i1}, TT_{i2}, \dots, TT_{i,nb}$ . Prin calculul indicatorilor  $G$  și  $G_e$  se măsoară gradul de utilizare a cuvintelor din subvocabularul  $SV$ , respectiv gradul de utilizare a altor cuvinte decât cele din subvocabular.

Textele bazate trebuie să fie omogene în raport cu modul în care referă cuvintele textului de bază.

Se calculează un indicator agregat al omogenității textelor bazate,  $\bar{G}$ :

$$\bar{G} = nb \sqrt[nb]{\prod_{i=1}^{nb} g_i}$$

unde  $g_i$  reprezintă valoarea indicatorului de apartenență  $G$  pentru  $TT_{ki}$ .

Gradul de omogenitate a entității  $TT_{kj}$  în raport cu apartenența la subvocabularul  $SV$ , se determină conform relației:

$$G_O = \frac{g_j - \bar{G}}{\bar{G}}$$

În același timp, textele  $TT_{i1}, TT_{i2}, \dots, TT_{i,nb}$  trebuie să fie ortogonale.

Pentru:

$V = \langle \text{ortogonalitate metrici alfabet mulțime matrice indicatori software} \rangle$
---

$SV = \langle \text{ortogonalitate metrici alfabet mulțime} \rangle$
--



se consideră textele bazate:

$TT_i = \langle \text{rezultă indicatori asemănare alfabet și software, precum și matrice ortogonalitate alfabet și software} \rangle$

$TT_j = \langle \text{se definește ortogonalitate mulțime precum și ortogonalitate alfabet și software} \rangle$

Tabelul frecvențelor de utilizare a cuvintelor din vocabularul V și subvocabularul SV este:

### Frecvențele de utilizare a cuvintelor în $TT_i$ și $TT_j$

Tabel 6

Vocabular	Subvocabular	$TT_i$	$TT_j$
ortogonalitate	ortogonalitate	1	2
metrici	metrici	0	0
alfabet	alfabet	2	1
mulțime	mulțime	0	1
matrice	-	1	0
indicatori	-	1	0
software	-	2	1

Ortogonalitatea textelor bazate  $TT_i$  și  $TT_j$  în raport cu utilizarea cuvintelor din vocabularul V se determină conform expresiei:

$$H(TT_i, TT_j) = \frac{\sum_{i=1}^{nv} \gamma_i}{nv}$$

unde:

$\gamma_i$  – variabilă booleană ce ia una din valorile:

0, dacă  $C_k$  este utilizat în  $TT_i$  și  $TT_j$ ; de asemenea, pentru cazul în care  $c_k$  nu apare nici în  $TT_i$ , nici în  $TT_j$ ;

1, în caz contrar;

$nv$  – numărul de cuvinte din V.

Pentru valorile din tabelul 6,  $H(TT_i, TT_j) = 0,42$ .

Dacă  $H(TT_i, TT_j)$  tinde spre 1 înseamnă că textele bazate  $TT_i$  și  $TT_j$  sunt ortogonale. Dacă  $H(TT_i, TT_j)$  tinde spre zero, atunci entitățile analizate sunt aproape identice.

Pentru exemplul considerat, diferențele nu sunt semnificative. Se afirmă că textele bazate  $TT_i$  și  $TT_j$  fac parte din același domeniu.

### Software pentru analiza textelor bazate

Există un tezaur de fizică:

$TZ = \langle \text{optica, foton, ondulatoriu, energetic, unde, fotometria, electromotor, rezonanța, lumina, polarizare, dispersia, radiolocația, sunetul, osciloscop, rezistor, circuit, bobina, condensator, curent, oscilație, principiu, propagare} \rangle$

Din cuvintele tezaurului de fizică se extrag termenii pentru o programă analitică.

Termenii formează subvocabularul:

$SV = \langle \text{optica, ondulatoriu, energetic, radiolocația, sunetul, rezistor, principiu, bobina, condensator, electromotor, rezonanta, circuit} \rangle$

iar programa analitică este dată de:

$T = \langle \text{optica ondulatoriu principiu radiolocația circuit} \rangle$

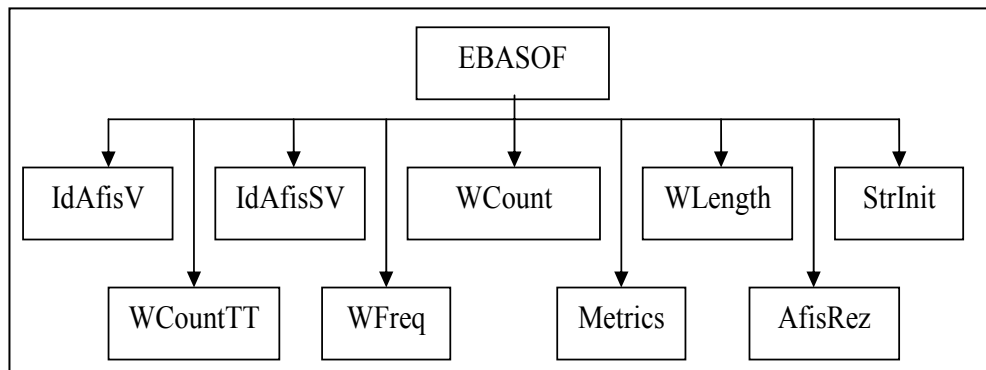
Autorii scriu manualele  $TT_1$ ,  $TT_2$ ,  $TT_3$  și acestea sunt texte bazate pentru că trebuie să aibă cuvintele din programă.

$TT_1 = \langle \text{circuit dispersie unde radiolocația} \rangle$

$TT_2 = \langle \text{principiu unde circuit bobina lumina rezistor foton} \rangle$

$TT_3 = \langle \text{optica propagare circuit ondulatoriu radiolocația bobina circuit condensator curent} \rangle$

Aplicația EBASOF are o structură modularizată, vezi figura 1.



**Figura 1 Structura aplicației EBASOF**

Determinarea indicatorilor privind gradele de încredere și de externalizare se realizează prin intermediul aplicației EBASOF, elaborată în limbajul C++. Aplicația cuprinde următoarele module:

**IdAfisV** realizează identificarea și afișarea cuvintelor din vocabularul V; constă în deschiderea fișierului care conține cuvintele vocabularului, parcurgerea fișierului pentru identificarea și afișarea cuvintelor; numele fișierului este furnizat de la tastatură;

**IdAfisSV** efectuează identificarea și afișarea cuvintelor dintr-un subvocabular SV<sub>i</sub>; presupune deschiderea fișierului care conține cuvintele din subvocabular și traversarea sa pentru identificarea și afișarea cuvintelor; numele fișierului se introduce de utilizator de la tastatură;

**WCount** realizează determinarea numărului de cuvinte din vocabular și subvocabular; constă în identificarea cuvintelor din vocabular și subvocabular, însoțită de contorizare; această se realizează prin identificarea și numărarea simbolurilor separator din șirul de caractere care formează entitatea subvocabular;

**WLength** implementează determinarea lungimilor cuvintelor din vocabular și subvocabular; lungimea cuvintelor se determină ca număr de caractere care sunt utilizate în construirea acestora;

**StrInit** permite inițializarea structurilor și verificarea apartenenței subvocabularului SV la vocabularul V; structurile utilizate pentru inițializare sunt de masive unidimensionale și bidimensionale; verificarea este necesară pentru a asigura calitate ridicată pentru analiza textelor bazate;

**WCountTT** realizează determinarea numărului de cuvinte din SV<sub>i</sub> utilizate în textele TT<sub>i</sub>; cuvintele subvocabularului SV<sub>i</sub> sunt identificate în textele TT<sub>i</sub>, date de utilizator de la tastatură; aceste frecvențe de apariție sunt utilizate în calculul metricilor de apartenență;

**WFreq** efectuează determinarea și afișarea frecvențelor de apariție a cuvintelor din V și SV<sub>i</sub> incluse în textele TT<sub>i</sub>; se determină numărul de cuvinte din V care apar în textele TT<sub>i</sub>; se afișează frecvențele de apariție ale cuvintelor din V și SV<sub>i</sub> sub formă tabelară;

**Metrics** realizează implementarea modelelor asociate metricilor; presupune pregătirea datelor de intrare prin considerarea elementelor structurale ale metricilor de includere și externalizare; modelele verifică apartenența cuvintelor din TT<sub>i</sub> care se regăsesc în V și SV<sub>i</sub> în raport cu cuvintele care se regăsesc numai în subvocabular sau care sunt în V și nu aparțin de SV<sub>i</sub>;

**AfisRez** efectuează afișarea rezultatelor; sunt furnizate valorile obținute prin aplicarea modelelor asociate metricilor pentru tabelele cu frecvențele de apariție construite.

Execuția aplicației EBASOF permite obținere gradelor de încredere:

$G(TT_1) = 0,40$
$G(TT_2) = 0,40$
$G(TT_3) = 0,80$

și gradele de externalizare:

$Ge(TT_1) = 0,29$
-------------------

$Ge(TT_2) = 0,25$
-------------------

$Ge(TT_3) = 0,57$
-------------------

pentru tezaurul de fizică T și textele bazate  $TT_1$ ,  $TT_2$  și  $TT_3$ .

Valorile indicatorului G sunt relativ ridicate. Interpretarea lor se rezumă la faptul că mare parte din cuvintele textelor  $TT_i$ ,  $i = 1, 2, 3$  fac parte din textul T.

Pentru valorile indicatorului  $G_e$ , cuvintele utilizate sunt din subvocabularul SV. Entitățile  $TT_i$ ,  $i = 1, 2, 3$  sunt entități bazate.

### Concluzii

Analiza textelor bazate este foarte importantă pentru determinarea asemănării a două texte având la baza același vocabular de bază sau același subvocabular.

Calculând gradul de încredere și gradul de externalizare al textelor se verifică gradul de proveniență al unui text dintr-un vocabular sau dintr-un alfabet.

Lucrările de specialitate comportă o mare utilizare a software-ului de analiză al textelor bazate, întrucât rezultatele furnizate de acesta probează apartenența lucrărilor la un domeniu de specialitate.

### Bibliografie

1. BALLOU, D. „Enhancing Data Quality in Data Warehouse Environments”, TAYI, G. K. *Communications of the ACM*, vol. 42, nr. 1, 1999, pp. 73 – 78
2. IVAN, I. *Entități text – dezvoltare, analiză, evaluare*, București, Editura POPA, M. ASE, 2005
3. IVAN, I. „Reingineria entităților text”, în *Revista Română de Informatică și Automatică*, vol. 15, nr. 2, 2005, pp. 15 – 28  
POPA, M. TOMOZEI, C.

4. IVAN, I.                    *Text Entities Representativeness*, Conferința Internațională  
POPA, M.                    „The Impact of European Integration on the National  
   Economy”, Babeș-Bolyai University, Cluj-Napoca, October 28  
   – 29, 2005
5. IVAN, I.                    *Operații pe entități text*, Conferința Științifică Internațională  
POPA, M.                    „Binomul sărăcie-bogație și integrarea României în Uniunea  
BOJA, C.                    Europeană”, Sibiu, 20 – 21 mai 2005, vol. 3, pp. 464 – 473
6. IVAN, I.                    *Clonarea informatică*, București, Editura ASE, 2003  
POPA, M.  
CAPISIZU, S.  
BREDA, L.  
FLORESCU, B.
7. POPA, M.S                „Text Entities Metrics”, în *Informatica Economică*,  
vol. 9, nr. 2, 2005, pp. 56 – 60
8. POPA, M.                *Structured Text Entities Metrics*, The Proceedings of „The 36th  
Informational Scientific Symposium of METRA”, București, 26 -  
27 mai 2005, pp. 486 – 491
9. POPA, M.                *Text Entities Topographic Characteristic*, Proceedings of the  
Seventh International Conference on Informatics in Economy  
„Information & Knowledge Age”, București,  
19 – 20 mai 2005, pp. 1115 – 1121
10. SMEUREANU,            *Programarea orientată obiect în limbajul C++*, București,  
I. DÂRDALĂ,                Editura CISON, 2002  
M.
11. \*\*\*                        [www.research.ibm.com](http://www.research.ibm.com)