Methodological aspects of data quality modeling

Otilia PÎRLOG
Ministerul Apărării Naţionale

Even though a fairly new field of economical science, data quality provides approaches that are both comprehensive and thorough. In order to state quality requirements and evaluate the final results, it is required to define the axiomatic system and introduce the objective and subjective aspects of the analyzed concept. The quality levels of data are also defined: attribute characteristics, dimensions and category. Thus, by investigating the quality levels, the category to which the deficiencies belong to is relieved. Improving the quality is associated with systems or processes belonging to the data category levels, the effort is focused towards the determination of the changes in the data acquisition and / or processing. A very important aspect in this field is represented by the data usage premises, the ultimate measure of the quality being the level of fulfillment of the end user's requirements.
Keywords: *data, quality, deficiency, usage, attribute, characteristic, dimension, category.*

Atributele de calitate ale datelor

Proiectarea conceptuală a colecţiilor de date vizează problematica aplicaţiei informatice, transpusă la nivelul entităţilor şi a formalizării relaţiilor dintre acestea. Maniera de abordare tradiţională a proiectării colecţiilor de date nu recurge la incorporarea explicită a aspectelor de calitate.

Următoarele problematici ale utilizării datelor conduc la concentrarea atenţiei asupra calităţii acestora:

- folosirea nu numai de către aplicaţia pentru care au fost iniţial proiectate;
- procesarea împreună cu date provenite din alte colecţii;
- punerea datelor la dispoziţia unor utilizatori noi, nefamiliarizaţi cu acestea.

Una din definiţiile calităţii datelor unanim acceptată este conformitatea la cerinţele exprimate de utilizatori. Din punct de vedere operaţional, aceasta este definită prin intermediul parametrilor şi indicatorilor corespunzători.

Un *indicator* este o dimensiune a calităţii datelor prin care se furnizează informaţii obiective despre aceasta. Astfel de indicatori pot fi: sursa, timpul de creare, metoda de colectare etc.

Scopul modelării calităţii datelor este acela de a permite utilizatorilor să-şi definească cerinţele privind calitatea datelor şi să stabilească indicatorii de calitate care sunt adecvaţi pentru o aplicaţie dată.

Valoarea unui indicator este o caracteristică măsurabilă a datelor stocate.

Valoarea unui parametru reprezintă valoarea determinată pentru un parametru de calitate, bazată pe interpretarea valorii indicatorilor aferenţi.

Translatarea valorii indicatorilor în valori parametrice este subiectivă şi se realizează prin funcţii utilizator, expresia acestora fiind dependentă de context.

Un *atribut* este un termen bidimensional, care pune în corelaţie parametrii şi indicatorii calitativi. Relaţia dintre atribute, parametri şi indicatori este prezentată în Figura 1.

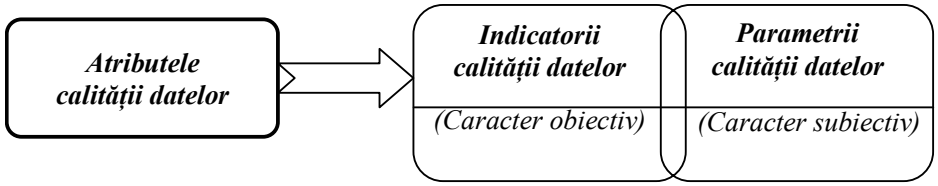

Fig.1. Definirea atributelor de calitate ale datelor

Cerinţele calităţii datelor specifică acei indicatori a căror valoare este necesar să se înscrie între anumite limite specificate, astfel încât să se asigure punerea la dispoziţia utilizatorilor a unor date cu un nivel calitativ acceptabil.

Administratorul datelor este o persoană (sau sistem) a cărei responsabilitate este asigurarea îndeplinirii de către datele stocate a cerinţelor exprimate.

În general, utilizatori diferiţi au cerinţe diferite de calitate, iar calitatea unor date diferite poate varia între limite extrem de largi. De aici decurg şi cele două categorii de premise de studiu referitoare la calitatea datelor, respectiv modelarea şi utilizarea datelor.

Premise privind modelarea calităţii

Modelarea calităţii datelor este o extensie a metodologiilor tradiţionale de modelare a datelor. Diferenţa dintre acestea constă în faptul că în timp ce modelarea datelor se adresează structurii şi semanticii datelor, modelarea calităţii datelor vizează aspectele structurale şi semantice aferente calităţii datelor.

Premisa a.1. Relaţia dintre atributele aplicaţiei şi ale calităţii

Nu este întotdeauna necesar ca atributele aplicaţiei şi ale calităţii să fie distincte. Pot fi identificate două domenii diferite de activitate din acest punct de vedere: utilizarea datelor şi administrarea calităţii.

Astfel, valoarea unui indicator poate să fie folosită în contextul procesării datelor pentru obţinerea unor informaţii livrabile la utilizator sau care să fie trecute într-o nouă etapă de procesare. Aceeaşi valoare poate fi interpretată printr-o funcţie dedicată pentru stabilirea nivelului de dezvoltare al unui parametru calitativ.

Unele informaţii cuprind detalii referitoare la procesul de prelucrare al datelor. Această categorie reprezintă numai indicatori de calitate.

Premisa a.2. Atributele de calitate nu sunt ortogonale

Calitatea datelor poate fi reliefată prin parametri extrem de diferiţi, funcţie de contextul în care acestea sunt analizate. Unii parametri sunt interdependenţi, necesităţile de interpretare impunând reliefarea unor aspecte intercorelate dintr-un anumit unghi de referinţă. Exemplul cel mai elocvent pentru acest caz îl constituie interdependenţa dintre oportunitatea şi volatilitatea datelor.

Premisa a.3. Heterogenitatea şi ierarhizarea calitativă a datelor furnizate

Calitatea datelor aparţinând unor colecţii, entităţi, atribute sau valori distincte poate fi diferită. De aici necesitatea menţinerii parametrilor de calitate în limitele admisibile stabilite funcţie de prelucrările ulterioare ale datelor şi a gradului de interdependenţă al acestora.

Premisa a.4. Indicatori de calitate recursivi

O premiză esenţială în studiul calităţii este stabilirea setului de indicatori de calitate relevanţi. Apare astfel necesitatea definirii *meta-indicatorilor de calitate* ca un obiectiv de bază în dezvoltarea unei perspective calitative în analizele ulterioare.

Premise privind utilizarea datelor

Utilizatorii datelor au opinii şi cerinţe diferite referitoare la calitatea datelor. Utilizatorul unui sistem informatic dat poate cunoaşte calitatea datelor acestuia. Prin exportarea acestui sistem la alţi utilizatori sau prin combinarea cu informaţii de calităţi diferite, calitatea datelor poate deveni necunoscută, rezultând necesităţi diferite privind atributele de calitate acoperitoare pentru utilizatorii şi domeniile aplicaţiei.

Premisa b.1. Specificitatea atributelor funcţie de utilizator

Parametrii şi indicatorii de calitate pot diferi de la un utilizator la altul. Astfel, pentru un manager parametrul de calitate critic pentru un raport de cercetare îl poate constitui costul, în timp ce pentru un comerciant credibilitatea şi oportunitatea sunt mult mai importante.

Premisa b.2. Utilizatorii au standarde de calitate diferite

Nivelurile acceptabile de calitate a datelor pot diferi în limite foarte largi de la un utilizator la altul. Funcţie de natura activităţii desfăşurate şi a datelor vehiculate, un acelaşi nivel al atributelor de calitate poate fi considerat acceptabil sau nu.

Cel mai adesea se face referire în acest sens la oportunitatea datelor. Timpul de întârziere de la preluarea datelor în sistemul informatic şi până la punerea lor la dispoziţia beneficiarilor poate să reprezinte un parametru critic pentru un anumit utilizator sau conex pentru altul.

Premisa b.3. Atribute şi standarde diferite pentru un singur utilizator

Funcţie de contextul folosirii datelor, acelaşi utilizator poate avea cerinţe diferite privind colecţiile, entităţile, atributele sau valorile propriu-zise ale datelor. Prin definirea cu claritate a acestora se oferă posibilitatea concentrării şi distribuirii efortului conform priorităţilor stabilite.

Dimensiunile şi categoriile calităţii datelor

În urma studiului efectuat de Richard Wang, Diane Strong şi Lisa Guarascio [Wang96] au fost identificate principalele atribute de calitate a datelor. Cercetarea a fost continuată prin analiza factorială a unui număr de 118 variabile, delimitându-se cincisprezece dimensiuni ale calităţii datelor.

În timp ce atributele susţin nivelul cel mai scăzut pentru care problemele datelor pot fi identificate şi înţelese, dimensiunile facilitează înţelegerea la cel mai înalt nivel. Altfel spus, atributele oferă mecanismul cu nivelul cel mai scăzut începând de la care problemele datelor devin evidente, în timp ce dimensiunile susţin condiţiile prin care se previne apariţia problemelor datelor. Aceasta implică evaluarea şi identificarea problemelor datelor prin analizarea atributelor şi apoi stabilirea domeniului deficienţelor constatate, prin gruparea atributelor funcţie de dimensiunile de apartenenţă. Oprirea analizei şi extragerea concluziilor la nivelul atributelor împiedică reliefarea problemelor de bază, sistematice.

Funcţie de contextul utilizării şi al scopului iniţial al proiectării datelor, nivelul de importanţă acordat diverselor dimensiuni ale acestora poate varia extrem de mult. Un alt aspect important îl reprezintă dependenţa calităţii datelor de procesele care le generează, transformă şi modifică.

Deşi datele pot avea probleme de calitate care aparţin uneia dintre dimensiuni în timp ce sunt satisfăcătoare pentru altă dimensiune, o singură cauză de bază poate afecta dimensiuni multiple. Prin gruparea atributelor în dimensiuni s-a facilitat structurarea problemelor calităţii datelor. Wang, Strong şi Guarascio au observat că dimensiunile formează câteva familii, sau categorii, aşa cum se prezintă în Tabelul 1. Fiecare din aceste categorii denotă existenţa unei omisiuni a procesului sau a unei defecţiuni în procesarea curentă referitoare la un anumit aspect major al calităţii datelor.

Categoriile reprezintă al treilea nivel de înţelegere al problemelor calităţii datelor. Pot fi depistate astfel lacunele de procesare a căror apariţie sau existenţă a fost facilitată de condiţiile asigurate la nivelul anterior. Conducerea analizei în acest caz permite identificarea şi interpretarea cauzelor de bază legate de procesele de achiziţionare şi manipulare a datelor.

Calitatea intrinsecă denotă faptul că datele au valoare prin însăşi existenţa lor. Aici se include nu numai acurateţea şi credibilitatea, care sunt evidente pentru specialiştii din domeniu, dar şi obiectivitatea şi reputaţia. Astfel se sugerează că, în mod contrar accepţiunii tradiţionale, consumatorii de date privesc de asemenea obiectivitatea şi reputaţia drept părţi integrante ale calităţii intrinseci a datelor.

Contextul în care vor fi folosite datele este funcţie de timp şi de consumatorii acestora. Consumatorul de date trebuie să specifice ce tip de lucrare va executa şi parametrii contextuali corespunzători. Astfel, în scopul de a se furniza informaţii oportune şi de acurateţe, cercetătorii din domeniul aviaţiei militare au recunoscut necesitatea incorporării explicite a calităţii contextuale a datelor în sistemele informatice: localizare, rezoluţie, condiţii atmosferice, tip de obiectiv.

Tabelul 1. Categoriile calităţii datelor

Categoria	Dimensiunile	Deficienţe de proces
Intrinsecă	Acurateţe, Obiectivitate, Credibilitate, Reputaţie	La crearea datelor care corespund valorilor actuale sau adevărate.

Contextuală	Valoare Adăugată, Relevanţă, Oportunitate, Completitudine, Cantitate adecvată de date	Cu privire la obţinerea datelor pertinente necesităţilor utilizatorului.
Reprezentaţională	Interpretabilitate, Uşurinţa înţelegerii, Consistenţa reprezentării, Conciziunea reprezentării	Referitoare la furnizarea unor date inteligibile şi clare, compatibile cu datele arhivate.
Raportată la abordabilitate	Accesibilitate, Securitatea accesului	Pentru asigurarea disponibilităţii rapide şi obţinerea datelor.

Calitatea reprezentaţională include aspecte legate de formatul datelor (reprezentare concisă şi consistentă) şi înţelesul datelor (interpretabilitate şi uşor de înţeles). Din punct de vedere al beneficiarului, datele sunt de calitate nu numai dacă sunt reprezentate concis şi consistent, dar posedă şi caracteristici de interpretabilitate şi uşor de înţeles.

Calitatea raportată la abordabilitate

Datorită restricţiilor fizice ale accesării datelor (copii hard în loc de date on-line), în etapa de început a studiului calităţii datelor s-a recurs la tratarea distinctă a accesibilităţii. Schimbarea radicală a condiţiilor de utilizare şi a performanţelor calculatoarelor au modificat opiniile utilizatorilor, în sensul considerării accesibilităţii ca fiind un aspect important al calităţii datelor.

Analiza cerinţelor de calitate

Analiza cerinţelor de calitate, *modelarea datelor*, reprezintă un proces similar cu cel al analizei tradiţionale, dar cu concentrarea efortului pe aspectele calitative ale datelor. Se poate stabili astfel o paralelă între ciclul de viaţă al datelor şi metodologia de determinare a cerinţelor.

Rezultatul final al modelării calităţii datelor, *schema calităţii*, documentează atât cerinţele de date ale aplicaţiei cât şi problemele calităţii datelor considerate importante de către echipa de proiectare. Diagrama din Figura 2 ilustrează această metodologie. Pentru fiecare etapă de analiză sunt evidenţiate intrarea, procesarea şi ieşirea corespunzătoare.

Deoarece una din restricţiile majore ale asigurării calităţii datelor impune stabilirea numai a acelor date care sunt strict necesare pentru aplicaţia curentă, prima etapă nu reprezintă activitatea tradiţională de determinare a cerinţelor de date pentru o anumită aplicaţie informatică.

Integrarea indicatorilor nu reprezintă o simplă comasare a subseturilor identificate pentru aplicaţiile particulare. Prin examinarea determinărilor reciproce ale acestora, se elimină redundanţele şi inconsistenţele rezultate iniţial.

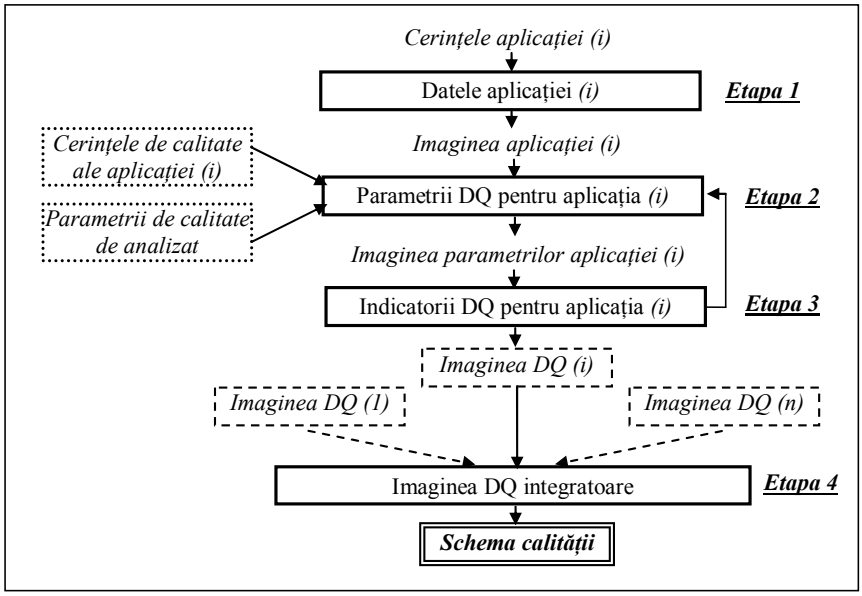

Fig.2. Procesul modelării calităţii datelor

Abordarea deficienţelor

Îmbunătăţirea calităţii datelor este asociată cu sistemele sau procesele ale căror deficienţe au fost reliefate la nivelul categoriilor din ierarhia calităţii datelor. Aceasta deoarece investigarea dimensiunilor scoate în evidenţă cărei categorii îi corespund deficienţele identificate. Efortul ultim este îndreptat spre realizarea schimbărilor corespunzătoare în procesele de achiziţionare şi / sau manipulare ale datelor. Acestora le sunt puse în corespondenţă metode performante, de achiziţionare eficientă a datelor, sau soluţii tehnologice, cu impact asupra procesării datelor. Relaţiile dintre deficienţele reliefate la nivelul categoriilor calităţii datelor şi abordările corespunzătoare de îmbunătăţire sunt prezentate în Figura 3.

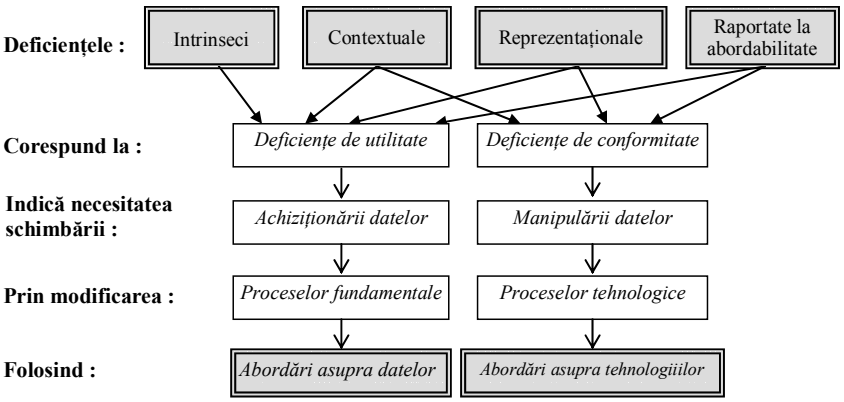

Fig.3. Relaţia dintre deficienţe şi abordările de îmbunătăţire

Aşa cum se arată în Figura 3, există două categorii de bază de îmbunătăţiri, care corespund necesităţii schimbării proceselor fundamentale şi a schimbării proceselor tehnologice.

Abordările asupra datelor constau în analiza economică, statistică, modelarea şi simularea datelor primare, ingineria cerinţelor precum şi modalităţi specifice care vizează diminuarea efectelor negative datorate acţiunii anacronice a factorilor umani. Abordările tehnologice sunt impuse ca urmare a instalării de echipamente noi, implementării unor produse informatice sau dezvoltării de noi interfeţe.

Pentru a îmbunătăţi utilitatea datelor este necesar să se implementeze soluţii asupra datelor, deoarece schimbările tehnologice nu afectează procesele de achiziţie. Deşi este evident că utilitatea datelor nu poate fi îmbunătăţită independent de procesele care le generează, abordările tradiţionale pentru rezolvarea problemelor calităţii datelor au vizat aplicarea strictă a unor soluţii tehnologice.

Bibliografie

[Bech00] - Bechtel Hanford Inc. Procedure: *Data Quality Objectives*, BHI-EE-01, Environmental Investigations Procedure: Procedure Number 1.2; Revision 3; September 30, 1999.

[Dill00] - Dillman, D. A.: Mail and Internet Surveys: *The Total Design Method*, New York: John Wiley & Sons, 2000.

[EPA00] - EPA: *Guidance for the Data Quality Objectives Process*; Office of Environmental Information, U.S. EPA, Washington D.C. 20460, August 2000.

[Han02] - Han, J., Kamber, M.: *Data Mining: Concepts and Techniques,* Publisher: Wiley, John & Sons, Incorporated, March 2002.

[Ivan99] - Ivan, I., Noşca, Gh., Tcaciuc, S., Pîrlog, O., Căciulă, R.: *Calitatea datelor,* Inforec, Bucureşti, 1999.

[Redm01] - Redman, T.C., Daugherty, M., Daugherty, M.: *Data Quality: The Field Guide,* Publisher: Elsevier Science & Technology Books, January 2001.

[Wang96] - Wang,R., Strong,D. and L.Guarascio:, *"Beyond accuracy: What data quality means to data consumer"*, Journal of Management Information Systems, Spring 1996.

[Wang01] - Wang, R.Y., Ziad, M., Lee, Y.W.: *Data Quality*, Kluwer, 2001.