Evaldas LEBEDYS

IMPLEMENTATION OF DATA QUALITY REQUIREMENTS IN INFORMATION SYSTEMS

Summary of Doctoral Dissertation
Technological Sciences, Informatics Engineering (07T)

1421

Vilnius VGTU LEIDYKLA TECHNIKA **2007**

VILNIUS GEDIMINAS TECHNICAL UNIVERSITY

Evaldas LEBEDYS

IMPLEMENTATION OF DATA QUALITY REQUIREMENTS IN INFORMATION SYSTEMS

Summary of Doctoral Dissertation
Technological Sciences, Informatics Engineering (07T)

Vilnius LEIDYKLA TECHNIKA 2007

Doctoral dissertation was prepared at Vilnius Gediminas Technical University in 2003–2007.

Scientific Supervisor

Prof Dr Olegas VASILECAS (Vilnius Gediminas Technical University, Technological Sciences, Informatics Engineering – 07T).

The dissertation is being defended at the Council of Scientific Field of Informatics Engineering at Vilnius Gediminas Technical University:

Chairman

Prof Dr Habil Gintautas DZEMYDA (Institute of Mathematics and Informatics, Technological Sciences, Informatics Engineering – 07T).

Members:

Prof Dr Habil Petras Gailutis ADOMĖNAS (Vilnius Gediminas Technical University, Technological Sciences, Informatics Engineering – 07T),

Assoc Prof Dr Romas BARONAS (Vilnius University, Physical Sciences, Informatics – 09P),

Prof Dr Habil Romualdas BAUŠYS (Vilnius Gediminas Technical University, Technological Sciences, Informatics Engineering – 07T),

Assoc Prof Dr Rimantas BUTLERIS (Kaunas University of Technology, Technological Sciences, Informatics Engineering – 07T).

Opponents:

Assoc Prof Dr Algis SAULIS (Vilnius Gediminas Technical University, Technological Sciences, Informatics Engineering – 07T),

Prof Dr Habil Rimantas ŠEINAUSKAS (Kaunas University of Technology, Technological Sciences, Informatics Engineering – 07T).

The dissertation will be defended at the public meeting of the Council of Scientific Field of Informatics Engineering in the Senate Hall of Vilnius Gediminas Technical University at 9 a. m. on 27 November 2007.

Address: Saulėtekio al. 11, LT-10223 Vilnius, Lithuania.

Tel.: +370 5 274 4952, +370 5 274 4956; fax +370 5 270 0112;

e-mail: doktor@adm.vgtu.lt

The summary of the doctoral dissertation was distributed on 26 October 2007.

A copy of the doctoral dissertation is available for review at the Library of Vilnius Gediminas Technical University (Saulėtekio al. 14, LT-10223 Vilnius, Lithuania) and at the Library of Institute of Mathematics and Informatics (Akademijos g. 4, LT-08663 Vilnius, Lithuania).

VILNIAUS GEDIMINO TECHNIKOS UNIVERSITETAS

Evaldas LEBEDYS

DUOMENŲ KOKYBĖS REIKALAVIMŲ REALIZAVIMAS INFORMACINĖSE SISTEMOSE

Daktaro disertacijos santrauka
Technologijos mokslai, informatikos inžinerija (07T)

Vilnius LEIDYKLA TECHNIKA 2007

Disertacija rengta 2003–2007 metais Vilniaus Gedimino technikos universitete.

Mokslinis vadovas

prof. dr. Olegas VASILECAS (Vilniaus Gedimino technikos universitetas, technologijos mokslai, informatikos inžinerija – 07T).

Disertacija ginama Vilniaus Gedimino technikos universiteto Informatikos inžinerijos mokslo krypties taryboje:

Pirmininkas

prof. habil. dr. Gintautas DZEMYDA (Matematikos ir informatikos institutas, technologijos mokslai, informatikos inžinerija – 07T).

Nariai:

prof. habil. dr. Petras Gailutis ADOMĖNAS (Vilniaus Gedimino technikos universitetas, technologijos mokslai, informatikos inžinerija – 07T),

doc. dr. Romas BARONAS (Vilniaus universitetas, fiziniai mokslai, informatika – 09P),

prof. habil. dr. Romualdas BAUŠYS (Vilniaus Gedimino technikos universitetas, technologijos mokslai, informatikos inžinerija – 07T),

doc. dr. Rimantas BUTLERIS (Kauno technologijos universitetas, technologijos mokslai, informatikos inžinerija – 07T).

Oponentai:

doc. dr. Algis SAULIS (Vilniaus Gedimino technikos universitetas, technologijos mokslai, informatikos inžinerija – 07T),

prof. habil. dr. Rimantas ŠEINAUSKAS (Kauno technologijos universitetas, technologijos mokslai, informatikos inžinerija – 07T).

Disertacija bus ginama viešame Informatikos inžinerijos mokslo krypties tarybos posėdyje 2007 m. lapkričio 27 d. 9 val. Vilniaus Gedimino technikos universiteto senato posėdžių salėje.

Adresas: Saulėtekio al. 11, LT–10223 Vilnius, Lietuva.

Tel.: (8 5) 274 4952, (8 5) 274 4956; faksas (8 5) 270 0112;

el. paštas doktor@adm.vgtu.lt

Disertacijos santrauka išsiuntinėta 2007 m. spalio 26 d.

Disertaciją galima peržiūrėti Vilniaus Gedimino technikos universiteto bibliotekoje (Saulėtekio al. 14, LT–10223 Vilnius, Lietuva) ir Matematikos ir informatikos instituto bibliotekoje (Akademijos g. 4, LT–08663 Vilnius, Lietuva).

VGTU leidyklos „Technika“ 1421 mokslo literatūros knyga.

General characteristic of the dissertation

Topicality of the problem. Even after all data management procedures are performed data in the databases may become corrupt. Thus constraints prohibiting the entry of erroneous data are not enough. Some of constraints may be represented in system models and the existing modelling tools allow automated implementation of these constraints in active database management systems. There are domain oriented tools suitable for data validation and clarification, but these tools require the manual specification of data quality requirements. Data quality requirements represented in system models can not be used for data validation. The repeated manual specification of data quality requirements leads to inconsistency of requirements, loss of human and financial resources.

Topicality of the problem. The importance and criticality of data quality is widely discussed in recent publications. Many external and internal factors impact the quality of data. Patterns for data quality problems identification are proposed in literature. The complexity of data quality control is common problem in every application domain, because data quality is multidimensional concept. Nearly 200 desirable features of qualitative data exist, thus making the control of data quality even more complex. Data quality assurance methods are defined and implemented in commercial tools, but in every application domain the concept of data quality is different and data quality control procedures have to be defined manually. Implementation of business rules in software systems as data quality assurance routines is common and not problematic by itself. Integrity constraints, stored procedures and triggers are used for years to increment data quality. The lack of mechanisms to record, review and assess the rules is the starting–point of the problems faced within data quality management.

Data validation is the first step in assessing data quality. Data validation is performed considering the requirements for data quality. Most data quality requirements are specified in system models while designing the system. Requirements express the rules that data should meet. Previous analysis showed, that a full and detail system model contains the rules that define most of data quality requirements that might be derived from model and implemented in data quality procedures.

Aim and tasks of the work. The main goal of this work is to improve the process of data quality assurance and propose the method for automated implementation of data quality requirements, expressed as business rules, in software systems.

The following tasks have to be implemented to reach the main goal:

1. The methods of data quality assurance presented in the literature have to be analysed.
2. Languages used to model business rules expressing data quality requirements have to be analysed. The methods of business rules implementation in software systems have to be analysed.
3. The method for automated implementation of business rules in software systems has to be proposed. The proposed method should improve data quality assurance.
4. A model of a software system has to be designed. Software systems has to be created on the basis of the designed prototype system model.
5. An experiment has to be performed to evaluate if the proposed method is suitable for automated business rules implementation in software systems and the created software system realises the proposed method.

Scientific novelty. The novelty of dissertation is the use of system models represented by UML for implementation of business rules in software systems. At the moment variety of methods and commercial tools are available that can be used to model business systems and implement data integrity constraints through the functionality of active database management systems. Unfortunately, these tools do not support data validation – the implementation of business rules as integrity constraints, triggers stored procedures is used only to avoid entry of erroneous data into the database. Regardless the use of data quality checks at the entry of data into the database, errors in data exist. The application of business rules approach in data quality assurance is widely discussed in the publications of recent years. Currently only domain specific data management tools support data validation, but these tools do not support system modelling at all or are suitable to model only some aspects of system. Therefore, there are no tools that support both system modelling and data validation.

Methodology of research

1. Comparative research and library research methods were used while analysing the existing business rules modelling languages, the methods of business rules implementation in software systems and the methods of data quality assurance.
2. The results of analysis were summarised and the approach was expounded using the research generalisation and logical induction methods.
3. The proposed method was implemented using the constructive research method.
4. An experiment was performed using the experimental research method.

Practical value. The results of the research demonstrated the practicality of the proposed method and created software prototype. The method for automated business rules implementation improved the development of information systems on the basis of business rules. Business rules can be derived from system models represented by UML. The derived business rules can be expressed as data quality examination queries and implemented using the created software prototype. The derived business rules can also be represented in natural language. The use of automated business rules implementation in data quality assurance, makes data quality control accessible for the business audience. The centralist management of business rules and data quality requirements makes data quality control simple.

Defended propositions

1. Data quality requirements are embodied in business rules that are represented in system models.
2. Business rules that embody data quality requirements can be derived from system models and placed in business repository.
3. Data quality requirements that are represented in system models can be implemented in software systems in the automated way.

The scope of the scientific work. The scientific work consists of the general characteristic of the dissertation, 4 chapters, conclusions, list of literature, list of publications and addenda. The total scope of the dissertation – 110 pages, 41 pictures, 12 tables and 3 addenda.

1. Business rules in information systems

At the moment variety of methods and commercial tools are available that can be used to model business systems and implement data integrity constraints through the functionality of active database management systems. Unfortunately, these tools do not support data validation – the implementation of business rules as integrity constraints, triggers stored procedures is used only to avoid entry of erroneous data into the database. Regardless the use of data quality checks at the entry of data into the database, errors in data exist. The application of business rules approach in data quality assurance is widely discussed in the publications of recent years. Currently only domain specific data management tools support data validation, but these tools do not support system modelling at all or are suitable to model only some aspects of system. Therefore, there are no tools that support both system modelling and data validation. Errors in software, unintended access to data and other considerations may be the sources of errors in data. These circumstances may

be crucial for the quality of data in certain domains, such as statistical data processing, clinical trials or telecommunications. Besides, even if the data are erroneous it may not be changed or rejected at the entry in the database in certain domains. For example, data have to be entered into the database exactly as it was collected in clinical trails. In this case, only after data are in the database, data validation can start to list the discrepancies and get confirmation from the origin of data. The need of the tool for validation of data is obvious. There are commercial and non commercial tools supporting validation of data, but the functionality of these tools is limited to the manual entry of data validation rules by the user. As system models can be used to retrieve domain knowledge, we state that a part of business rules can be derived from system models and implemented as data validation scripts.

Constraints in software systems were used to avoid errors in data for many years. Constraints are effective to avoid user errors and are widely used up until now. Although, constraints are implemented in most of information systems, errors in data still occur. Constraints are not suitable to eliminate errors that occur due to erroneous program code, security problems and unexpected access to the database, inaccurate mass data updates, inadequate data representation. Data may be erroneous due its origin and data processing steps, for example genome data. The whole of desirable data characteristics are analysed in the context of data quality. Although, "fitness for use" is the most widely adopted concept of data quality. Data quality can be defined as a lack of intolerable defects in data. Spatial data quality is evaluated in regard to lineage, positional, attribute and temporal accuracy, completeness and logical consistency.

M. Strong, W. Lee and R. Wang define the following data quality categories and dimensions:

1. Intrinsic DQ – Accuracy, Objectivity, Believability, Reputation.
2. Accessibility DQ – Accessibility, Access security.
3. Contextual DQ – Relevancy, Value-Added, Timeliness, Completeness, Amount of data.
4. Representational DQ – Interpretability, Ease of understanding, Concise, representation, Consistent representation.

2. The use of business rules represented in UML diagrams for automation of data quality assurance

Data validation is the first step in assessing data quality. Data validation is an analytic and domain-specific process that extends the evaluation of data beyond method, procedural, or contractual compliance to determine the analytical quality of a specific data set. Data validation is performed considering the requirements for data quality. Requirements express the rules

that data should meet. We state, that a full and detail system model contains the rules the define most of data quality requirements. Thus, data quality requirements might be derived, if system model is available. Generally the rules of the following types may be derived from system models and implemented in data validation procedures:

1. The checks of mandatory values, so called not null constraints.
2. The checks of the ranges of values.
3. If … then … else rules.
4. The cross checks between values of different items.

The primary intention, the model is created with, is to depict the system from the system owner point of view. Thus, application domain model reflects consumers point of view and data quality requirements derived from system model represent consumer's desirable data features.

Previous analysis showed that business rules might be derived from systems models represented by UML. The Unified Modelling Language was chosen for analysis, because it is a general-purpose modelling language that can be used with all major object and component methods, and that can be applied to all application domains. UML diagrams can be classified into three different classes:

1. Diagrams describing the roles and obligations of system users generally (Use Case diagrams).
2. Diagrams describing structural system aspects (class and object diagrams).
3. Diagrams describing the internal and external behaviour of system (state transition diagrams, activity diagrams, sequence and collaboration diagrams.

Business rules in Use Case diagrams mostly appear as statements describing system actor competence boundaries and obligations. Business rules in Use Case diagrams are represented as the statements defining the tasks of system users. Each Use Case can be depicted in details using sequence and collaboration diagrams. Sequence and collaboration diagrams represent how system actors act and exchange information to execute the tasks they are assigned. Sequence and collaboration diagrams include business rules describing the exact order of actions to be executed to perform a task. State transition diagrams are used to specify the sequences of changes of states of business objects. Event-Condition-Action (ECA) rules are mostly represented in state transition diagrams. UML activity diagrams can be used to model the logic of the operations captured by a use case or a few use cases. Activity diagrams represent both the basic sequence of actions s well as the alternate sequence of actions. ECA rules are mostly represented in activity diagrams. Class diagrams include rules that express constraints of business objects, properties of

relationships between business objects. The rest of UML diagrams are used to represent aspects of the development and implementation of software systems and are not analysed further.

3. The design and development of software prototype

The purpose of the system being created is automation of the business rules implementation in software systems as data validation queries. The software prototype is to be used for data quality assurance in application domains where the means of active database management systems are not applicable. The system being created should be suitable to perform the following tasks:

1. The import of data from system model represented by UML using Sybase® PowerDesigner to rule repository.
2. The import of business rules components represented in UML diagrams into the rule repository.
3. Representation of business rules as natural language sentences.
4. Generation of SQL queries for data quality examination.
5. Validation of data in the chosen database against the rules represented in the UML model.

The system must be able to import UML model components from model created using Sybase® PowerDesigner. Rule repository has to be implemented in the Microsoft® SQL Server active database management system.

The requirements were specified for the software prototype being created. The model of software prototype was designed using UML. Software prototype was developed considering the specified requirements and the model. Testing instructions were prepared and the created software system was tested to evaluate its conformity to requirements. Testing results were positive and software prototype was ready to use. An experiment was planned to check the efficiency of the proposed method and the created software prototype.

4. The experimental research of the proposed method

The experiment was performed in the following steps:

1. A simplified model of the income declaration system was designed. A subsystem for examination of data was chosen for analysis.
2. Business rules expressing data quality requirements were derived from the experiment model and stored in business rules repository. The created software prototype was used.
3. An experimental database was created for the analysed simplified income declaration system.

4. Test data were recorded in the created database.
5. The created software prototype was used to generate SQL queries from the business rules stored in rules repository.
6. The data stored in experimental database was validated against the rules derived form system models. The created software prototype was used.

The results of the experiment showed that automation of implementation of data quality requirements in software systems greatly improves the control of data quality. Data validation procedures can be created at least 1.5 times faster using the proposed method and the created software prototype.

5. General conclusions

After developing the method for automated implementation of data quality requirements, expressed as business rules, in software systems, as well as after performing a computer experiment and after verifying and checking it in a nature, there were formulated the following scientific and practical conclusions:

1. Analysis of recent researchers in data quality area showed that data quality is relevant for each organisation and due to its complexity is a problematic research area.
2. Researchers made a theoretical background of the relevant data quality issues. Variety of tools supporting data quality assurance are currently available in the market, but most of tools are domain specific and are mostly purposed for assuring data quality in data warehousing.
3. The research showed, that the available tools for data quality assurance are used to define data quality requirements manually and are centred on intuitive data quality approach.
4. On the basis of the previous research we decided that data quality requirements might be derived from system models represented by UML. Thus we proposed a method for automated data quality checking that is based on automatic derivation of data quality requirements from system models.
5. The analysis showed that the rules stored in rules repository can be used for automated generation of data quality check queries.
6. The research showed that the rules derived from system models can be used for data validation and representation of rules in natural language.
7. It was presented that the automation of data validation process makes this process understandable for business audience as rules are stored in the repository and can be expressed in natural language.
8. The experiment showed the advantage of the proposed method in comparison with the hand work based methods. The automated method is faster than the hand work based methods at least 1.5 times.

List of published works on the topic of the dissertation
In the reviewed scientific periodical publications

1. VASILECAS, O.; LEBEDYS, E. Application of business rules for data validation. *Information technologies and control*, Technologija, Kaunas, 2007, p. 273–277. ISSN 1392-124X.
2. VASILECAS, O.; LEBEDYS, E. Moving business rules from system models to business rules repository. *Infocomp*, June 2006, Vol 5, No 2, p. 11–17. ISSN 1807-4545.
3. VASILECAS, O.; LEBEDYS, E. Repository for Business Rules Represented in UML Diagrams. *Izvestia of the Belarusian Engineering Academy*, Minsk, 2005, Vol 1 No 19/2, p. 187–192.
4. VASILECAS, O.; LEBEDYS, E. The usage of system models for business rules modelling. *Informacijos mokslai* (*Information sciences)*, Vilnius University publishing house, 2005, p. 189–194. ISSN 1392–0561.
5. VASILECAS, O.; LEBEDYS, E. The use of formal languages for business rules realisation in information systems. *Lietuvos matematikos rinkinys (Lithuanian Mathematical Journal)*, Vilnius, Vo. 45, 2005, p. 174–179. ISSN 0132-2818.
6. VASILECAS, O.; LEBEDYS, E. Representation of Business Rules Using Formal Logic. *Scientific Proceedings of Riga Technical University*, 5th Series, Computer Science, Applied Computer Systems, Vol 22, Riga, RTU Publishing, 2005, p. 34–40. ISSN 1407-7493.
7. VASILECAS, O.; LEBEDYS, E. Realization of business rules visualized using formal logic sentences in Information Systems. *Lietuvos matematikos rinkinys* (*Lithuanian Mathematical Journal)*, Vilnius, Vol 44, 2004, p. 327–331. ISSN 0132-2818 (in Lithuanian).
8. LAUCIUS, J.; LEBEDYS, E.; VASILECAS, O. Realisation of ECA rules by ADBVS triggers. *Informacijos mokslai (Information sciences)*, Vilnius University publishing house, 2003, p. 129–133. ISSN 1392–0561 (in Lithuanian).

In the other editions

9. VASILECAS, O.; LEBEDYS, E. Application of business rules approach in clinical trials. In B. Rachev, A. Smirkarov (eds.). *Proc. of the International Conference on Computer Systems and Technologies "CompSysTech'06"*, Veliko Tarnovo, Bulgaria, 2006, p. II.8-1–II.8-6.
10. VASILECAS, O.; LEBEDYS, E. Business rules in clinical trials. In G. Magyar, G. Knapp (eds.). *Proc. of the Fifteenth International Conference on*

Information Systems Development 2006 (ISD'2006), Budapest, Hungary, Springer, 2006, p. 437–447.
11. LEBEDYS, E. UML in clinical trial design. In O. Vasilecas, A. Caplinskas, J. Eder (eds.). *Databases and Information Systems. Communications. Materials of Doctoral Consortium (Baltic DB&IS 2006)*, Vilnius, Technika, 2006, p. 323–327.
12. VASILECAS, O.; LEBEDYS, E. Business rules for automation of clinical trial data clarification. In *Proc. of the 20th International Conference "Systems for Automation of Engineering and Research" (SAER'2006)*, Varna, Bulgaria, p. 165–170.
13. VASILECAS, O.; LEBEDYS, E. The use of UML models for the development of information systems on the basis of business rules. In *Proc. of the conference "Informacinės technologijos 2006"*, Kaunas, Technologija, 2006, p. 639–644 (in Lithuanian).
14. VASILECAS, O.; LEBEDYS, E. Business rules repository for business rules represented using UML. In B. Rachev, A. Smirkarov (eds.). *Proc. of the International Conference on Computer Systems and Technologies "CompSysTech'05"*, Varna, Bulgaria, 2005, p. II.5-1–II.5-6.
15. VASILECAS, O.; LEBEDYS, E.; LAUCIUS, J. Formal methods for representation of business rules specified using UML. In R. Simutis (eds.). *Proc. of International Conference on Information Technologies for Business 2005*, Kaunas, 2005, p. 41–47.
16. VASILECAS, O.; LEBEDYS, E. Formal business rules specification for automation of engineering process. In *Proc. of the 19th International Conference "Systems for Automation of Engineering and Research" (SAER'2005)*, Varna, Bulgaria, p. 30–34.
17. VASILECAS, O.; LEBEDYS, E. The use of formal business rules specification in the development of information systems. In *Proc. of the conference "Informacinės technologijos 2005"*, Kaunas, Technologija, 2005, p. 636–643 (in Lithuanian).
18. VASILECAS, O.; LEBEDYS, E. Specification and implementation of business rules on the basis of formal languages. In *Proc. of the 8th conference of young Lithuania scientists "Lietuva be mokslo–Lietuva be ateities"*, Vilnius, Technika, 2005, p. 217–222 (in Lithuanian).
19. VASILECAS, O.; LEBEDYS, E. Analysis of the business rules modeling languages. In *Proc. of the conference "Informacinės technologijos 2004"*, Kaunas, Technologija, 2004, p. 487–494 (in Lithuanian).
20. VASILECAS, O.; LEBEDYS, E. Representation of business rules using formal logic. In *Proc. of the 7th conference of young Lithuania scientists*

"Lietuva be mokslo–Lietuva be ateities", Vilnius: Technika, 2004. p. 169–187. (in Lithuanian).

About the author

Evaldas Lebedys was born in Simnas, on 6th April 1979. Evaldas Lebedys started studying in the Faculty of Fundamental Sciences of Vilnius Gediminas Technical University in 1997. In 2001 he defended bachelor's thesis "Order management system database" in Informatics. In 2003 Evaldas Lebedys finished master's studies and defended thesis "Business rules approach in the design of information systems as a part of business system modeling" in Informatics. He started doctoral studies in 2003 in the Information Systems Department.

Evaldas Lebedys was participating in the seminars held in Information Sciences Laboratory during the course of doctoral studies. He presented the results of research in the seminars and publications.

Evaldas Lebedys was PhD student of Vilnius Gediminas Technical University in 2003–2007.

DUOMENŲ KOKYBĖS REIKALAVIMŲ REALIZAVIMAS INFORMACINĖSE SISTEMOSE

Mokslo problema. Dėl įvairių priežasčių duomenų bazėse saugomi duomenys genda, todėl vien riboti įvesti klaidingus duomenis nepakanka. Duomenų kokybė užtikrinama draudžiant įvesti klaidingus duomenis ir tikrinant bei valant duomenų bazėse saugomus duomenis. Klaidingų duomenų įvedimo ribojimai yra aprašomi sistemos modeliuose ir gali būti automatizuotai realizuoti duomenų bazių valdymo sistemose naudojant egzistuojančios sistemų modeliavimo priemones. Egzistuojančios duomenų tikrinimo ir valymo sistemos pritaikytos naudoti specifinėse dalykinėse srityse, o duomenų kokybės reikalavimai šiose sistemose gali būti aprašyti tik programiniu kodu. Duomenų kokybės reikalavimai turi būti specifikuoti pakartotinai – aprašant klaidingų duomenų įvedimo ribojimus ir aprašant duomenų tikrinimo ir valymo taisykles. Dėl pakartotino reikalavimų specifikavimo atsiranda reikalavimų nesuderinamumas, prarandama daug laiko ir lėšų.

Mokslo problemos aktualumas. Duomenų bazėse saugomų duomenų kokybė lemia informacinių sistemų kokybę. Tačiau informacinių sistemų komponentų kokybė daro įtaką duomenų kokybei, nes duomenys įvedami, apdorojami ir vaizduojami naudojant programų sistemas, kurios yra sudėtinės informacinių sistemų dalys. Duomenų kokybė ir su duomenų kokybės užtikrinimu susijusios problemos pastaruoju metu plačiai analizuojamos. Įtaką

duomenų kokybei daro vidiniai ir išoriniai faktoriai. Dažnai duomenų kokybę užtikrinti sudėtinga, nes duomenų kokybė yra daugiamatė. Literatūroje išskiriama daugiau kaip 200 kriterijų, kuriais gali būti vertinama duomenų kokybė ir siūlomi įvairūs metodai, tačiau skirtingose dalykinėse srityse duomenų kokybė apibrėžiama skirtingai, todėl ne visada galima pritaikyti universalius įrankius. Moksliniuose darbuose plačiai analizuojamas duomenų kokybės užtikrinimas aktyvių duomenų bazių valdymo sistemų priemonėmis. Duomenų darnos ribojimai, trigeriai ir saugomos procedūros naudojamos duomenų kokybei užtikrinti, tačiau literatūroje nepakankamai dėmesio skiriama centralizuotam ir automatizuotam duomenų kokybės užtikrinimui. Pastaruoju metu organizacijos susiduria su problemomis, kurios atsiranda dėl metodų, skirtų automatizuotai realizuoti verslo taisykles informacinėse sistemose, ir juos realizuojančių priemonių trūkumo. Ilgus metus verslo sistemos buvo modeliuojamos aprašant verslo sistemų struktūrą, elgseną, funkcinius vienetus ir jų tarpusavio ryšius bei kitus verslo sistemos aspektus, tačiau ne visada buvo skiriamas pakankamas dėmesys taisyklėms, kuriomis aprašomi reikalavimai duomenų kokybei. Šios taisyklės yra kiekvienos sistemos sudėtinė dalis ir organizacijų veiklos kontekste vadinamos verslo taisyklėmis. Verslo taisyklės susijusios su visais verslo sistemų aspektais: sistemos struktūra, sistemos vidine ir iš išorės stebima elgsena, sistemos sąveika su aplinka, sistemos aktorių pareigomis ir atsakomybės ribomis.

Sudėtingose sistemose egzistuoja daug verslo taisyklių, todėl šių taisyklių valdymas, kontrolė ir tinkamas vykdymas yra sudėtingas uždavinys. Verslo taisyklių, tame tarp ir taisyklių išreiškiančių duomenų kokybės reikalavimus, valdymą ir vykdymą siekiama maksimaliai automatizuoti. Tokiu būdu didinamas sistemų patikimumas ir padedama užtikrinti kokybės kontrolę. Remiantis atlikta analize ir eksperimentų rezultatais, galima teigti, kad suformulavus metodą automatizuotai perkelti verslo taisykles, išreiškiančias duomenų kokybės reikalavimus, iš verslo sistemos modelių, pavaizduotų naudojant tam tikrą modeliavimo kalbą, į verslo taisyklių saugyklą būtų galima automatizuoti duomenų kokybės užtikrinimą realizuojant taisykles duomenų kokybės tikrinimo užklausomis.

Darbo tikslas ir uždaviniai. Pagrindinis darbo tikslas yra patobulinti duomenų kokybės užtikrinimo procesą ir išvystyti metodą, skirtą automatizuotai realizuoti duomenų kokybės reikalavimus išreiškiančias verslo taisykles, programų sistemose.

Siekiant įgyvendinti darbe iškeltą tikslą, reikia išspręsti šiuos uždavinius:

1. Išanalizuoti duomenų kokybės reikalavimus išreiškiančių verslo taisyklių modeliavimo kalbas ir realizavimo programų sistemose metodus.

2. Išanalizuoti literatūroje siūlomus duomenų kokybės užtikrinimo metodus.
3. Pasiūlyti metodą, skirtą automatizuotai realizuoti verslo taisykles programų sistemose siekiant užtikrinti duomenų kokybę.
4. Suprojektuoti ir sukurti programų sistemą, įgyvendinančią siūlomą verslo taisyklių realizavimo metodą.
5. Atlikti eksperimentą siekiant įvertinti, ar pasiūlytas metodas tinkamas automatizuotai realizuoti verslo taisykles, išreiškiančias duomenų kokybės reikalavimus, programų sistemose ir ar sukurta programų sistema įgyvendina pasiūlytą metodą bei yra pritaikoma praktiškai.

Mokslinis naujumas. Šios disertacijos mokslinis naujumas pagrįstas verslo taisyklių, išreiškiančių duomenų kokybės reikalavimus ir atvaizduotų UML modeliuose, panaudojimu automatizuojant verslo taisyklių realizavimą programų sistemose. Darbe siūlomas verslo taisyklių realizavimo duomenų kokybės tikrinimo užklausomis metodas, kuris turėtų supaprastinti duomenų kokybės užtikrinimą bei padaryti šį procesą prieinamą specialių žinių neturintiems veiklos dalyviams.

Paskutiniu metu vykdoma daug tyrimų informacinių sistemų modeliavimo verslo taisyklių požiūriu. Tai svarbu, nes taisyklės, tarp jų ir duomenų kokybės reikalavimus išreiškiančios taisyklės, veikia visus verslo sistemų aspektus. Nors verslo taisyklės paskutiniu metu tapo plačiai literatūroje nagrinėjama tema, nėra priemonių aprašyti ir realizuoti konkrečios dalykinės srities visą verslo taisyklių sistemą. Literatūroje vyrauja nuomonė, kad nėra vienos modeliavimo kalbos, tinkančios sukurti išsamų koncepcinį dalykinės srities modelį ir aprašyti visų klasių verslo taisykles. Manoma, kad, norint automatizuoti programų sistemų kūrimą, verslo taisyklės turi būti užrašytos naudojant formalią kalbą, kurią galėtų suprasti intelektualios sistemos.

Darbe siūlomas metodas, skirtas automatizuotai išrinkti skirtinguose verslo sistemos modeliuose aprašytas verslo taisykles išreiškiančias duomenų kokybės reikalavimus. Siūloma verslo taisyklių saugykloje esančias taisykles vaizduoti natūralia kalba, skirta verslo auditorijai. Darbe taip pat siūloma generuoti duomenų kokybės tikrinimo užklausas naudojant saugykloje laikomas taisykles.

Tyrimų metodika. Darbe aprašomam tyrimui vykdyti buvo naudojami šie tyrimo metodai:

1. Lyginamosios analizės ir bibliotekinio tyrimo metodai buvo naudojami vykdant egzistuojančių verslo taisyklių modeliavimo kalbų, verslo taisyklių realizavimo metodų bei duomenų kokybės užtikrinimo metodų analizę.
2. Analizės rezultatams apibendrinti ir siūlymams formuluoti buvo naudojami sąvokų apibendrinimo bei loginės indukcijos metodai.

3. Siūlomas taisyklių išrinkimo iš UML modelių ir realizavimo programų sistemose metodas buvo įgyvendintas naudojant konstravimo metodą.
4. Siekiant patikrinti siūlomo metodo veikimą buvo atlikti bandymai naudojant eksperimentinės analizės metodą.

Praktinė vertė. Tyrimo rezultatai parodė, kad disertacijoje pasiūlytas metodas ir sukurtas programų sistemos prototipas gali būti naudojami automatizuoti informacinių sistemų kūrimą verslo taisyklių pagrindu. Turint konkrečios sistemos UML modelį ir naudojant sukurtą programų sistemos prototipą, taisykles, pavaizduotas modelyje ir aprašančias duomenų kokybės reikalavimus, galima perkelti į taisyklių saugyklą. Čia perkeltos taisyklės tampa prieinamos įvairiapusiškam apdorojimui: gali būti vykdoma taisyklių sistemos sudėtingumo analizė, atliekamas taisyklių tarpusavio darnos tikrinimas, taisyklės gali būti pakartotinai panaudotos. Sukurta programų sistema gali būti naudojama generuoti duomenų kokybės tikrinimo užklausas iš taisyklių, saugomų taisyklių saugykloje. Automatizavus duomenų kokybės reikalavimų realizavimą programų sistemose, duomenų kokybės užtikrinimo valdymas tampa prieinama ir sistemos vartotojams, kadangi duomenų kokybės reikalavimai tvarkomi centralizuotai ir gali būti išreikšti natūralia kalba. Buvo atlikti tyrimai siekiant pritaikyti siūlomą metodą e-mokymų dalykinėje srityje. Tyrimų rezultatai aprobuoti ir publikuoti tarptautinės recenzuojamos konferencijos medžiagoje.

Ginamieji teiginiai

1. Sistemos modeliuose vaizduojamos verslo taisyklės išreiškiančios duomenų kokybės reikalavimus.
2. Duomenų kokybės reikalavimus aprašančios verslo taisyklės gali būti automatizuotai perkeltos iš sistemos modelių į taisyklių saugyklą.
3. Sistemos modeliuose pavaizduotus duomenų kokybės reikalavimus galima automatizuotai realizuoti programų sistemose.

Darbo apimtis. Darbą sudaro bendra darbo charakteristika, 4 skyriai, išvados, literatūros sąrašas, publikacijų sąrašas ir priedai. Bendra disertacijos apimtis – 110 puslapių, 41 iliustracija, 12 lentelių ir 3 priedai.

Antrame disertacijos skyriuje aprašytas informacinių sistemų modeliavimas verslo taisyklių pagrindu. Šiame skyriuje aprašyti aspektai, kuriais gali būti modeliuojamos verslo sistemos, informacinės sistemos be programų sistemose. Nagrinėtos sistemų modeliavimo kalbos, bei išanalizuotas skirtingų modeliavimo kalbų tinkamumas modeliuoti sistemas skirtingais požiūriais. Čia taip pat aprašytos verslo taisyklės vaizduojamos skirtingus

sistemų aspektus aprašančiuose modeliuose. Pagrindinis dėmesys darbe skiriamas UML kalbai ir skirtingose UML diagramose pavaizduotoms verslo taisyklėms. Šiame skyriuje išskiriami skirtingose UML diagramose pavaizduoti verslo taisyklių tipai.

Trečiame skyriuje aprašytas UML modeliuose pavaizduotų verslo taisyklių panaudojimas automatizuojant duomenų kokybės tikrinimą. Aptartos pagrindinės klaidų atsiradimo duomenyse priežastys, išanalizuoti skirtingi duomenų apdorojimo procesai bei duomenų kokybės užtikrinimo metodai. Trečiame skyriuje pateikiama schema, kaip verslo taisyklės gali būti išrinktos iš UML modelių ir perkeltos į taisyklių saugyklą. Šiame skyriuje nagrinėtas siūlomas duomenų kokybės užtikrinimo metodas ir siūlomo metodo pranašumas egzistuojančių metodų atžvilgiu

Ketvirtame skyriuje suformuluoti reikalavimai kuriamam programų sistemos prototipui. Aprašyti funkciniai reikalavimai bei nefunkciniai reikalavimai sistemos savybėms. Suformuluotų reikalavimų pagrindu buvo kuriamas programų sistemos prototipas bei atliekamas sukurto prototipo testavimas. Šiame skyriuje aptartas programų sistemos prototipo modeliavimas bei kūrimas, pateiktos sistemos struktūrinius komponentus bei elgseną vaizduojančios UML diagramos, aprašyti apibendrinti sistemos testavimo rezultatai.

Penktame skyriuje nagrinėtas eksperimentui pasirinktos verslo sistemos veiklos modelis bei jame pavaizduotos verslo taisyklės. Šiame skyriuje pavaizduotas sukurto programų sistemos prototipo veikimas perkeliant UML modelyje pavaizduotas verslo taisykles į taisyklių saugyklą. Čia taip pat pademonstruota, kaip sukurtoji programų sistema generuoja duomenų korektiškumo tikrinimo užklausas bei vykdo duomenų tikrinimą nurodytoje duomenų bazėje.

Darbo rezultatų aprobavimas

Pagrindiniai tyrimo rezultatai paskelbti (2003–2007):

- 3 recenzuojamuose užsienio mokslo žurnaluose;
- 5 recenzuojamuose Lietuvos mokslo žurnaluose;
- 2 tarptautinėse duomenų bazėse referuotuose žurnaluose;
- 2 ISI konferencijų darbų sąraše (proceedings list);
- 3 tarptautinių recenzuojamų konferencijų leidiniuose;
- 5 kituose mokslo leidiniuose.

Bendrosios išvados

1. Sistemų modeliavimo kalbų ir verslo taisyklių vaizdavimo kalbų analizė parodė, kad skirtingus sistemų aspektus aprašančios verslo taisyklės gali

būti vaizduojamos UML kalba. Nustatyta, kad verslo taisyklės gali būti išrinktos iš sistemų modelių, pavaizduotų naudojant UML, ir perkeltos į taisyklių saugyklą.
2. Tyrimo metu nustatyta, kad verslo taisyklės, vaizduojamos UML modeliuose, yra įgyvendinamos duomenų bazių valdymo sistemose arba programų sistemose programiniu kodu. Analizės rezultatai parodė, kad nėra priemonių skirtų centralizuotai valdyti sistemos modeliuose pavaizduotas verslo taisykles.
3. Išanalizavus duomenų kokybės užtikrinimo metodus, nustatyta, kad egzistuojantys metodai yra pagrįsti ribojimų, draudžiančių išsaugoti klaidingus duomenis, kūrimu arba rankiniu duomenų kokybės reikalavimų realizavimu programų sistemose skirtose duomenims valyti.
4. Sukūrus siūlomą metodą įgyvendinančią programų sistemą ir atlikus eksperimentą, nustatyta, kad turint sistemos modelį ir naudojant sukurtą programų sistemą, taisykles galima išrinkti ir perkelti į taisyklių saugyklą.
5. Tyrimas parodė, kad saugykloje saugomas taisykles galima panaudoti generuojant duomenų tikrinimo užklausas ir patikrinti, ar nurodytoje duomenų bazėje saugomi duomenys nepažeidžia taisyklių.
6. Eksperimentu įrodyta, kad sukurta programų sistema gali generuoti duomenų tikrinimo užklausas, išreikšti duomenų kokybės reikalavimus natūralia kalba bei vykdyti duomenų kokybės tikrinimą atsižvelgiant į verslo taisykles.
7. Nustatyta, jog automatizavus duomenų kokybės reikalavimų įgyvendinimą programų sistemose, duomenų kokybės kontrolė tampa prieinama ir sistemos vartotojams, kadangi duomenų kokybės reikalavimai tvarkomi centralizuotai ir gali būti išreikšti natūralia kalba.
8. Eksperimentu pademonstruota, kad naudojant pasiūlytą metodą duomenų tikrinimo užklausos gali būti sukurtos ~1,5 greičiau negu rankiniu būdu.

Trumpos žinios apie autorių

Evaldas Lebedys gimė 1979 m. balandžio 6 d. Simne, Alytaus rajone. Studijas Vilniaus Gedimino technikos universiteto Fundamentinių mokslų fakultete jis pradėjo 1997 m. 2001 m. apgynė informatikos bakalauro baigiamąjį darbą tema „Užsakymų valdymo sistemos duomenų bazė“. Evaldas Lebedys įstojo į magistratūros studijas 2001 m. ir 2003 m. apgynė informatikos mokslo magistro baigiamąjį darbą tema „Verslo modeliavimas verslo taisyklių pagrindu IS kūrimo procese“. 2003 m. jis pradėjo doktorantūros studijas Informacinių sistemų katedroje.

Evaldas Lebedys

IMPLEMENTATION OF DATA QUALITY REQUIREMENTS IN INFORMATION SYSTEMS

Summary of Doctoral Dissertation
Technological Sciences, Informatics Engineering (07T)

Evaldas Lebedys

DUOMENŲ KOKYBĖS REIKALAVIMŲ REALIZAVIMAS INFORMACINĖSE SISTEMOSE

Daktaro disertacijos santrauka
Technologijos mokslai, informatikos inžinerija (07T)

2007 10 26. 1,25 sp. l. Tiražas 100 egz.
Vilniaus Gedimino technikos universiteto
leidykla „Technika“, Saulėtekio al. 11, 10223 Vilnius
http://leidykla.vgtu.lt
Spausdino UAB „Baltijos kopija“,
Kareivių g. 13B, 09109 Vilnius, www.kopija.lt