International Journal of Software Engineering and Computer Systems (IJSECS)
ISSN: 2289-8522, Volume 4 Issue 1, pp. 95-105, February 2018

https://doi.org/10.15282/ijsecs.4.1.2018.8.0041

INFORMATION SYSTEMS REENGINEERING APPROACH BASED ON THE MODEL OF INFORMATION SYSTEMS DOMAINS

Maria Glava[1], Valery Malakhov[2]

[1]Department of Economic Cybernetics and Information Technologies,
Odessa National Polytechnic University, Odessa, Ukraine.
[2]Department of Mathematical Support of Computer Systems,
Odessa I. I. Mechnikov National University, Odessa, Ukraine.
Email: glavamaria@mail.ru, valery.e.malakhov@gmail.com

ABSTRACT

The paper considers current problems of integration of Information Systems (IS), limitations of current methods of IS Reengineering and limitations of existing approaches for Data Integration in Relational Databases. The performed analysis shows the drawbacks of existing integration approaches. To increase effectiveness of IS integration at the conceptual level, the paper proposes a new technique for the integration of relational databases on the base of the model of an IS Domain (ISD). The approach allows to automate the synthesis of conceptual schemas for the different layers of ISDs and results in the increasing effectiveness (saving labour and time efforts) of the problem solution. The technique is based on the mapping values of the properties of instances of composing ISD objects.

Keywords: Information system; Information systems reengineering; Information systems domain; Model of information systems domain.

INTRODUCTION

Information systems technologies influence any business today. The success in a business depends on the speed of response to IT changes and effectiveness of an IS use. Any alteration of a government policy or expanding a business scope requires rapid and weighed solutions for the business processes optimization.

Redesign of business processes accordingly to the definition of M. Hammer and J. Champi (1993) is called a business process reengineering (BPR). BPR is a fundamental reconsideration and a radical change of business processes to achieve a rapid improvement of the core indicators of enterprise activities, such as durations, quality, expenses and services.

Business problems, which should be solved by reengineering, are typically characterized by high degree of complexity and responsibility. They can be solved by change of the sequence of actions (process steps); modification of distribution of tasks between departments (employees); adjustment of the material and information flows, circulating in the company etc. Because any enterprise or organization use an IS now, the question how the change in the business processes will affect the work of IS becomes crucial.

Information technologies provide services to manage interaction between business organizations and clients. They also allow to integrate the efforts of the different departments, involved in a business process, and to improve results of their

work. BPR is not possible without IS restructuring, because any modern organization supports IS infrastructure.

Integration of the data will have significant value for organization development and improve its competitiveness on the market, because a corporate IS will result in the innovative structure of the organization.

In this regard, the problem how to change effectively an existing IS infrastructure for the optimization of business processes arose. The solution requires study both of business processes reengineering and IS Reengineering (ISR), because any IS operates in an enterprise, which needs to be managed in the process of reengineering.

The relevance of development of new ISR methods follows from the fact that organizations usually spend from 20 to 40% of their IT budget for data migration (change of location of data), conversion (change of a form or a data structure) or cleaning (deleting of repeated data entries) (Aiken, Allen, Parker & Mattia, 2007). Practice of reengineering shows that more than two thirds of total time and money are spent on attempts to combine IS modules written by the different people, in the different time in different languages and technologies, an under different platforms.

PROBLEMS OF INFORMATION SYSTEMS REENGINEERING

Reengineering of an information system is its improvement through redesign, revision of its operation with the purpose to increase the key productivity indicators.

According to the standard "ISO/IEC 2382:2015 Information technology", an IS is a system, intended for storing information and processing the appropriate organizational resources (human, technic, financial etc.), which provides and disseminates information. A modern IS contains both stored in the database (DB) information and the technology for information processing.

Currently, ISs, based on the relational DB model, are the most widespread in a business environment. Relational databases are suitable for solution of a wide class of problems, as they provide a simple interface, show good performance, and have the most elaborated mathematical apparatus for database manipulation.

To start a BPR, an organization needs take into account following constraints:

– an enterprise has different databases in various departments;

– databases, which form the basis of an IS, as a rule are constructed on the outdated platforms;

– business process reengineering will lead to the considerable changes of information flows;

– during further development of an organization or their possible merge, the need for integration of miscellaneous IS arose etc.

IS integration first of all meets difficulties of connection and analyzing information from miscellaneous sources, frequently isolated from each other. The solution here is consolidating existing information into unified information space. Integration of ISs is needed not at the layer of an external schema, i.e. a user view, but at the data layer, which is the model of information system domain.

Even having similar scope organizations can have problems with ISs integration due to their different design. In the case of integration of IS of the different domains, stored in the different DB data have completely different view and are not consistent.

Today there are different approaches, methods and technologies are directly or indirectly correlated with IS reengineering activities. However, they are not integrated

at the level of methodologies and modern development processes. As result, there is big amount of ISR approaches, which focuses on the IS development "from scratch", but practically absent solutions for the IS reengineering problem (Akhtyrchenko & Sorokvasha, 2003).

Let's also note here big among the researchers, involved in the data integration problems - M. Kogalovsky (2010), L. Globa, M. Ternovoy and E. Shtogrina (2011), A. Berko (2010), L. Chernyak (2009), P. Ziegler and K. R. Dittrich, (2004, 2007), M. Lenzerini (2002) and others.

Lets consider existing approaches to data integration - consolidation, federalization, data distribution, hybrid and service approaches (Beloshitsky, 2013).

By using *consolidation*, data are retrieved from several ISs and are placed in a single data warehouse. The process of filling the warehouse is unidirectional and is divided into three phases - extraction, transformation and loading. There are several modifications of this approach, which can be classified by the categories of structures transfer and integration. The transfer also includes integration of data structures. The integration process consists of linking the data models, metadata and data in a new IS. To minimize costs, developers of organizations use structures transfer. It allows them to reduce the number of servers and price of solution consequently.

In the case of *federalization*, a single virtual information space is formed and an integration takes place in a real time. If the request contains an access to multiple data sources, it is decomposed into several separate running subqueries. To obtain the final answer, results of subqueries are composed together. Disadvantage of federalization method is a low productivity, which prevents using servers for many tasks, and additional costs to access multiple data sources at runtime.

The *data distribution* method is an information transfer from one IS to another at certain events. A distinctive feature of this method is quick data exchange. Data can be transferred both synchronously and asynchronously. The disadvantage is an inability to perform common analytic queries, as it may be necessary to use temporary storage-analyzer, which is not provided in this approach.

The *hybrid* approach is simultaneous application of several methods for data integration. For example, data consolidation ("customer data") and federalization ("orders").

Data services, combined into a single layer, abstract business logic from data delivery applications of their different sources and data conversion. The level of data services allows encapsulating all considered data integration technologies into the components, available for reuse by various applications in the different scenarios of data integration. Note, that the concept of data management service approach is still under development.

Analyses of existing data integration approaches outcomes in the fact that in order to minimize the data duplication (in terms of storage) and time (to receive information by the user), it is necessary to identify common information elements of all databases. This why the model of the common entities in the different database representation to be developed, e.g. in the form of the projections on the corresponding ISD (Malakhov, 2006, 2007). This also require mathematical integration of the ISD models, describing each of the combined databases.

Thus, existing data integration approaches needs careful revision. There is a need for the development of the IS integration technology, based on the mathematically elaborated model of IS domain.

ANALYSIS OF THE EXISTING APPROACHES AND PROBLEMS OF DATA INTEGRATION

During IS reengineering, specialists need solve following most common problems of data integration (Glava & Vasylieva, 2015):

1) Use of the data profiling tool for the analysis and assessment of the data quality, taken from different sources, or use of data from the environment for the development of integration logic.

2) Establishing the criteria for admissibility of individual application systems. All participating in the integration project business parties have to define common usage scenarios.

3) Checking quality and accuracy at each stage of the integration project.

4) Development of standard interfaces for loading and exporting ISs, which provide data on a regular basis. The integration solution must be universal for various systems.

There are several alternative layers of data integration: integration at the physical, logical and semantic level. Before their analysis, let us briefly review the criteria to be taken into account for the data integration: data updating, performance, cost, encapsulation, data synthesis, data access, management, "secondary effects", data integrity and scalability.

The criterion of *data updating* is based on how new data from the source database are transferred into the target database. Depending on the integration technique, there can be time delays in transferring data from one database (a source, available to users), and their use in a target database. The criterion of *efficiency* is based on the speed of the integration process. The *cost* criterion includes not only the effort for integrating products, but also costs of implementation. The *encapsulation* criterion shows how well-integrated solution hides the physical location of data from the users and other applications. The criterion of the *data synthesis* considers how users work with data from several databases. The criterion of *data access* focuses on what type of access the integration solution implements for users and applications. The data access includes creation, reading, updating and deleting data and defines whether the access is uni or bidirectional. The *management* criterion defines the effect, which the integration solution will have on the administration of a database. "*Secondary effects*" – this criterion is used for classifying a technical effect of an integration to other applications and computer environment parts. The criterion of the data integrity checks how well the integration technique manages transactions, related to several databases. The *scalability* criterion assesses how the integration method performs if the number of integrated databases increases (Glava & Vasylieva, 2015).

Let’s consider the layers of data integration in more details.

Integration of data at the physical layer is a converting data into the uniform format, which is the simplest method. This method has many advantages: cost is the least expensive here; special integration tools for data export and import are not required, because existing database tools are sufficient; only the minimum changes in a database are needed. The criteria that satisfy this technique are performance, encapsulation, data synthesis, administration, and "secondary effects". This method also has disadvantages: the data updating, data access, data integrity and scalability.

Syntactically or logically, data integration is based on the resemblance of merged data; enables access to the data in the terms of a uniform global schema and considers structural properties of data from various sources.

Structural distinctions of data schemes can cause following data integration problems:

– heterogeneity, when the different models of data from the various sources are used;

– the problem of names, when the schemes use different terminology, leading to naming homonymy and synonymy;

– semantic problems, when different levels of abstraction for the modeling similar entities of the real world are used;

– structural problems, when the same entities are presented in the different sources by dissimilar data structures (Kogalovsky, 2010).

The semantic layer of integration is based on the similarity of the merging data. Semantic integration is knowledge-based approach, taking into account the nature of the data. For it, data have to be stored together with metadata, which is more difficult in the plan of implementation, but significantly increases the effectiveness of work.

There are following problems at the level of semantic integration (Wache et al., 2001):

– contradiction in the definitions of the concepts;

– ambiguity or discrepancy of the names;

– applying of the inconsistent metrics;

– contradiction in the definition of the data relationships;

– contradiction of the constraints and axioms;

– ambiguity of the interpretation results.

During the database integration, one of the most important tasks is to keep the data and avoid duplication. As we noted above, a database from the same information domain can be constructed in completely different way. There are two types of data inconsistency: structural and lexical. The structural inconsistency occurs when the fields are structured differently in various databases. For example, in a database, a customer address may be recorded in a field named "addr", while in another database, the same information can be stored in several fields, such as "street", "town", "state" and "zip code". Lexical inconsistency occurs, when the tuples have the same field structure in various databases, but use different data representation to describe the same objects of the real world (Elmagarmid, Ipeirotis & Verykios, 2007).

In the context of our research, the most promising level of the integration is the semantic layer, which allows to integrate not only structurally identical data, but also having the same meaning. As result it will allow to build the most adequate mathematical model of the information domain (Malakhov & Vostrov, 2010). Let us note here, that universal approach to IS integration at semantic layer of the data is under development now.

Currently, the following approaches solve the problem of ISs integration (Glava & Vasylieva, 2015):

1) The integration using Batch-service, which represents the physical layer of integration.

2) Integration at the application-layer: applications are able to communicate directly with several databases; they can be modified to obtain information directly from other systems. This technique satisfies the criteria for the data updating, administration and "secondary effects". Disadvantages of this technique include encapsulation, data synthesis, data integrity and scalability criteria.

3) Middleware integration also links the databases at the application layer but using some intermediate software. Middleware provides a mechanism for

communication of applications with database, which is not supported by the application itself. Whereas a layer of immediate communication of application with data is removed, appearing some performance lost. While middleware software cannot eliminate the problem of encapsulation and data synthesis, as integration at the application layer, it can reduce cost of software development and maintenance.

4) Integration at the database layer: the method has many advantages, such as minimal cost, encapsulation, "secondary effects", data synthesis, data integrity and scalability. Two drawbacks of this method – the cost and administration. Middleware, necessary for the integration at the database layer, is much more expensive and usually requires a significant time for database administration.

5) Data migration is a complex process of transition from one data storage to another. This method solves an integration problem by transferring different computer systems, which have to be integrated, into a joint database.

Obviously, that none of these approaches is deprived of drawbacks, the most important of which is the lack of mathematical models of ISD.

Some researchers and developers have proposed a number of the methods, focused on the "technical" integration of the existing databases.

A method for the data schemes integrating (Komar & Pogodaev, 2008) is based on the semantic describing of attributes as set of string templates. Based on these attributes the semantic similarity is assessed, and the similarity measure of the database relations is calculated. This method assumes that semantically identical attributes have the same occurrences of the attribute values, satisfying the set of templates. However, some string templates can be repeated in the semantically different attributes, e.g. a city name and a family name. In addition, this method does not describe a procedure for the comparison of attributes, whose types is not a string.

The method of identifying previously unknown functional dependencies (Radchenko & Tanyansky, 2012) is based on the analysis of the set of data of relational database. First step is to obtain a set of functional dependencies for each relation. On the second step, similar operation is performed for the universal relation of considered relational database. At this step, it is possible to identify functional dependencies between attributes of the different relations – the relationships between data, which have been established in the operation of a relational database. There is method for determining the information novelty, which is the checking of membership of the functional dependencies of a universal relation in the closure of the union of sets of the functional dependencies of the individual relations. However, this method does not take into account the data semantics; therefore, probability of generating random functional dependencies is high. In addition, the problem of matching universal relations of the merged databases is not solved.

Yesin (2012) proposes an approach to integrate databases by constructing a universal (standard) data model, which is based on the semantic model of data "event-object", set theory and logical calculus. In the "event-object" model, all objects, processes, phenomena of any domain are described by meta-ontology (meta-ontology serves as a model paradigm, focused on the description of the elements of any subject domain).

The common solution for the integration problem is based on the IS metadata description, and mapping entities and relationships of an IS in the terms of a common ISD ontology (Vagin & Mikhailov, 2008).

Conceptual models of information systems are created in accordance with XML and RDF schema standards. They used to create a common metamodel that combines the representation of entities of two or more data warehouses.

An ontology represents data dictionary and includes both the terminology and a model of a behavior. Since each conceptual model of an ISD is a subset of the ontology, the problem of integration of the ISs is reduced to the integration of the IS metamodels, i.e. development of the mapping between metamodels in ontology terms.

Having IS metamodel all integration problems such as data retrieval to be solved. Analysis of the ontology-driven research shows that proposed methods require further development to use in the enterprises, now the problem was solved only for partial cases.

For the most of database integration methods at the semantic layer to confirm the correctness of the result, an attraction of experts is necessary. It is impossible to apply current methods without involving experts to analyze the ontologies, created by various working groups. This is a major drawback of the proposed methods.

PROPOSED APPROACH FOR THE INTEGRATION OF RELATIONAL DATABASES

After analyses of existing solutions for the ISs integration, it can be concluded that considered methods do not complete solve the problem of data integration. The problem of databases integration on the conceptual layer remains actual, this why this paper proposes an approach for the integration of relational databases as domain models. The approach will allow us to automate the creation of conceptual schemas for multiple layers of ISD, which will give an opportunity to save work and time costs.

Each information system domain in the classical relational data model, can be described by the pair

$$SD = <E, R> \quad (1)$$

where E – set of objects (entities), R – the set of relationships between objects.

The procedure for combining domain models depends on the similarity of information systems domains. For the ISDs, which describe different business domains, an integration is possible only for the typical objects (defining objects) (Glava & Malakhov, 2016) - representatives of many ISDs (e.g., ‘contractors’, ‘employees’ etc.). For ISDs, reflecting the same kind of business, we need find matching entities and integrate them. For domains, that describe the same area of business should be matched both the objects that nondefining the domain and objects which can belong to both compared domains, so-called "borderline" objects (Fig. 1).

For the definition of the of information domains similarity, it is offered to use the ISD model proposed by Malakhov (2010)

$$ISD = <E, R, P> \quad (2)$$

where P – the problem to be solved in a specific information domain.

Glava and Malakhov (2016) presented the technology to find the projections of the same universal entities on ISD, where objects are compared on the base of the values of properties of the objects’ instances. Algorithms for comparing depends on the data type of specific properties.

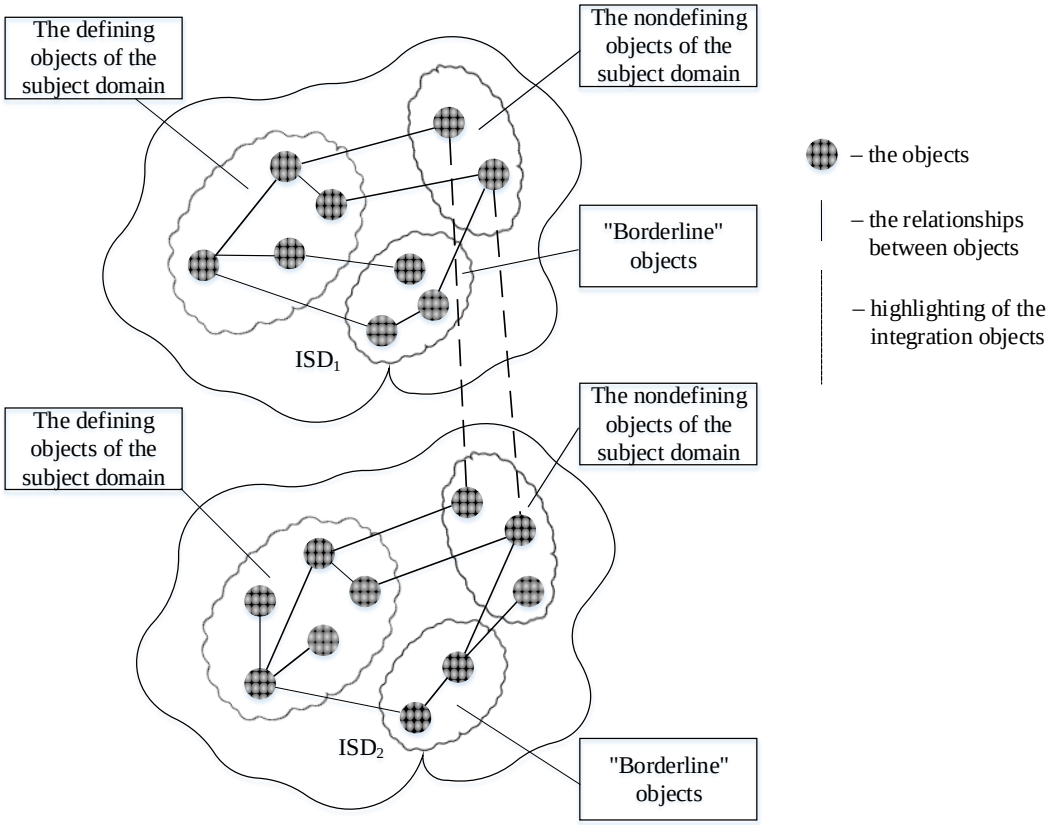

Figure 1. Domain object types

Accordingly to proposed technology, objects of the potentially similar ISD must be prepared for the matching:

– allocate the essential properties, based on the information amount of each property and expert assessment;

– rank the objects of each compared ISD by significance, based on the number and importance of a particular object relationships with others in the same ISD and the number of significant properties, measured by a certain scale (ordinal, nominal, numeric);

– sort corteges according to the values of ordinal and nominal properties, keeping obtained previously properties rank.

Next, the match of the potentially similar objects by their types is performed:

– analysis of the subgroup of ordered properties will allow to bring together the tuples, matching in potentially similar objects and aligning their number. For it, empty tuples are added to some objects, depending on the result of the values matching. Then comparison of the values of tuples of the ordered properties is performed by some statistical method (e.g. correlation);

– for comparing the nominal types of properties it is proposed to build an ontology model, which characterizes any nominal properties. To process each nominal property of every object of potentially similar ISDs, to fill in the individuals (instances) of the concepts (classes) of the ontology. Next, the ontology concepts of two compared domains should be pairwise matched (Glava, 2016);

– for comparing numerical properties we propose to apply digital filters, e.g., wavelet or discrete cosine transform. The result to be compared by statistical methods (e.g. correlation) (Fig. 2).

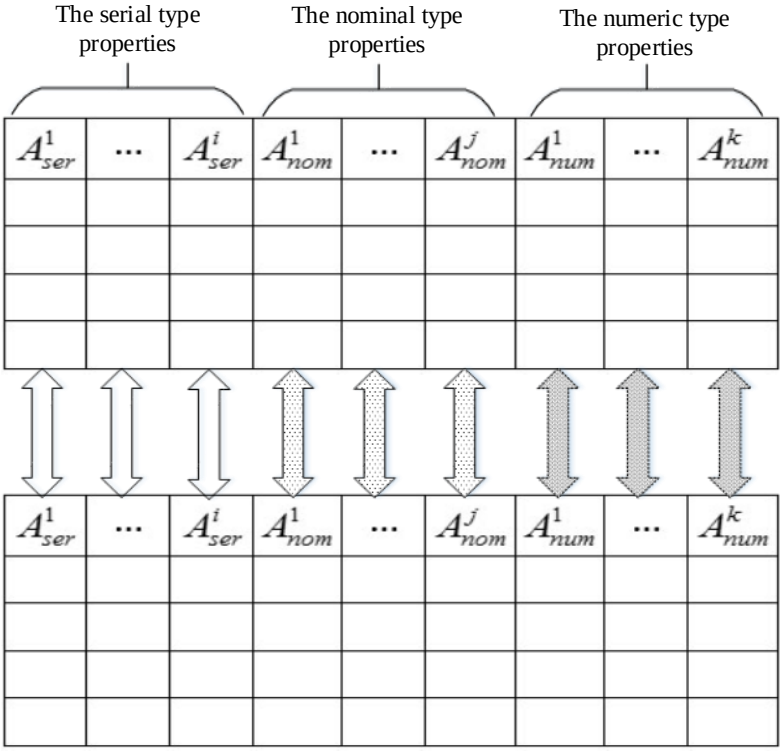

Figure 2. Matching objects by their types

Then the ratio the of objects' similar properties to their total number can be determined. The resulting ratio is compared to a threshold, given by an expert and concludes, whether the compared objects are the similar.

Finally, a common model of information systems domains is proposed by including from the most similar objects the most similar properties and expanding them by different properties, ordered by importance.

CONCLUSION

The paper analyses the problems of the information systems reengineering and considers existing methods for solving this problem. The technology for the integration of relational databases, based on a matching of the values of their objects is proposed.

A proposed approach does not completely solve the problem of data integration; more research are needed to integrate entities into unique database, as well as on the integration of general problems to be solved on the ISDs, which are subject to the integration.

Proposed approach allows to automate the development of the ISD model (conceptual schema), reducing the work efforts and time required for the database integration. It also allows reducing the number of problems to be solved by experts, thus decreasing subjectivity of decision.

REFERENCES

Aiken, P., Allen, M. D., Parker, B. & Mattia, A. (2007). Measuring Data Management Practice Maturity: A Community's Self-Assessment. *Computer*, *04* (40), 42-50. doi: 10.1109/MC.2007.139.

Akhtyrchenko, K. V. & Sorokvasha, T. P. (2003). Methods and technologies of IS reengineering. *Proceedings of the Institute for System Programming of the RAS, 4*, 141-162.

Beloshitsky, D. A. (2013). Data integration in the information systems. *The youth bulletin of science and technology, 8*.

Berko, A. Ju. (2010). The structural and semantic data integration based on the factual relational model. *Bulletin of the Lviv Polytechnic National University, 663*, 60-69.

Chernyak, L. (2009). Data Integration: Syntax and Semantics. *Open Systems, 10*.

Elmagarmid, A. K., Ipeirotis, P. G. & Verykios, V. S. (2007). Duplicate Record Detection: A Survey. *IEEE transactions on knowledge and data engineering, 19* (1), 1-16. doi: 10.1109/TKDE.2007.250581.

Glava, M. (2016). Comparison of the nominal type properties of objects of different subject subdomains in relational databases. *Informatics and Mathematical Methods in Simulation, Vol. 6 (2016), No. 3,* 302-309.

Glava, M. & Malakhov, E. (2016). Searching Similar Entities in Models of Various Subject Domains Based on the Analysis of Their Tuples. *2016 International Conference on Electronics and Information Technology (EIT),* 97-100. doi: 10.1109/ICEAIT.2016.7501001.

Glava, M. G. & Vasylieva, T. P. (2015). Major problems and methods of databases integration. *First Independent Scientific Journal, 1*, 28-32.

Globa, L. S., Ternovoy, M. Ju. & Shtogrina, E. S. (2011). Using the ontologies to integrate databases and knowledge bases. *Intelligent Analysis of Information IAI-2011*, 34-38.

Hammer, M. & Champy, J. (1993). *Reengineering the Corporation: A Manifesto for Business Revolution*. New York: HarperCollins.

Kogalovsky, M. R. (2010). Methods of data integration in the information systems. *Electronic Socionet depositor*.

Komar, F. V. & Pogodaev, A. K. (2008). The method of the data schemes integration based on attributes semantic description. *Software & Systems, 1*, 53-55.

Lenzerini, M. (2002). Data Integration: A Theoretical Perspective. *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, 233-246. doi: 10.1145/543613.543644.

Malakhov, E. V. (2006). Representation of the objects in a set of subject domains. *Eastern-european journal of enterprise technologies, 2/2* (20), 20-23.

Malakhov, E. V. (2007). Manipulation of the subject subdomains. *Eastern-european journal of enterprise technologies, 5/3* (28), 7-11.

Malakhov, E. V. (2010). The information technology for complex structures subject domains simulating in the organizational management systems (theory and implementation). *Synopsis of doctorate thesis (Sc.D.), Odessa, 11*.

Malakhov, E. V. & Vostrov, G. N. (2010). Criteria of adequacy mathematical models the subject domains. *Eastern-european journal of enterprise technologies, 6/2* (48), 18-20.

Nor Faiz Muhammad Noor, Omar Zakaria & Puteri N. E. Nohuddin. (2016) A proposed framework to control rumour propagation on twitter for critical national information infrastructure organisations. *International Journal of Software Engineering and Computer Sciences*, Volume 2, pp. 1-9.

Radchenko, V. O. & Tanyansky, S. S. (2012). Discovery of hidden data relationships in tasks of information systems reengineering. *Information processing systems, 3* (101), 203-205.

Vagin, V. N. & Mikhailov, I. S. (2008). Developing the methods for information systems integration based on metamodeling and domain ontology. *Software & Systems*, *1*, 22-26.

Wache, H., Vogele, T., Visser, U., Stuckenschmidt, H., Schuster, G., Neumann, H. & Hubner, S. (2001). Ontology-Based Integration of Information – A Survey of Existing Approaches. *Proceedings of the IJCAI-01 Workshop on Ontologies and Information Sharing, Seattle,* 108-118. doi: 10.1.1.12.8073.

Yesin, V. I. (2012). Reengineering of existing databases. *Information processing systems, 3* (101), 188-191.

Ziegler, P. & Dittrich, K. R. (2004). Three Decades of Data Intecration – all Problems Solved? In: *Jacquart R. (eds) Building the Information Society. IFIP International Federation for Information Processing, Springer, Boston, MA, 156*, 3-12. doi: 10.1007/978-1-4020-8157-6_1.

Ziegler, P. & Dittrich, K. R. (2007). Data Integration – Problems, Approaches, and Perspectives. *Conceptual Modelling in Information Systems Engineering*, 39-58. doi: 10.1007/978-3-540-72677-7_3.