UDC 004.652

Maria G. **Glava**[1], Senior Lecturer of the Department of the Information Systems, E-mail: glavamg@gmail.com, ORCID: 0000-0002-9596-9556
Eugene V. Malakhov[2], Doctor of Technical Sciences, Professor, Head of the Department of the Mathematical Support of Computer Systems, E-mail: eugene.malakhov@onu.edu.ua, ORCID: 0000-0002-9314-6062
Olena O. Arsirii[1], Doctor of Technical Sciences, Professor, Head of the Department of the Information systems, E-mail: e.arsiriy@gmail.com, ORCID: 0000-0001-8130-9613
Borys F. Trofymov[1], Candidate of Technical Sciences, Associate Professor, Associate Professor of the Department of the Information systems, E-mail: btrofimoff@gmail.com, ORCID: 0000-0002-8590-8223
[1]Odessa National Polytechnic University, Shevchenko Avenue, 1, Odessa, Ukraine, 65044
[2]Odessa I. I. Mechnikov National University, Dvoryanskaya str., 2, Odessa, Ukraine, 65082

INFORMATION TECHNOLOGY FOR COMBINING THE RELATIONAL HETEROGENEOUS DATABASES USING AN INTEGRATION MODELS OF DIFFERENT SUBJECT DOMAINS

Abstract. *The work is devoted to solving the problem of combining heterogeneous relational databases based on integration models of different subject domains. The paper proposes methods for analyzing objects and their properties when combining models of subject domains, a method of combining integration models of different subject domains based on consistent rank evaluations of objects and the values of their typed essential properties. The model of the subject domain object is improved, which, unlike the classical one, takes into account the integration components that are important for combining: the sets of values of consistent ranks of properties and the sets of typed essential and non-essential properties of the object and their values determined on the basis of them. The subject domain model has been improved, which, unlike the existing one, takes into account certain combining scenarios and consistent ranking assessments of objects. Based on the proposed models and methods, an information technology for combining relational heterogeneous databases has been developed, which has increased the reliability of detection of subject domain objects and their properties to be combined, while simultaneously reducing the number of comparison operations for automated creation of a combined integration model of the subject domain.*

Keywords: *database; subject domain; object of the subject domain; subject domain model; object of subject domain model; object property*

Introduction

One of the strategic directions in the area of information technology (IT) is the creation of a single information space for the effective management of modern enterprises. But the trends of previous years in the development and implementation of independent information systems (IS), automating the activities of individual enterprises or their divisions, in practice led to a situation where information is stored in relational heterogeneous databases (DB) of local information systems for functional or organizational purposes. The existing redundancy, inconsistency and semantic heterogeneity of significant amounts of accumulated heterogeneous data in the DB of independent information systems impede data processing and promptly management decision-making.

Formulation of the problem. Previous studies show that using existing technology solutions, such as developing the data replication system, implementation of distributed databases or application programming interfaces (APIs) for accessing Enterprise Resource Planning (ERP) systems, allows to integrate information systems at the data level only by creating additional software. But such approaches do not provide prompt processing of mismatched and semantically heterogenic data. Therefore, it is considered effective to create a single information space of an enterprise using the information technology of combining relational heterogeneous databases into a single logical database based on integration models of particular subject domains (SDs) that determine the rules for structuring data for individual enterprises or their subdivisions.

Survey of prior research. Enterprises typically spend between 20 and 40 per cent of their IT budget for evolvement their data through migration (changing the locations of data), conversion (changing data into other forms or states) or scrubbing (recoding or rekeying data to prepare it for subsequent usage) [1]. The practice of integrating of information systems shows that more than two-thirds of all resources in IT (tending, time and costs) are devoted to attempts of combining (achieving the interaction of) modules written by different people at different times, in different languages and technologies, powered by different platforms. This is primarily due to data heterogeneity.

The main factors of heterogeneity of data and their sources are [2]:

– various types of data (logical, integer, real, object, etc.);

– various nature of the data (numeric arrays, texts);

– various database models - relational, hierarchical, object-oriented, network, multidimensional, etc.;

– various data presentation formats;

– differences in the degree of distribution of data storage systems;

– differences in the degree of reliability and accuracy of data measured at different scales and units of measurement;

– differences in the degree and form of data structuring, etc.

The use of heterogeneous "components" can cause difficulties both in solving problems of enterprise management or information exchange and in managing these components themselves, their support and administration. All this leads to the need to resolve the issue of compatibility of different systems.

Research in this area is quite dynamic and popular. Their main results are given in [3-15]. Most of the researchers suggested various classifications at different stages of data integration.

K. R. Dittrich [10] proposed a classification of data integration technologies. The scheme K. R. Dittrich allows to link together the integration of data with the integration of information – gradually moving upwards; simple elementary data acquire semantic content, become accessible to understanding and turn into useful information presented in a convenient form.

In [3], the integration of data at the physical, logical and semantic level is considered. The integration of data at the physical level is reduced to the conversion of data from various sources into the required uniform format of their physical representation. Integration of data at the logical level provides for the possibility of access to data contained in various sources in terms of a single global scheme describing their joint presentation taking into account structural and, possibly, behavioral (using object models) data properties. In this case, the semantic properties of the data are not taken into account. The support of a unified presentation of data, considering their semantic properties in the context of a unified ontology of the subject domains, is attained through data integration at the semantic level.

A classification, interpretation of uncertainties and an ontological approach to the integration of incomplete and inaccurate data were proposed in [7]. The above-mentioned methods allow to avoid possible contradictions in the integration of information resources that may arise due to the different nature of uncertainties, and also to determine the ways and procedures for processing integrated data includes the uncertainty.

P. Ziegler [9] proposed to consider data sources as structured, semi-structured and unstructured, as well as an approach that complements the existing integration approaches, suitable for situations with significant heterogeneity of data.

One of the general solutions to the integration problem is based on the description of the DB metadata within the framework of the developed methodology and the implementation of the mapping of entities and relationships of the databases in terms of a common information field, which is defined by the subject domain ontology [16, 17].

Conceptual database models are created in accordance with the standards of XML and Resource Description Framework (RDF) Schemas. They are then used to create a common metamodel that combines the representations of the entities of two or more data stores [18].

An ontology is a data dictionary that includes both terminology and a system behavior model [19]. Since each conceptual subject domain model is a subset of ontology, the task of combining the database is reduced to the task of combining the metamodels of the database that is, building mappings between these metamodels, in terms of ontology.

When combining database metamodels, similar problems arise in the search for denoted data to be combined in order to avoid their redundancy [20]. Analysis of the ontology comparison studies proves that the currently proposed methods mainly need the improvement for further use in the integrations of databases reflecting other subject domains; the task is solved mainly for individual cases and requires additional research.

A number of methods for combining relational heterogeneous databases based on data schemas are also proposed.

The method of integrating data schemas is based on the semantic description of attributes in the form of a set of symbolic patterns, on the basis of which the semantic similarity of attributes is assessed, and on the basis of this assessment in its turn, a measure of database relations converging is calculated [21]. This method assumes that semantically identical attributes have an equal number of occurrences of attribute values matching the criteria of a set of patterns. But any character pattern can be repeated in semantically different attributes: for example, the name of the city and person surname.

Also, this method does not describe the approach to matching attributes of non-character types.

The method of detecting previously unknown functional dependencies is based on the analysis of a variety of relational database data [22]. The first step is getting a set of functional dependencies for each relationship. In the second step, a similar operation is performed for the universal relationship of the given relational database. At this stage, it becomes possible to identify functional dependencies between the attributes of various relationships i.e. relationships defined during the operation of a relational database. A method for determining their informational novelty is proposed, which consists in checking the membership of the functional dependencies of the universal relation in the closure of the sets of functional dependences of the particular relations. This method does not take into account the semantics of the data, a high probability of obtaining random functional dependencies, and also does not consider the problem of comparing the universal relations of the combined databases.

In [23], an object representation was proposed that would adequately depict a relational database. Using the vocabulary of the subject domain to build the object representation of a relational database makes it possible to establish a single and understandable terminology for naming objects and attributes. The proposed mechanism of identifying attributes allows setting up the correspondence between the elements of the object representations of the integrated databases. Development of software for the implementation of this method requires considerable material and time costs, and also depends on the subject domains being combined and requires studying the structure of each database. At the same time, the software is complex and not universal.

There is offered [24; 25] to combine databases using the formulation of a universal (standard) data model based on the semantic "object-event" data model, set theory and logical calculus. The universal data model, on the one hand, is a set of standard mathematical relationships used to describe the data, the relations between them, and the constraints that are imposed on them by any subject domain. On the other hand, according to the definition of the data model and the selected modelling object, it is a modelling tool for any subject domain that is easily implemented within the framework of the relational data model and can be used, among other things, to build a database model. In the "object-event" model, all objects, processes, and events of any subject domain are described using meta-ontologies.

In most methods of database combining at the semantic level, to confirm the correctness of the result, it is necessary to involve experts. Using existing methods, it is impossible to integrate ontologies created by different working groups without the participation of experts. This is the main disadvantage of the proposed methods.

Therefore, based on the above, when integrating heterogeneous databases into a single database, it is necessary to combine subject domain models, and in order to avoid data redundancy, to identify both identical subject domain objects and their properties. Studies show that the classical subject domain model [26], represented by a tuple of objects sets E and relationships R between them, and each object, in turn, with a set of properties A – needs to be refined, because it allows to identify the same objects of subject domain and their properties only by name. To implement the operations of manipulating the subject domain models [27; 28], the SD model was expanded by introducing the concept of mass problems P [29], solved over the subject domain and influencing the model formation of this SD:

$$d = \langle E, R, P \rangle, \quad (1)$$

where: $E = \left\{ e_j \mid j = \overline{1,l} \right\}$; e_j – j-th object of SD; determined as $e_j = \left\{ a_{ji} \mid j = \overline{1,l},\ i = \overline{1,f_j} \right\}$; a_{ji} – the name of i-th property of j-th object; f_j – number of properties of j-th object; l – number of objects of SD, $R = \left\{ r_i \mid i = \overline{1,v} \right\}$; v – number of relationships, $P = \left\{ p_i \mid i = \overline{1,c} \right\}$; c – number of mass problem, solved over SD.

In the framework of the proposed operation of combining SD models, the formal definitions of objects being compared and objects to be combined and their properties are not presented.

The need to create integration models is caused by the fact that the use of the classic "entity-relationship" subject domain model as the basis for combining the SD models allows to successively match all the objects of the SD to each other only by the name of all their properties, taking into account the existing relationships. According to the existing method of comparing objects based on the "entity-relationship" model, those are considered similar for which there is a direct correspondence between the names of objects and their properties or the presence of their synonyms in previously created vocabularies of such names.

The analysis of existing approaches to matching objects has shown that, in particular, an approach based on the creation of vocabularies of synonyms for

object names and their properties is quite a laborious, complicated and nontrivial process, depending on the qualifications of experts, since it requires the creation of corresponding vocabularies of synonyms and analysis of all names of objects and their properties.

The general purpose. The general purpose of the study is to increase the reliability of detection of SD objects and their properties to be combined, while simultaneously reducing the number of operations to comparing them in the process of combining relational heterogeneous databases by creating appropriate information technology based on the developed integration models of subject domains.

Research methods. When solving the study issues, methods of non-parametric mathematical statistics and methods of mathematical processing of expert estimates were used to determine consistent ranking estimates of the SD objects and their properties as well as methods of cluster, histogram, correlation and structural analysis in the process of determining regular expressions to compare the properties of the SD objects with their typified values and methods of object-oriented design and programming in the development of information system for the connections of heterogeneous relational databases along with the computer simulation methods in the development of the IT components.

Detailed report of the main research matter

The IT has been developed for combining relational heterogeneous databases (CRHDB), the block diagram of which is presented in Fig. 1. The proposed technology is implemented as an IS CRHDB software.

To develop IT CRHDB the following tasks were set and solved:

– an information model of functioning databases has been developed;

– a method for identifying the essential properties of SD objects has been developed;

– a method for determining the ranking of objects of the SD has been developed;

– a method of combining integration models of particular SDs has been developed;

– a common model of the SD and the combined database have been developed;

– the approbation of the developed technology carried out.

As a part of the first task solution, the data stored in the functioning combined databases was preliminary processed. Preliminary processing of data means reducing to the same representation (placing data in one or several properties) in both databases such properties as person name, surname, patronymic, address, ID data, etc. Third-party software is used to implement this stage. For example, SQL Server Integration Services.

Next, the *information models* of both SD are built, to be combined in the form of (1), using standard database management system tools that support the corresponding databases.

The method proposed in the framework of solving the second *task for identifying the essential properties of SD objects consists* of six steps and is implemented as follows [30; 31].

Step 1. Data collection using standard tools of collecting statistical data for a certain period of database functioning to obtain the matrix of scores *Ch* of the statistical characteristics Ch_i of each property a_i of each object *e* of SD:

$$Ch_i = \left\{ g_i, sl_i, wh_i, jn_i, tr_i, vw_i, pr_i \mid i = \overline{1, f} \right\}, \quad (2)$$

where: sl_i, wh_i, jn_i are estimates of the number of addressing in events implementing relational operations of projection (select), selection (where) and joining (join), respectively;

tr_i, vw_i, pr_i are estimates of the number of occurrences of a property in the body of triggers or trigger functions, views, and stored procedures, respectively.

Step 2. Line by line processing of the *Ch* score matrix to convert the Ch_i (2) score values into a rank scale to obtain r_i^{Ch}:

$$r_i^{Ch} = \left\{ r_i^{g}, r_i^{sl}, r_i^{wh}, r_i^{jn}, r_i^{tr}, r_i^{vw}, r_i^{pr} \mid i = \overline{1, f} \right\}. \quad (3)$$

Step 3. Checking the consistency of rank scores of r_i^{Ch} based on the Kendall's W coefficient of concordance for rejecting random estimation results. Testing the significance of *W* using statistics distribution of the Pearson χ^2 test.

Step 4. Processing on the matrix r^{Ch} columns in order to consist the ranks of each property a_i using the methods of median ranks cr_i^M and Kemeny's median cr_i^K, as well as calculating the generalized consistent rank $cr_i = \min\left(cr_i^M, cr_i^K\right)$.

Step 5. Ranking properties a_i in order of increasing values cr_i^M, and cr_i^K. Comparison of the elements of the ranked sequences $a_{cr_i^M}$ and $a_{cr_i^K}$ and the determination of the threshold rank of the essential properties of *z* in one of the following options:

– equal to the made consistent rank, which correspond to different properties;

– set by an expert in a certain SD in the case of impossibility of automated determination.

Step 6. Clustering the properties into the sets of essential A_c and unessential A_u, the properties of the object:

$$\forall a_i = \begin{cases} a_i \in A_c \,|\, cr_i \le z,\ i = \overline{1, f} \\ a_i \in A_u \,|\, cr_i > z,\ i = \overline{1, f} \end{cases}, \qquad (4)$$

where: z is the set threshold rank.

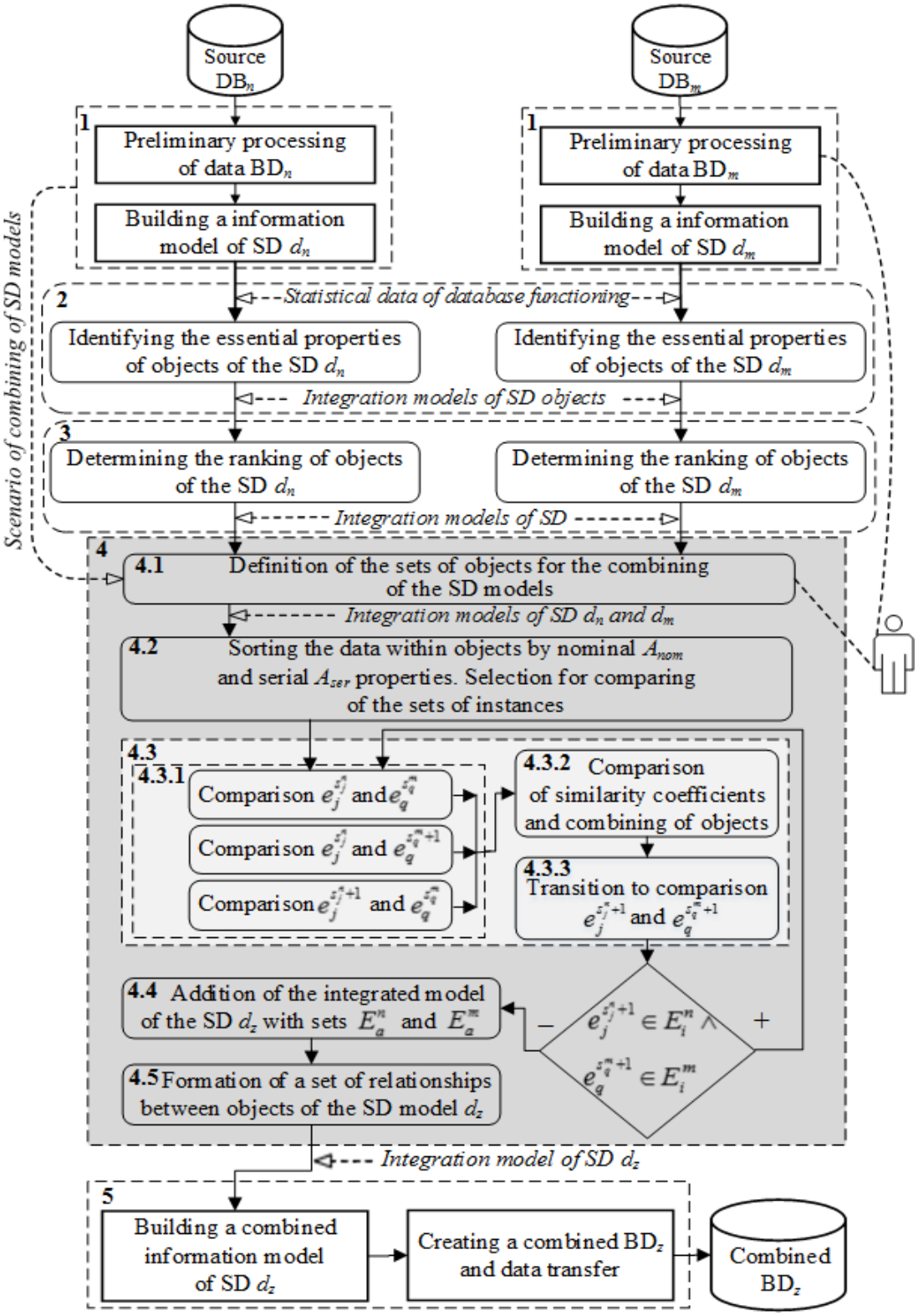

Fig. 1. Block diagram of information technology CRHDB:

1 – building information models of functioning databases; 2 – the method for identifying the essential properties of SD objects; 3 – the method for determining the ranking of objects of the SD; 4 – method of combining the integration models of particular SDs; 5 – building common model of the SD and the combined database

In order to make it possible to compare object models of an SD in the process of combining them not only by property names, but also taking into account the values of these properties, it is proposed to consider the object model *e* of SD in a tipple which components are the set of names of its relational properties *A* and the *K* values of these properties:

$$e = \langle A, K \rangle, \quad (5)$$

where: $K = \left\{ k_{ib} \mid i = \overline{1, f},\ b = \overline{1, g_i} \right\}$; k_{ib} is *b*-th value of *i*-th property; g_i is a number of values of *i*-th property.

Taking into account the method of analyzing the properties of the SD objects, the *e* (5) object model takes the following form:

$$e = \langle CR, A_c, A_u, K \rangle, \quad (6)$$

where: $CR = \left\{ cr_i \mid i = \overline{1, f} \right\}$; $\underline{cr}_i$ is made consistent rank of the *i*-th property.

The method for determining the ranking of objects of the SD is implemented as follows.

Step 1. Data collection using standard tools of collecting statistical data for a certain period of database functioning to obtain a matrix of estimates *Che* of the statistical characteristics Che_j of each object e_j of SD:

$$Che_j = \left\{ \begin{matrix} ct_j, fk_j, sv_j, in_j, up_j, tin_j, tup_j, tf_j, \\ mp_j \mid j = \overline{1,l} \end{matrix} \right\}, \quad (7)$$

where: ct_j is the number of instances;
fk_j is the number of foreign keys and the following estimates of the number of addressing;
sv_j – in relational projection operation (“select”) and views, in total;
in_j, up_j – in the data manipulation operators “insert” and “update”, respectively;
tin_j, tup_j – in the “insert” and “update” operators that activate the trigger, respectively;
tf_j – in the body of the trigger or trigger function,
mp_j – in the materialized view.

Step 2. Line by line processing of the *Che* score matrix to convert the Che_j (7) score values into a rank scale to obtain s_j^{Che}:

$$s_j^{Che} = \left\{ s_j^{ct}, s_j^{fk}, s_j^{sv}, s_j^{in}, s_j^{up}, s_j^{tin}, s_j^{tup}, s_j^{tf}, s_j^{mp} \mid j = \overline{1,t} \right\}. \quad (8)$$

Step 3. Checking the consistency of rank scores of s_j^{Che} based on the Kendall's *W* coefficient of concordance for rejecting random estimation results. Testing the significance of *W* using statistics distribution of the Pearson χ^2 test.

Step 4. Processing on the matrix s^{Che} columns in order to consist the ranks of each property e_j using the methods of median ranks s_j^M.

Step 5. Ranking properties e_j in order of increasing values s_j^M.

Step 6. Assigning to the objects e_j of SD of the values of consistent ranking estimates s_j, starting with one.

Consequently, taking into account the made consistent ranking estimates, the object in model (6) has the form $e_j^{s_j^d}$.

And the SD model (1), taking into account the method for determining the rank estimates of SD objects, takes the following form:

$$d = \langle E, R, P, S \rangle, \quad (9)$$

where: $S = \left\{ s_j^d \mid j = \overline{1,l} \right\}$; s_j^d is the rank estimate of the *j*-th object in the SD's *d*.

The method of combining integration models of particular SDs is based on a pair-wise comparison of integration models of SD objects. Suppose that there are set of objects $E^n = \left\{ e_1^{s_j^n}, ..., e_t^{s_j^n} \right\}$ and $E^m = \left\{ e_1^{s_q^m}, ..., e_l^{s_q^m} \right\}$ in the SD models d_n and d_m, where: *j* and *q* are the numbers of the rank rating made consistent, *t* and *l* are the numbers of objects of SD models d_n and d_m, respectively.

In order to reduce the number of object comparison operations in the process of combining the SD models, it was proposed to choose one of two possible scenarios *C* for the detection of objects to be compared. Scenario *C* is selected taking into account the mass problems *P* solved over *SD*, which models are combined.

When combining the SD models in the first scenario, it was proposed to compare all the objects that are potentially similar according to the made consistent ranking estimates while in the second scenario – only the objects that aren’t peculiar to a certain SD.

In order to combine the d_n and d_m SD models, it is necessary to compare only the sets of objects E_i^n and E_i^m corresponding to the scenario, and to supplement the combined model with sets of objects E_a^n and E_a^m that are not to be combined.

The method of combining the integration models of particular SDs is implemented as follows.

Step 1. Definition of the sets of objects E_i^n and E_i^m for the combining (Fig. 2).

Step 2. Clustering the set of essential properties of A_c in order to increase the assurance of detecting the properties of objects to be combined in the process of integrating heterogeneous databases into

subsets of the nominal A_{nom}, numeric A_{num} and serial A_{ser} data type:

$$A_c = \langle A_{nom}, A_{num}, A_{ser} \rangle . \quad (10)$$

Sorting the data within objects E_i^n and E_i^m by nominal A_{nom} and serial A_{ser} properties. Selection for comparing of the sets of instances of the same cardinality with the absence of NULL values within the objects E_i^n and E_i^m.

Step 3. Combining objects E_i^n and E_i^m (Fig. 3):

Step 3.1. Pair-wise comparison of objects [32] is carried out on the basis of a comparison of the corresponding essential properties of objects of each data type according to (10):

a) with the same ranking scores made consistent in both SDs $e_j^{s_j^n}$ and $e_q^{s_q^m}$;

b) with the current ranking estimate made consistent in the SD d_n $e_j^{s_j^n}$ and one unit greater in the SD d_m $e_q^{(s+1)_q^m}$;

c) with the current rank estimate made consistent in the SD d_m $e_q^{s_q^m}$ and one unit greater in the SD d_n $e_j^{(s+1)_j^n}$.

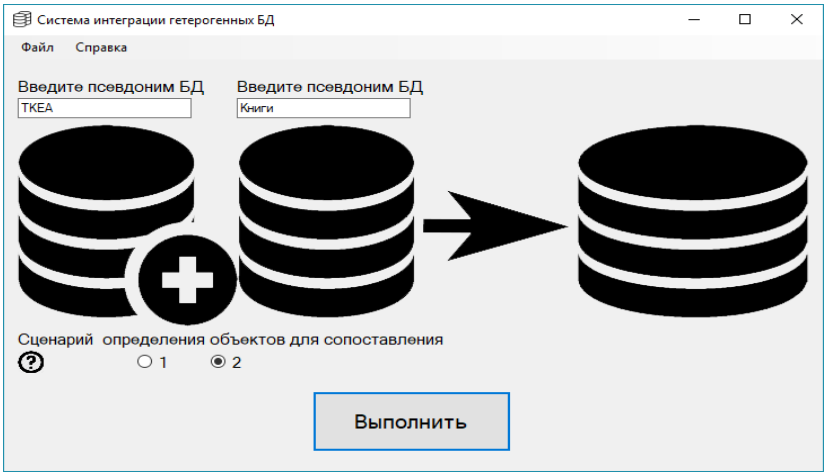

Fig. 2. The choice of databases and scenario of their combining

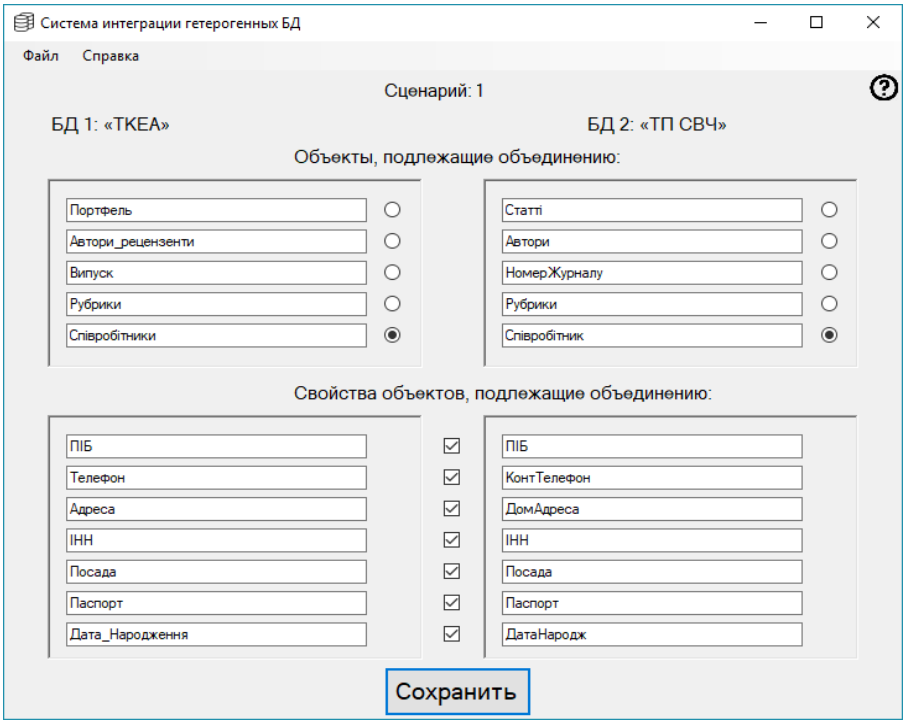

Fig. 3. The result of the detection of SD objects and their properties to be combined

To implement the operation of comparison the properties of each data type, appropriate procedures have been proposed.

To implement the procedure for comparing the nominal properties of A_{nom} objects of different SD, it was proposed to use the estimates of the structural characteristics of the property values obtained by using regular expressions [33]:

$$MD_i = \left\{ \begin{matrix} sp_{ib},\ cl_{ib},\ dt_{ib},\ cm_{ib},\ hp_{ib},\ ab_{ib},\ pt_{ib}, \\ ps_{ib} \mid i = \overline{1, |A_{nom}|}, b = \overline{1, g_i} \end{matrix} \right\}, \quad (11)$$

where: MD_i is the set of values of structural characteristics;

sp_{ib} is the number of spaces;

cl_{ib} is the number of capital letters;

number of punctuation marks: dt_{ib} of ".", cm_{ib} of ",", hp_{ib} of "-";

ab_{ib} is the presence of abbreviations;

pt_{ib} is the presence of quotes;

ps_{ib} is part of speech;

i is the number of the property of the j-th object of the SD d;

b is the number of the value of the i-th property.

The decision on the similarity of the obtained estimates of the structural characteristics of the property values is made on the basis of multidimensional statistical processing by the method of "Correspondence analysis".

The procedure for comparing the numerical properties of A_{num} consists of the following stages [34]: checking the coincidence of the distribution law for the values of potentially similar properties, their grouping using k-means and histograms, making decisions about the similarity of properties based on a comparison of the corresponding centers of the formed clusters.

The procedure for comparing the ordinal properties of A_{ser} involves analyzing properties with the data type "date" and primary keys of a numeric type containing a semantic characteristic using correlation analysis. Property values with the data type "date" are subject to preprocessing by separating the year from the property values.

If several objects have the same consistent ranking estimate, *Step 3.1* is repeated for each of these objects.

Step 3.2. Comparison of similarity coefficients of the objects mapped in step 3.1. If similar objects:

a) were: not revealed – transfer of objects $e_j^{s_j^n}$ and $e_q^{s_q^m}$ with the current rank assessment made costistent to the SD model d_z unchanged;

b) were: found – combining of objects with the maximum coefficient of similarity in the SD model d_z.

Step 3.3. Transition to object comparison (*Step 3.1*) with the following ranking estimates made consistent in the SD d_n and d_m.

The number of repetitions of *Step 3* is equal to the cardinality of the set of objects, according to which the models of the SD are combined according to the selected scenario.

Step 4. Addition of the integrated integration model of the SD d_z with sets of objects E_a^n and E_a^m not subject to combining.

Step 5. Formation of a set of relationships between objects R_z of the combined integration model of the SD d_z by combining the sets of the relationships of both models of the SDs.

According to the classical object model of the SD $e = \left\{ a_i \mid i = \overline{1, f} \right\}$ and taking into account (6), (10) the integration model of the object e of SD takes the following form:

$$e = \langle CR, A_{nom}, A_{num}, A_{ser}, A_u, K \rangle. \quad (12)$$

And taking into account scenarios C, the integration model of the SD (9) has the following form:

$$d = \langle E, R, P, S, C \rangle. \quad (13)$$

Building a combined information model of SD d_z and creating a combined database. On the basis of the obtained combined information model of the SD d_z, a combined database is created. Tables and relationships between them are created by standard database building tools based on the resulting model. Data transfer is performed in stages:

1) instances of objects that cannot be combined according to scenario C;

2) by properties that are defined as similar with the exception of duplication of instances;

3) instances are supplemented with values that do not match the properties of the database tables being combined.

Approbation of the results

Approbation of the proposed information technology for combining relational heterogeneous databases has been carried out on existing databases of the book and magazine publishing house "Politehperiodika", which uses in its work several DB developed at different times to solve various problems. The "TDEE" database was created for keeping records and data storage of the scientific journal "Technology and Design in Electronic Equipment" and has a data scheme presented in Fig. 4. The "Books" database automates work with orders for the production of other printed products. The scheme of the "Books" database is presented in Fig. 5. The scheme of the combined database is shown in Fig. 6.

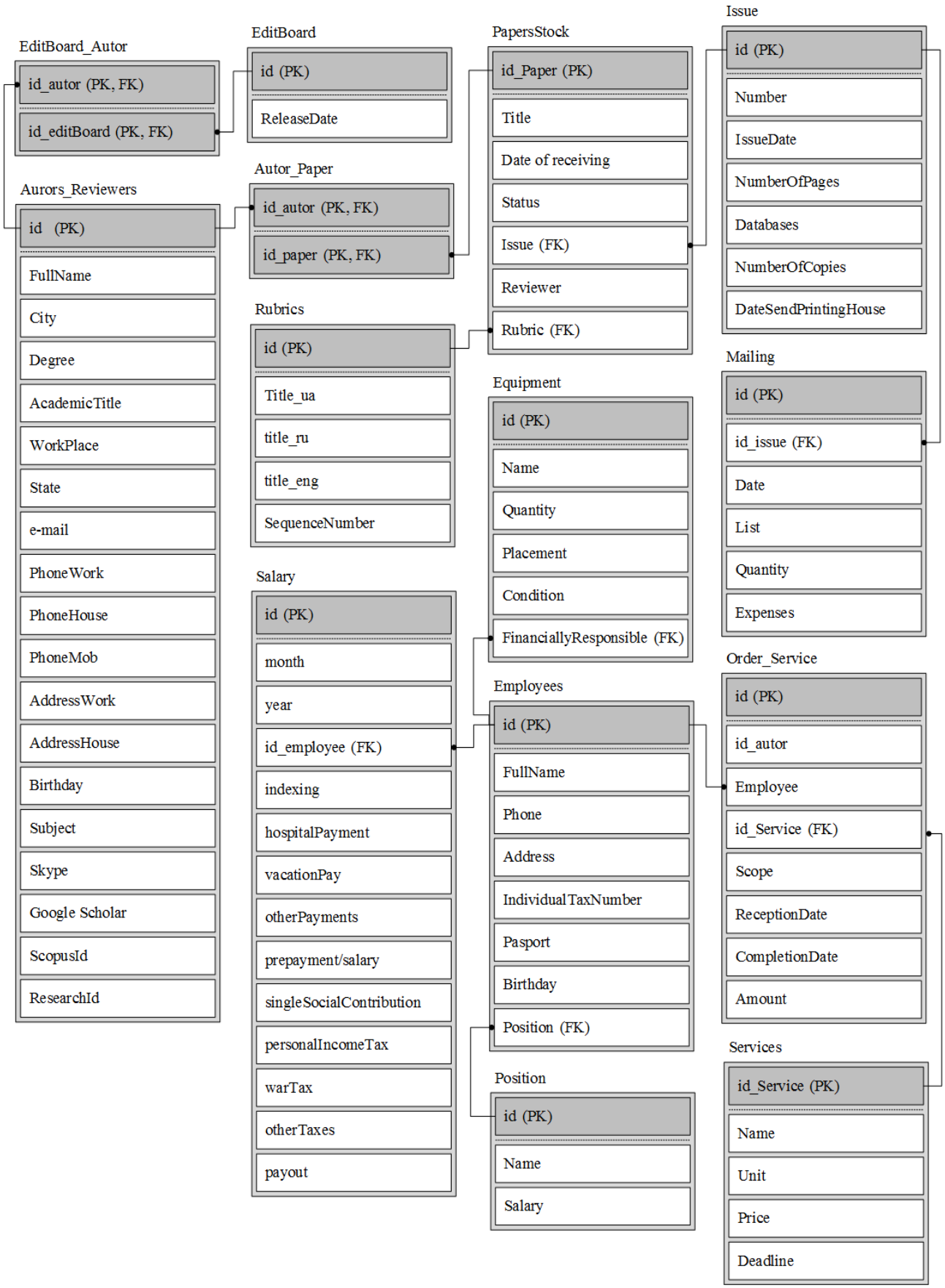

Fig. 4. Scheme of the database “TDEE”

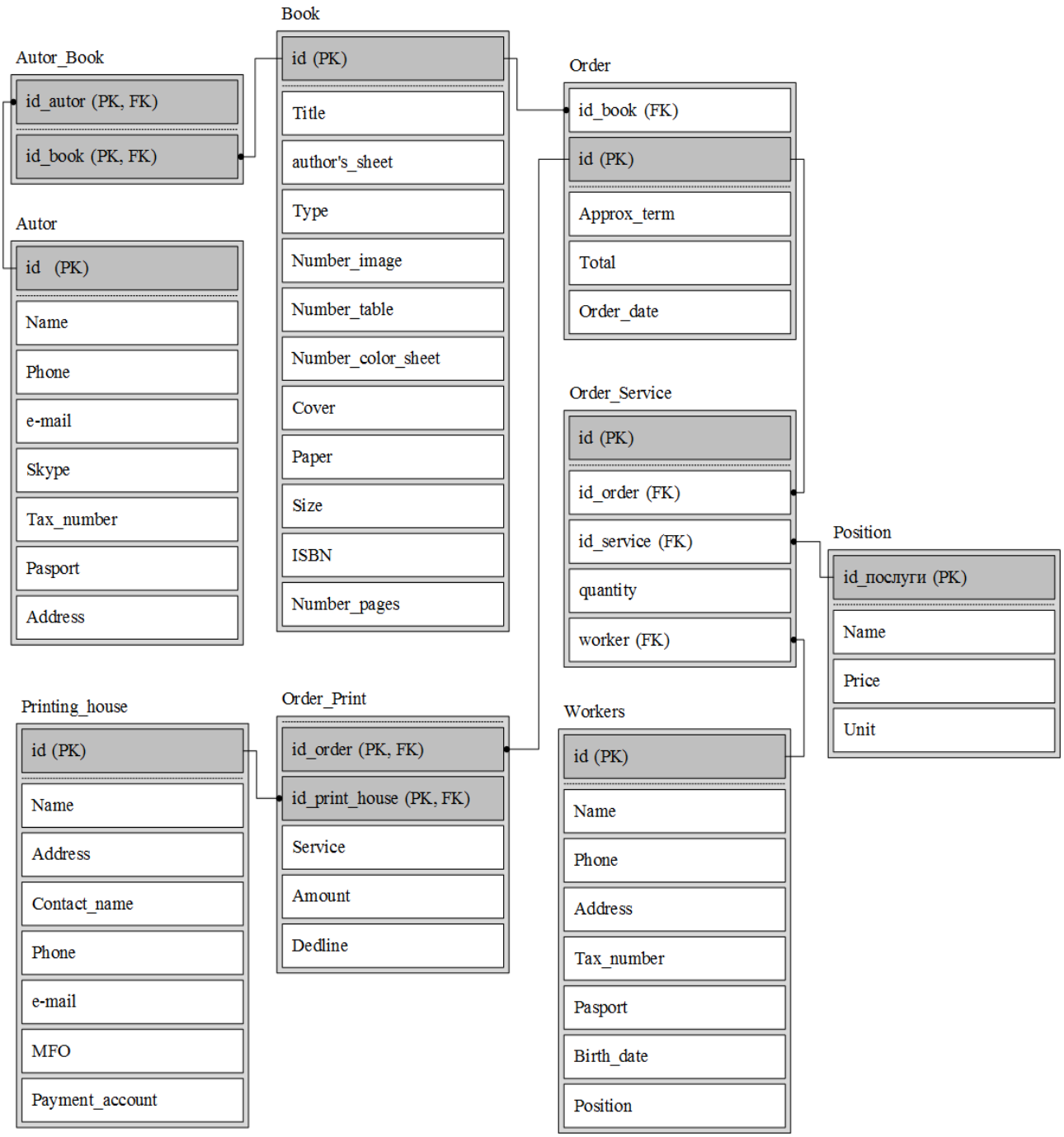

Fig. 5. Scheme of the database “Book”

To compare the results of combining functioning databases on the basis of the proposed and existing solutions, the definition of the number of operations of comparison of SD objects and their properties for the full enumeration method by the name of all objects and their properties is formalized – Q_{E_1} and Q_{A_1}, accordingly:

$$Q_{E_1} = l^n \times l^m, \tag{14}$$

$$Q_{A_1} = \sum_{j=1}^{l^n} f_j \times \sum_{j=1}^{l^m} f_j, \tag{15}$$

where: l^n, l^m is the number of objects of the SD model *n* and *m*, respectively.

Also the numbers of operations of comparison objects of SD and their properties are determined on the basis of the exhaustive search method using previously created vocabularies of object names and their properties – Q_{E_2} and Q_{A_2}, accordingly:

Fig. 6. Scheme of the combined database

$$Q_{E_2} = t_e \times \left(l^n + l^m \right), \quad (16)$$

$$Q_{A_2} = \sum_{y=1}^{t_s} th_y \times \left(\sum_{j=1}^{l^n} f_j + \sum_{j=1}^{l^m} f_j \right), \quad (17)$$

where: t_e is the number of values in the vocabulary of synonyms of object names;

th_y is the number of values in the vocabulary of synonyms for the names of the properties of the y-th object;

t_s is the number of objects in the vocabulary of synonyms of object names.

The reliability of the selection of the SD objects and their properties to be combined was calculated by [35]:

1) shares of true-positive rates of object classifications TPR_E and object properties TPR_A:

$$TPR = \frac{TP}{TP + FN}, \quad (18)$$

where: TP is the number of true-positively detected objects or their properties;

FN is the number of false-negatively detected objects or their properties;

2) shares of false-positive rates of object classifications FPR_E and object properties FPR_A:

$$FPR = \frac{FP}{TN + FP}, \quad (19)$$

where: FP is the number of false-positively detected objects or their properties;

TN is the number of true-negatively detected objects or their properties.

The numerical values of the indicators for calculating the shares of true positive and false positive objects and their properties when integrating the database of the publishing house "Politehperiodika" are presented in Table 1 and Table 2 for the exhaustive method and the exhaustive method using previously created vocabularies of object names and their properties, respectively, and in Table 3 – using the proposed IT CRHDB.

Table 1. The results of classifications for calculating the reliability of detection of objects and their properties by the method of complete enumeration (units)

$TP_E = 2$	$FP_E = 0$
$FN_E = 2$	$TN_E = 5$
$TP_A = 6$	$FP_A = 7$
$FN_A = 12$	$TN_A = 33$

Table 2. The results of the classification of indicators of the reliability of the detection of objects and their properties by the method of enumeration using previously created vocabularies (units)

$TP_E = 3$	$FP_E = 2$
$FN_E = 1$	$TN_E = 3$
$TP_A = 7$	$FP_A = 3$
$FN_A = 11$	$TN_A = 37$

Table 3. The results of the classification of indicators of the reliability of the detection of objects and their properties using IT CRHDB (units)

$TP_E = 2$	$FP_E = 0$
$FN_E = 2$	$TN_E = 5$
$TP_A = 10$	$FP_A = 0$
$FN_A = 8$	$TN_A = 40$

The calculation of the number of comparison operations for objects and their properties and the shares themselves for the complete enumeration method (1), the enumeration method using previously created object name dictionaries and their properties (2) and using the proposed IT CRHDB (3) – in Table 4 and Table 5, respectively.

Table 4. The values of the number of operations comparison objects and their properties (units)

1	2	3
$Q_E = 126$	$Q_E = 644$	$Q_E = 11$
$Q_A = 5510$	$Q_A = 21879$	$Q_A = 103$

Table 5. Values of reliability indicators for detection of objects and their properties (per-cents)

1	2	3
$TPR_E = 50$	$TPR_E = 75$	$TPR_E = 50$
$FPR_E = 0$	$FPR_E = 40$	$FPR_E = 0$
$TPR_A = 33,3$	$TPR_A = 38,9$	$TPR_A = 55,26$
$FPR_A = 17,5$	$FPR_A = 7,5$	$FPR_A = 0$

Conclusions and prospects for further research. The proposed solutions as part of the developed information technology and the automated system of combining subject domain models with the

integration of databases of functioning distributed information systems of the "Politehperiodika" publishing house made it possible to increase the accuracy of detecting objects and their properties to be combined, while simultaneously reducing the necessary comparison operations. At the same time, the proportion of falsely positively detected objects to be combined decreased by 40 %, and the properties of such objects – by 7,5 %. The share of truly positively detected properties of objects increased by 16,7 %, but the objects themselves decreased by 25 %. At the same time, the number of object property comparison operations decreased by an average of 18 %, while the number of object comparison operations decreased by more than five times.

References

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

2. Ananchenko, I. V., Gaykov, A. V., & Musaev, A. A. (2013). Tehnologii sliyaniya geterogennoy informatsii iz raznorodnyih istochnikov (data fusion), [Data fusion technologies for the heterogeneous information from diverse sources]. Bulletin of St PbSIT (TU), Vol. 19 (45) (in Russian).

3. Kogalovsky, M. R. (2010). Metodyi integratsii dannyh v informatsionnyh sistemah, [Methods of data integration in the information systems], Electronic Socionet depositor (in Russian).

4. Globa, L. S., Ternovoy, M. Ju. & Shtogrina, E. S. (2011). Ispolzovanie ontologiy dlya integratsii baz dannyh i baz znaniy, [Using the ontologies to integrate databases and knowledge bases], Intelligent Analysis of Information IAI-2011, pp. 34-38 (in Russian).

5. Globa, L., Ternovoy, M., & Shtogrina, O. (2011). Intehratsiia baz danykh ta baz znan na osnovi ontolohii, [Databases and Knowledge Bases Integration Based on Ontology], Collection of scientific paper MITIT NTUU "KPI", No. 1, pp. 43-47 (in Ukrainian).

6. Berko, A. Ju. (2010). Strukturno-semantychna intehratsiia danykh na osnovi faktolohichnoi reliatsiinoi modeli, [The structural and semantic data integration based on the factual relational model], Bulletin of the Lviv Polytechnic National University, No. 663, pp. 60-69 (in Ukrainian).

7. Berko, A. Y., & Vysotska, V. A. (2009). Semantychna intehratsiia nepovnykh ta netochnykh danykh, [Semantic integration of incomplete and imperfect data], *Information Processing Systems*, Vol. 7 (81), pp. 93-98 (in Ukrainian).

8. Chernyak, L. (2009). Integratsiya dannyh: sintaksis i semantika, [Data Integration: Syntax and Semantics], Open Systems, No. 10, http://www.osp.ru/os/2009/10/11170978 (in Russian).

9. 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, No. 156, pp. 3-12, https://doi.org/10.1007/978-1-4020-8157-6_1.

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

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

12. Filatov, V., & Radchenko, V. (2015). "Reengineering relational database on analysis functional dependent attribute". Proceedings of the X Intern. Scient. and Techn. *Conf. "Computer Science & Information Technologies"* (CSIT'2015), 14-17 sept., Lviv, Ukraine, pp. 85-88. DOI:10.1109/STC-CSIT.2015.7325438.

13. Filatov, V., Konstantinov, S. M., & Ponomarenko, Yu. L. (2016). Chastkove vidobrazhennia modelei danykh pry intehratsii informatsiinykh system, [Partial display of data models when integrating information systems], *Economicmathematical Modeling of Socio-economic Systems*, Vol. 21, pp. 140-158 (in Ukrainian).

14. Kungurtsev, A., Kovalchuk, S., Potochniak, Ia., & Shirokostup, M. (2016). Pobudova slovnyka predmetnoi oblasti na osnovi avtomatyzovanoho analizu tekstiv ukrainskoiu movoiu, [Creating the domain vocabulary on the basis of automated analysis of ukrainian texts], *Technical sciences and technologies*, No. 3 (5), pp. 164-174 (in Ukrainian).

15. Kungurtsev, A., Gavrilova, A., Leonhard, A., & Potochniak, Ia. (2016). Uchet mezhfrazovyh svyazey pri avtomatizirovannom postroenii tolkovogo slovarya predmetnoy oblasti, [Accounting of interphrase connections in automated development explanatory dictionary of some subject area], (2016). *Informatics and Mathematical Methods in Simulation*, Vol. 6 (2016), No. 2, pp. 173-183 (in Russian).

16. Vagin, V. N., & Mikhailov, I. S. (2008). Razrabotka metoda integratsii informatsionnyh sistem na osnove metamodelirovaniya i ontologii predmetnoy oblast, [Developing the methods for information systems integration based on metamodeling

and domain ontology], *Software & Systems*, No. 1, pp. 22-26 (in Russian).

17. Bubareva, O. A. (2014). Model, algoritmy i programmnoe obespechenie integratsii dannyh informatsionnyh sistem na osnove ontologiy (na primere VUZa), [Model, algorithms and software for the data integration of information systems based ontology (by the example of the university)], Thesis for the degree PhD. 05.13.11, Biysk (in Russian).

18. Mikhailov, I. S. (2008). Issledovanie i razrabotka metodov i programmnyh sredstv obespecheniya strukturnoy i semanticheskoy interoperabelnosti informatsionnyh sistem na osnove metamodeley, [Research and development of methods and software for ensuring structural and semantic interoperability of information systems based on metamodels], *Proceedings of the XI National Conference on Artificial Intelligence*, Dubna, No. 2, pp. 207-209 (in Russian).

19. Tusovsky, A. F., Chirikov, S. V., & Yampolskiy, V. Z. (2005). Sistemy upravleniya znaniyami (metody i tehnologii), [Knowledge management systems (methods and technologies)], Tomsk, Russian Federation, *NTL* (in Russian).

20. Saenko, I., Brunilin, A., Efimov, V., & Yasinsky, S. (2016). Organizatsiya informatsionnogo vzaimodeystviya raznorodnyh avtomatizirovannyh sistem: ontologicheskiy podhod, [Organizing the Information Interaction between Heterogeneous Automated Systems: an Ontological Approach], *Information and Space*, No. 2, pp. 60-64 (in Russian).

21. Komar, F. V., & Pogodaev, A. K. (2008). Metod integrirovaniya shem dannyh na osnove semanticheskogo opisaniya atributov, [The method of the data schemes integration based on attributes semantic description], *Software & Systems*, No. 1, pp. 53-55 (in Russian).

22.Radchenko, V. O., & Tanyansky, S. S. (2012). Vyiyavlenie skrytyh zavisimostey mezhdu dannymi v zadachah reinzhiniringa informatsionnyh system, [Discovery of hidden data relationships in tasks of information systems reengineering], *Information processing systems*, Vol. 3 (101), Iss. 2, pp. 203-205 (in Russian).

23. Kungurtsev, A. B., & Neizvestny, A. S. (2007). Matematicheskaya model ob'ektnogo predstavleniya relyatsionnoy bazy dannyh, [Mathematical model of objective representation of a relation database], Odes'kyi Politechnichnyi Universytet. Pratsi, Odesa, Ukraine, No. 1 (27), pp. 130-134 (in Russian).

24. Yesin, V. I. (2012). Reinzhiniring suschestvuyuschih baz dannyh, [Reengineering of existing databases], *Information processing systems*, Vol. 3 (101), Iss. 2, pp. 188-191 (in Russian).

25. Yesin V. I. (2012). Metody razrabotki baz dannyh dlya informatsionnyh system, [Methods of databases development for the information systems], Bulletin of Kharkiv national university, No. 1037, pp. 64-72 (in Russian)

26. Date, C. J. (2005). Vvedenie v sistemy baz dannyh, [An Introduction to Database Systems], *Viliams* (in Russian).

27. Malakhov, E. V. (2010). Informatsiina tekhnolohiia modeliuvannia skladnostrukturovanykh predmetnykh oblastei v systemakh orhanizatsiinoho upravlinnia (teoriia ta realizatsiia), [The information technology for complex structures subject domains simulating in the organizational management systems (theory and implementation)], Synopsis of doctorate thesis (Sc.D.), Odessa (in Ukrainian).

28. Filatova, T., & Glava, M. (2016). "Mathematical Models of Information Manipulation in the Subject Field of Intellectual Production in Educational Institutions", *International Conference on Electronics and Information Technology (EIT)*, May 23-27, 2016, Odesa, Ukraine, pp. 92-96. doi: 10.1109/ICEAIT.2016.7501000; eid: 2-s2.0-84979554925.

29. Malakhov, E. V. (2010). Rasshirenie operatsiy nad meta-modelyami predmetnyh oblastey s uchyotom massovyh problem, [Expansion of operations on meta-models of subject domains in view of mass problems], *Eastern-european journal of Enterprise Technologies*, No. 5/2 (47), pp. 20-24 (in Russian).

30. Malakhov, E. V., Vostrov, G. N., & Mikulinska, M. G. (2010). Metody opredeleniya stepeni vazhnosti svoystv suschnostey predmetnyh oblastey, [Methods of subject domains objects properties importance definition], *Refrigeration engineering and Technology*, No. 4 (126), pp. 73-77 (in Russian).

31. Glava, M., & Malakhov, E. (2018). Metod vydilennia vlastyvostei, yaki kharakteryzuiut obiekt predmetnoi oblasti, [Method of isolating the properties that characterize the domain object], *Refrigeration Engineering and Technology*, No. 54(2), pp. 68-72. https://doi.org/10.15673/ret.v54i2.1048 (in Ukrainian).

32. Glava M., & Malakhov E. (2016). "Searching Similar Entities in Models of Various Subject Domains Based on the Analysis of Their Tuples". *International Conference on Electronics and Information Technology (EIT)*, May 23-27, 2016, Odesa, Ukraine, pp. 97-100, doi: 10.1109/ICEAIT.2016.7501001; eid: 2-s2.0-84979503116.

33. Glava, M. (2016). Sravnenie svoystv nominalnogo tipa ob'ektov razlichnyh predmetnyh podoblastey v relyatsionnyh bazah dannyh, [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, pp. 302-309 (in Russian).

34. Glava, M. G., & Malakhov, E. V. (2018). "Comparison of numeric properties of objects of different data domains in relational databases", *Electronics and Control Systems*, No. 2 (56), pp. 99-105, doi: 10.18372/1990-5548.56.12943.

35.Gus'kov, S. Yu., & Lyovin, V. V. (2015). Intervalnye doveritelnye otsenki dlya pokazateley kachestva binarnyh klassifikatorov – ROC-krivyh, AUC dlya sluchaya malyh vyborok, [Confidence interval estimation for quality factors of binary classifiers – ROC curves, AUC for small samples], *Engineering Journal: Science and Innovation*, Vol. 3, http://engjournal.ru/catalog/mesc/idme/ 1376.html (in Russian).

Received 20.12.2019

УДК 004.652

[1]**Глава, Марія Геннадіївна,** старший викладач каф. інформаційних систем, E-mail: glavamg@gmail.com, ORCID: 0000-0002-9596-9556
[2]**Малахов, Євгеній Валерійович,** доктор техніч. наук, професор, зав. каф. математичного забезпечення комп'ютерних систем, E-mail: eugene.malakhov@onu.edu.ua, ORCID: 0000-0002-9314-6062
[1]**Арсірій, Олена Олександрівна,** доктор техніч. наук, професор, зав. каф. інформаційних систем, E-mail: e.arsiriy@gmail.com, ORCID: 0000-0001-8130-9613
[1]**Трофимов, Борис Федорович,** кандидат техніч. наук, доцент, доцент каф. інформаційних систем, E-mail: btrofimoff@gmail.com, ORCID: 0000-0002-8590-8223
[1]Одеський національний політехнічний університет, пр. Шевченка, 1, м. Одеса, Україна, 65044
[2]Одеський національний університет імені І. І. Мечникова, вул. Дворянська, 2, м. Одеса, Україна, 65082

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ ОБ'ЄДНАННЯ РЕЛЯЦІЙНИХ ГЕТЕРОГЕННИХ БАЗ ДАНИХ НА ОСНОВІ ІНТЕГРАЦІЙНИХ МОДЕЛЕЙ ОКРЕМИХ ПРЕДМЕТНИХ ОБЛАСТЕЙ

Анотація. *Робота присвячена вирішенню задачі об'єднання реляційних гетерогенних баз даних на основі інтеграційних моделей окремих предметних областей. В роботі запропоновано методи аналізу об'єктів та їх властивостей при об'єднанні моделей предметних областей, метод об'єднання інтеграційних моделей окремих предметних областей на підставі узгоджених рангових оцінок об'єктів та значень їх типізованих суттєвих властивостей. Удосконалено модель об'єкта предметної області, яка на відміну від класичної враховує важливі при об'єднанні інтеграційні складові: множини значень узгоджених рангів властивостей та визначені на їх основі множини типізованих суттєвих і несуттєвих властивостей об'єкта та їх значень. Удосконалено модель предметної області, яка на відміну від існуючої враховує визначені сценарій об'єднання та узгоджені рангові оцінки об'єктів. На основі запропонованих моделей і методів розроблено інформаційну технологію об'єднання реляційних гетерогенних баз даних, що дозволила збільшити достовірність визначення об'єктів предметних областей та їх властивостей, що підлягають об'єднанню, з одночасним зменшенням кількості операцій їх зіставлення при автоматизованому створенні об'єднаної інтеграційної моделі предметної області.*

Ключові слова: *база даних; предметна область; об'єкт предметної області; модель предметної області; модель об'єкта предметної області; властивість об'єкта*

[1]**Глава, Мария Геннадьевна,** старший преподаватель каф. информационных систем, E-mail: glavamg@gmail.com, ORCID: 0000-0002-9596-9556
[2]**Малахов, Евгений Валерьевич,** доктор технич. наук, профессор, зав. каф. математического обеспечения компьютерных систем, E-mail: eugene.malakhov@onu.edu.ua, ORCID: 0000-0002-9314-6062
[1]**Арсирий, Елена Александровна,** доктор технич. наук, профессор, зав. каф. информационных систем, E-mail: e.arsiriy@gmail.com, ORCID: 0000-0001-8130-9613
[1]**Трофимов Борис Федорович,** кандидат технич. наук, доцент, доцент каф. информационных систем, E-mail: btrofimoff@gmail.com, ORCID: 0000-0002-8590-8223
[1]Одесский национальный политехнический университет, пр. Шевченко, 1, г. Одесса, Украина, 65044
[2]Одесский национальный университет имени И. И. Мечникова, ул. Дворянская, 2, г. Одесса, Украина, 65082

ИНФОРМАЦИОННАЯ ТЕХНОЛОГИЯ ОБЪЕДИНЕНИЯ РЕЛЯЦИОННЫХ ГЕТЕРОГЕННЫХ БАЗ ДАННЫХ НА ОСНОВЕ ИНТЕГРАЦИОННЫХ МОДЕЛЕЙ ОТДЕЛЬНЫХ ПРЕДМЕТНЫХ ОБЛАСТЕЙ

Аннотация. *На основе проведенного анализа существующих подходов к разработке и внедрению независимых информационных систем, которые автоматизируют деятельность отдельных предприятий или их подразделений создана информационная технология объединения реляционных гетерогенных баз данных на основе разработанных интеграционных моделей предметной области и ее объекта. Разработка интеграционных моделей предметной области и ее объекта базируется на предложенных методах выявления существенных свойств объектов предметных областей и определения ранговых оценок объектов предметных областей. Интеграционная модель объекта предметной области учитывает важные при объединении составляющие: множества значений согласованных рангов свойств и определенные на их основе множества типизированных существенных и несущественных свойств объекта и их значений. Интеграционная модель предметной области учитывает определенные сценарий объединения и согласованные ранговые оценки объектов. Предложена структура информационной технологии, составными элементами которой являются разработанные интеграционные модели предметной области и ее объекта и методы их анализа и сопоставления. При апробации разработанной информационной технологии объединения реляционных гетерогенных баз данных на примере реализации сценария интеграции существующих реляционных баз данных отдельных подразделений книжно-журнального издательства увеличено достоверность определения объектов предметных областей и их свойств, подлежащих объединению, с одновременным уменьшением количества операций их сопоставления в процессе интеграции.*

Ключевые слова: *база данных; предметная область; объект предметной области; модель предметной области; модель объекта предметной области; свойство объекта*