ISSN 2226-3780

UDC 004.9
DOI: 10.15587/2312-8372.2018.123527

Sahaida P.

DEVELOPMENT OF METHODOLOGY FOR DATA AND KNOWLEDGE WAREHOUSE DESIGN IN COMPUTER SYSTEMS FOR INTELLECTUAL DATA PROCESSING

Розроблено методологію проектування сховищ даних і знань для рішення завдань обробки й аналізу даних на основі категоріально-онтологічних моделей. Методологія поєднує проектування з використанням різних діаграмних методик і мов моделювання. Це дозволило усунути недоліки й доповнити переваги різних підходів до проектування й одержати структуру сховищ, близьку до оптимальної.

Ключові слова: *комп'ютерна система, сховище даних і знань, категоріально-онтологічна модель.*

1. Introduction

At present, it is necessary to develop and study models and methods for designing effective data and knowledge warehouses (DKW) of modern computer systems for intellectual data processing (CS for IDP) at enterprises and organizations [1, 2]. These models and methodologies should be combined into an end-to-end methodology based on a mathematically grounded approach that will ensure the provability and validity of design results. At the same time, it is necessary to take into account the accumulated knowledge of the subject area (SA) of specialists and knowledge engineers who do not have much experience in designing the CS, but can formulate a thesaurus or ontology describing the concepts and relationships between them. At the same time, the ontological model, constructed in general form on the basis of accumulated information about the work of the domain, has a number of significant drawbacks. When designing it, the analyst does not have to follow formal rules and restrictions. Despite the fact that a number of recommendations have been developed at the moment [3, 4], their value depends on the effectiveness of applied ontological modeling practices and is subjective. The most important possibility to introduce mathematical foundations, verification of results and the demonstrative power of topological design patterns into the knowledge engineering process is the category theory (CT). The mathematical apparatus of this theory makes it possible to develop commutative diagrams, including for the formalization of knowledge [5, 6]. Logical development of the CT – the theory of sketches [7, 8] – makes it possible to develop mathematical constructions in the composition of graphs, theories and categorical models.

The architecture of the projected DKWs in the CS composition should ensure the requirements of the database design theory [9], which implies the relationship of the relational data model to the constraints imposed on them by normal forms. At the same time, the skills and abilities of the project team members are also related to the graphical formalization of the results of system analysis using the appropriate diagrammatic techniques. This formalization of SA knowledge provides an opportunity to visualize the information received during the analysis, ensure effective interaction with colleagues and customers, move on to the following, more formal stages of the CS development.

Therefore, it is important to study the possibilities and disadvantages of modern approaches to formalizing knowledge about SA, diagrammatic techniques of information and data modeling in the development of the CS DKW for IDP. The result of such studies is development of a design methodology that summarizes the achievements in this field on the basis of a single mathematical apparatus.

2. The object of research and its technological audit

The object of research is development of a methodology for designing effective data and knowledge warehouses of modern computer systems for intelligent data processing using the categorical-ontological (CO) approach developed by the author [10]. The use of CO models as a metalanguage simulation provides a means for verifying the DKW design process and its results. SA analysts and DKW developers during the creation of a CS for IDP usually use at different stages of work disparate techniques and technologies for designing and implementing components of the CS. These techniques and technologies, based on heterogeneous approaches to formalizing and modeling aspects of SA, will be considered below. In this case, semantic and linguistic barriers arise in the SA perception, understanding of problems and methods for their solution, from the point of view of choosing an effective diagram technique in the design. Schematically, the nature and topology of the emerging barriers are shown in Fig. 1. In addition, the differences are also in the formats and technologies of data storage and exchange, in the models for building requests for DKW.

The main drawback of the current theory and practice of DKW design is that customers, analysts and developers do not have the opportunity to use one end-to-end methodology that is suitable for teamwork and knowledge sharing.

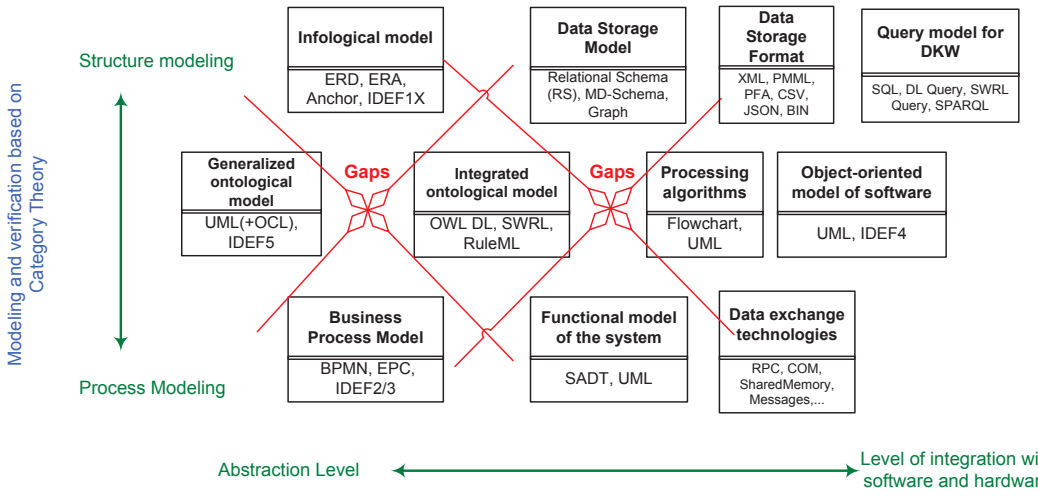

Fig. 1. Solution of the problem of semantic and linguistic gaps at the stages of implementation and in the process of functioning of computer systems for intellectual data processing at enterprises and organizations

Such methodology should be understandable and objective from the point of view of the provability and validity of models, be suitable for a full range of research and design work. Models created at different stages of design and using different methods should, in the framework of such a methodology, be displayed to each other without loss. These requirements are not currently met by any applied technique in itself. It is necessary to integrate them on the basis of mathematical theory and using the methods of knowledge engineering, in particular, ontological approach.

3. The aim and objectives of research

The aim of this research is improvement of the DKW design process by developing a methodology for their design based on the ontological approach to modeling the subject area of data processing automation and data analysis.

To achieve this aim it is necessary to:

1. Analyze the existing models and methods of formalizing knowledge about the SA of CS operation for IDP and DKW design, perform their classification, determine the features and possibilities of integration.

2. Develop a methodology for the DKW design and its formal categorical-ontological (CO) model, on the basis of which to justify the required procedure for applying the methodologies included in the unified methodology.

4. Research of existing solutions of the problem

Traditional approaches to the conceptual modeling and design of DKW can be classified as such that formalize, basically, Relational-oriented Approach (RoA), Attribute-oriented Approach (AoA) attributes and objects (Object-oriented Approach – OoA). And most of the design technologies are of a mixed nature. Thus, the technology using Entity-Relationship-Attribute (ERA) Diagrams [11] and Anchor technology [12] represent a combined (RoA-AoA) approach, and only the design based on the Functional Dependencies Diagrams is attribute-oriented (AoA). The ontological model of this group of technologies, more focused on the representation of attributes in models, is presented in Fig. 2.

The higher-level modeling technologies used in the design of the knowledge warehouse (KW), however, using the combined (RoA-AoA) approach, are the Entity-Relationship Diagrams (ERD) model [4, 9], and the framework Semantic Web, the Resource Description Framework (RDF) [13]. The ontology of these simulation technologies of the SA in the DKW design, more oriented to the representation of relations in models, is shown in Fig. 3.

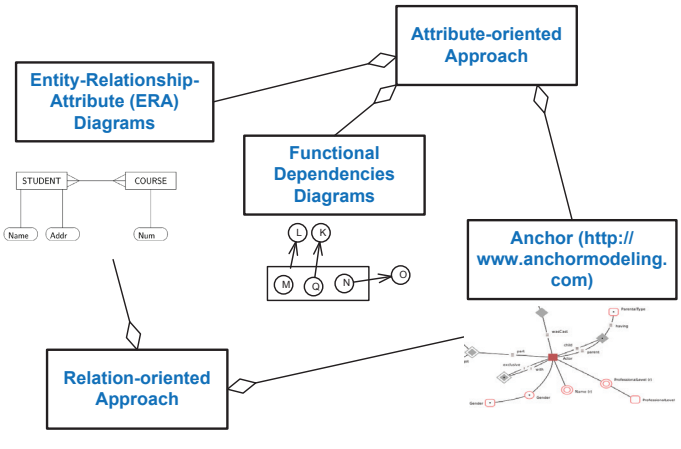

Fig. 2. Ontology of domain modeling technologies for the design of data warehouses and knowledge, more focused on the presentation of attributes

Historically, one of the first technologies for modeling software for designing a DW is to consider Object Role Modeling (ORM). Based on the ORM, the Integration DEFinition for information modeling (IDEF1X) technology, which is an integral part of the family of IDEF diagram techniques, was then developed (in the direction of the reduction presented in the knowledge models) to formalize various aspects of the SA. The result of the merging of a large number of similar approaches to modeling and formalization, as well as the design of hardware and software systems, was the Unified Modeling Language (UML) modeling technology [14], which provides ample opportunities for modeling classes, their instances and relationships between them.

ISSN 2226-3780

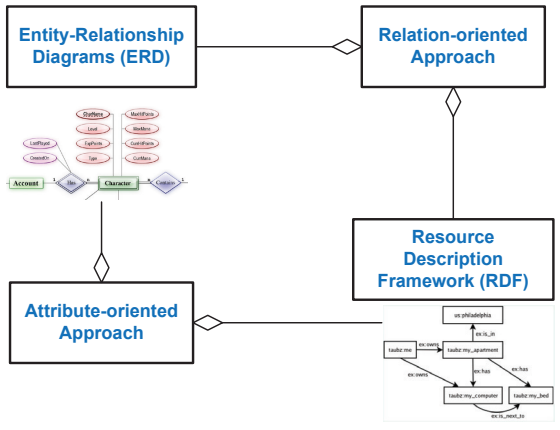

Fig. 3. Ontology of domain modeling technologies for the design of data warehouses and knowledge, more relationship-oriented

However, to represent business rules and interaction algorithms in objects, it was necessary to develop a specialized language based on UML diagrams – Object Constraint Language (OCL) [14]. The use of OCL made it possible to use for this purpose expressions that are close in expression to first-order logic, but specific and complex from the point of view of specialists in the PR. The ontology of the considered simulation technologies of the SA in the design of the data warehouse, oriented to the representation of objects and relations, is shown in Fig. 4.

To increase the level of abstraction in modeling and automating the CS development, in recent years, the Meta-Object Facility (MOF) initiative has been developed and is being implemented [15]. MOF involves the use of a standard for model-driven development prepared by the Object Management Group (OMG) on the basis of UML. The analysis shows that this approach is a metamodel over UML and has no other useful properties for the tasks solved in this paper.

Preferred as the basis for the developed methodology are technologies based on the categorical approach to the SA modeling, since this approach allows verification of information models, usually expressing the subjective point of view of their authors, on the basis of a formal mathematical apparatus. These technologies, based both on the representation of attributes and on the approach of category theory, include Olog Diagrams [6] and Sketches, which are the development of ERA technology [11]. The authors of these technologies have developed languages and means of automating the obtaining of database schemas from diagrams, respectively, Algebraic Query Language (AQL) and EASIC (graphical modeling of EA sketches and views). The ontology of modeling technologies for the design and development of attribute-oriented DKWs and the theory of category theory is presented in Fig. 5.

The disadvantage of these technologies is the absence in their composition of means for presenting rules for complex SA of operations of enterprises and organizations.

Thus, in order to overcome the considered semantic and linguistic barriers, it is necessary to integrate diverse design methods based on ontological models verified on the basis of category theory to construct an effective, theoretically grounded methodology, based on:

– possibilities of the theory of categories and sketches [5–8];
– approach to categorical-ontological modeling of the SA and the processes that take place in them [10];
– methods for the mutual mapping of ER and FDs diagrams [4] for the joint use of information and data models in the design of databases;
– methods for presenting queries to databases and knowledge, usually formalized in different query languages (SQL, DL Query, SPARQL, SQueryWRL) [13], using a subject-oriented language based on categorical-ontological models.

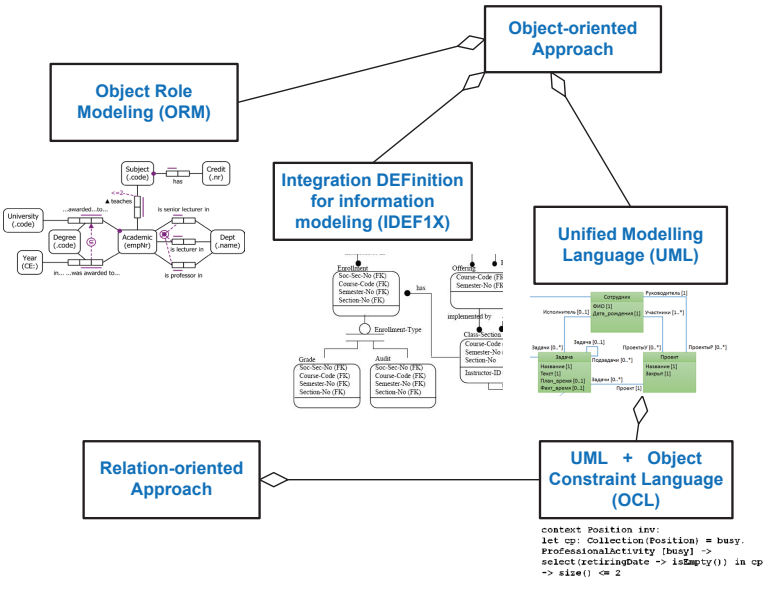

Fig. 4. Ontology of domain modeling technologies for the design of data warehouses oriented to the representation of objects and relations

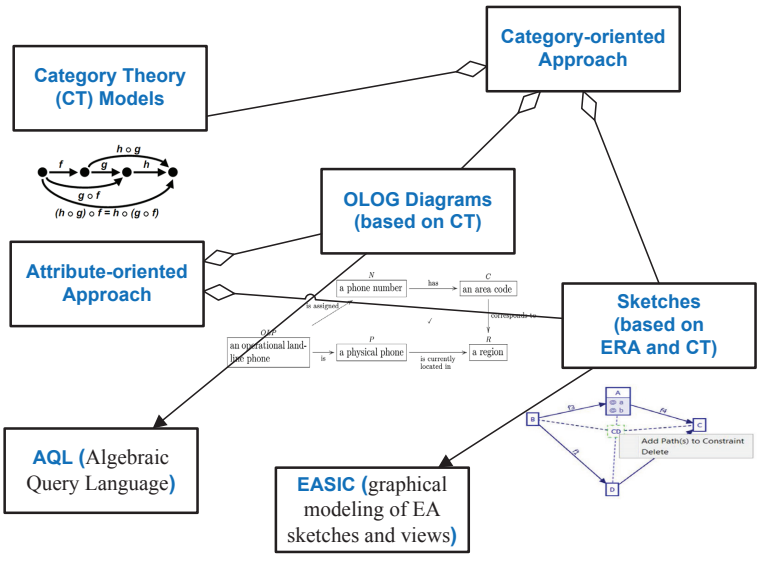

Fig. 5. Ontology of domain modeling technologies for the design of data warehouses and attribute-oriented knowledge and category theory approach

5. Methods of research

In this paper, a categorical-ontological approach is used to SA modeling, in which the generic ontological model is mathematically rigorous, by imposing constraints on the objects and morphisms of the category theory on the concepts and relationships that are presented. The one shown in Fig. 6 the diagram is a commutative diagram of the category theory constructed in accordance with the graph in sketch theory [16].

Since the arrows in category theory are function-transformations (mappings) of some objects into others, commutativity means the equality of paths (compositions of morphisms) by means of which the results of transformations are achieved.

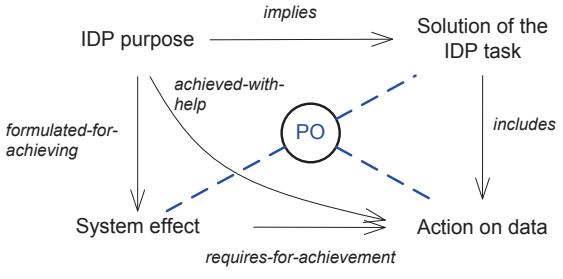

Fig. 6. Commutative diagram of category theory for a fragment of an ontological model for solving the problem of intelligent data processing

The imposition of the requirements of commutativity and the use of concepts in category theory made it possible to reveal for the given diagram a topological template mathematically described and justified as necessary for the given semantics of the domain, namely pushout, denoted in the diagram as «PO». The formulation of such restrictions corresponds in the theory of sketches to the development of cones and cocones for the diagram and their introduction into the composition of the sketch [17].

A generalized diagram of the methodology developed by the author for the DKW design based on categorical-ontological models is shown in Fig. 7. Fig. 7 shows successively the stages of transformation of a generic ontological model into a categorical-ontological (CO) model, that is, SA model, verified on the basis of category theory. On the basis of this model, further design procedures are performed: joint information (in the form of an Entity-Relationship – ER diagram) and a datalogic (in the form of Functional Dependencies – FDs diagram) are developed. This joint approach allows to supplement and justify both diagrams using the provisions of the theory of relational database design. And also build on their basis a conceptual model (relational schema) of the database, close to optimal from the point of view of the absence of problems leading to data integrity violation during operation.

In addition, the CO model of SA allows directly obtain the knowledge base model for the projected SA, in the form of OWL DL axioms and SWRL rules [13]. Obtaining the conceptual DKW models based on the CO model of SA guarantees the completeness and correctness of the entities (classes) and relations of the PR that are represented. The general scheme of the methodology also includes the additional possibility of reduced models obtaining on the basis of the CO model for simulating the SA operation (system dynamic models, Petri nets, fuzzy cognitive maps). The technology of simulation using fuzzy cognitive maps is developed in [18].

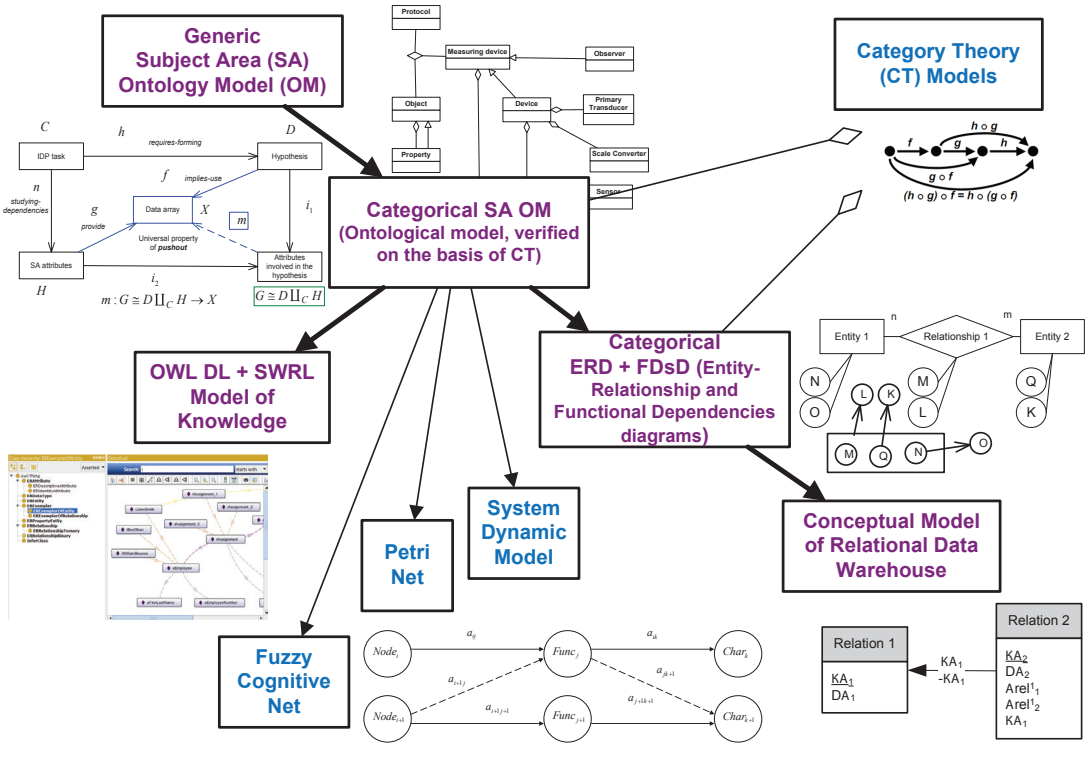

Fig. 7. Generalized scheme of the methodology for the design of data and knowledge warehouses for computer systems based on categorical-ontological models

ISSN 2226-3780

6. Research results

For the categorical-ontological representation of the developed methodology and the sequence of transformations performed on the data, a corresponding model is developed, which is shown in Fig. 8.

This model also represents the formalization of SA procedural models (business logic) during the CS for IDP design, the rules for the SA operation, the requirements and restrictions imposed on the DKW contents. The model reflects the following features of the implementation of the developed methodology. The categorical-ontological model COM^{Met} is the result of computing the pushout (PO) object:

$$COM^{Met} \cong Ont^{Met} \coprod_{K^{Met}} CTM^{Met} \coprod_{K^{Met}} FRC^{Met}, \quad (1)$$

which simulates the process of mapping knowledge about the SA K^{Met} into the COM^{Met} model by building a generic ontology Ont^{Met} verified on the basis of the category theory model CTM^{Met}, taking into account the SA requirements and constraints (business logic) FRC^{Met}. Also the model in Fig. 8 represents the process of obtaining a relational schema of data stores of a projected CS, close to optimal. This process is implemented on the basis of the joint use of the «entity-relationship» (ER) model and the model based on the concept of functional dependencies (FDs) EFM^{Met} and using the UML+OCL model UOM^{Met}. The data warehouse schema is derived from EFM^{Met} and UOM^{Met} on the COM^{Met} basis, calculating the corresponding PO object:

$$RS^{Met} \cong EFM^{Met} \coprod_{COM^{Met}} UOM^{Met}. \quad (2)$$

Other square faces of the cube on the right side of the model are not PO objects, since the model of the SA rules in the form of statements of first-order logic RM^{Met} is constructed not only on the basis of the model UOM^{Met}, but also directly on the basis of FRC^{Met}. RM^{Met} model is then used to build the KW, in particular, using rules in the language of SWRL or in the form of processing algorithms implemented in the Stored Procedure Language. This restriction is due to the fact that not all business logic of SA can be expressed by means of the language of Object Constraint Language in the model UOM^{Met}.

The result of applying the developed methodology is the construction of data warehouses and knowledge of the projected CS with a structure close to optimal by computing the coproduct object:

$$KW^{Met} \cong ODM^{Met} \coprod RI^{Met} \coprod RS^{Met}. \quad (3)$$

The result KW^{Met} is the final stage in the DKW construction for the information support of the CS operation for IDP, the structure of which is stored in them, the formalized ontological SA model and the business rules for data collection and processing represent the accumulated knowledge of the SA.

This result is obtained on the basis of the categorical-ontological model COM^{Met} as the nodal stage of modeling and design within the framework of the developed methodology.

7. SWOT analysis of research results

Strengths. The strengths of research are that, as a result of the development and application of this methodology, semantic and linguistic barriers have been overcome in the CS design. Customers, analysts and developers have an opportunity to use an end-to-end methodology, suitable for collective work and knowledge sharing, that is understandable and objective in terms of the provability and validity of models. This methodology can be used for a full range of research and design work, regardless of the formats of data presentation and diagrammatic techniques for formalizing the SA knowledge. Categorical-ontological approach to modeling and design makes it possible to formally substantiate subjective results of knowledge engineering and use objects of category theory in the form of design patterns at a high level of abstraction.

Weaknesses. Weak sides of research results are an increase in the requirements for the level of abstract thinking of analysts engaged in SA ontological modeling. When using the developed methodology, they must master the mathematical foundations of the theory of categories and sketches and the skills of using their positions when presenting the results of knowledge engineering. As with any method of formalization of knowledge, qualitative results can be obtained only on the basis of experience of long practical application of the developed methodology.

Opportunities. Additional opportunities that arise in achieving the aim of this study are solutions of the problems of intellectualization of computer systems, improve the quality of development and the effectiveness of using such systems based on the methods of knowledge engineering. This allows to develop productive CS for IDP, which perform the tasks of prompt and relevant data processing, effective knowledge enginee-

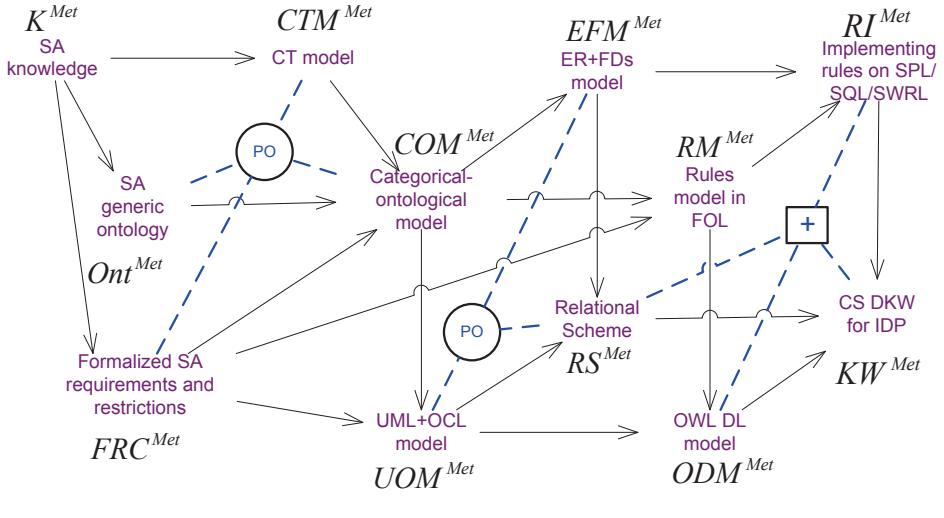

Fig. 8. Categorical-ontological representation of the developed methodology and sequence of transformations performed on data

ring and extraction of adequate models from data sets. Components of such CSs are implemented at a number of industrial enterprises, in scientific activity and educational process.

Threats. The difficulties in implementing the obtained results in the CS design are related to the following factor. Specialists who apply a technique or design methodology for a long time and receive satisfactory results may have corresponding biases to the developed methodology and doubt its practical utility. However, the increasing complexity of CS for IDP and the tasks solved with their help, increasing competition in the field of information technologies, require the development of new approaches to design based on modern achievements of scientific thought.

8. Conclusions

1. Comparative analysis and ontological modeling of approaches and techniques used in the design of data and knowledge warehouses in computer systems are performed. It is determined that the models created at different stages of design and using different techniques do not allow performing their mapping without losses. It requires the development and application of an end-to-end methodology suitable for collective work and knowledge sharing, integrating existing techniques and based on a unified mathematical apparatus.

2. A methodology is developed for designing data and knowledge warehouses for solving data processing and analysis problems based on categorical-ontological models. This methodology unites, in contrast to existing methods, design using information and data models with various diagrammatic techniques and modeling languages. This approach has helped to eliminate shortcomings and to supplement the merits of different design approaches and to obtain a storage structure close to optimal.

As a result of the development and application of this methodology, the semantic and linguistic barriers that arise between the members of the project team during the CS design have been overcome. Using a categorical-ontological approach to modeling and design makes it possible to formally substantiate the subjective results of knowledge engineering and use the objects of category theory in the form of design patterns at a high level of abstraction.

References

1. Data mining: practical machine learning tools and techniques. Ed. 2 / ed. by Witten I. H., Eibe F. Burlington: Morgan Kaufmann Publishers, 2005. 525 p.
2. Sahaida P. I. Modelirovaniye problemnoy oblasti komp'yuterizirovannykh informatsionnykh sistem dlya intellektual'noy obrabotki dannykh s ispol'zovaniyem inzhenerii znaniy // Naukovi pratsi DonNTU. Seriya: Obchislyuval'na tekhnika ta avtomatizatsiya. 2017. Vol. 1 (30). P. 78–87.
3. Palagin A. V., Kryvyy S. L., Petrenko N. G. Ontologicheskiye metody i sredstva obrabotki predmetnykh znaniy: monograph. Lugansk: ENU named after V. Dalya, 2012. 324 p.
4. Sahaida P. I. Ontologicheskiy podkhod k proyektirovaniyu baz dannykh informatsionnykh system: proceedings // Sovremennoye obrazovaniye i integratsionnyye protsessy. Kramatorsk: DGMA, 2012. P. 313–318.
5. Walter R. F. C. Categories and Computer Science. Cambridge: Cambridge Universities Press, 1991. 166 p.
6. Spivak D. I. Category theory for the sciences. MIT Press, 2014. 435 p.
7. Barr M. Models of sketches // Cashiers Topologie Geom. Differentielle. 1986. Vol. 27. P. 93–107.
8. Wells C. A generalization of the concept of sketch // Theoretical Computer Science. 1990. Vol. 70, No. 1. P. 159–178. doi:10.1016/0304-3975(90)90158-e
9. Date C. J. An Introduction to Database Systems. Ed. 8. Pearson, 2003. 1024 p.
10. Sahaida P. I. Kategorial'no-ontologicheskoye modelirovaniye intellektual'noy obrabotki dannykh dlya matematicheskogo obosnovaniya rezul'tatov inzhenerii znaniy // Vimiryuval'na ta obchislyuval'na tekhnika v tekhnologichnikh protsesakh. 2017. Vol. 4. P. 149–158.
11. Johnson M., Rosebrugh R., Wood R. J. Entity-relationship-attribute designs and sketches // Theory and Applications of Categories. 2002. Vol. 10, No. 3. P. 94–112.
12. About Anchor Modeling. URL: http://www.anchormodeling.com (Last accessed: 25.12.2017).
13. Ontology Management: Semantic Web, Semantic Web Services, and Business Applications / ed. by Hepp M. et al. Springer, 2007. 293 p.
14. Larman, C. Applying UML and Patterns: An Introduction to Object-Oriented Analysis and Design and Iterative Development. Addison Wesley Professional, 2004. 736 p.
15. ISO/IEC 19502:2005 (Information technology – Meta Object Facility (MOF). 2005. URL: http://webstore.iec.ch/preview/info_isoiec19502%7Bed1.0%7Den.pdf (Last accessed: 25.12.2017).
16. Wells C. Sketches: Outline with References. 2009. URL: http://www.cwru.edu/artsci/math/wells/pub/pdf/Sketch.pdf (Last accessed: 25.12.2017).
17. Wojtowicz R. L. A Categorical Approach to Knowledge Management. Computational Category Theory Workshop. National Institute of Standards and Technology. September 29, 2015. URL: http://www.bakermountain.org/talks/nist.pdf (Last accessed: 25.12.2017).
18. Sahaida P. I. Informatsionnaya tekhnologiya i programmno-metodicheskiy kompleks dlya modelirovaniya slozhnykh obiektov proyektirovaniya s ispol'zovaniyem nechetkikh kognitivnykh kart // Visnik Donbas'koi derzhavnoi mashinobudivnoi akademii. 2013. Vol. 2. P. 50–58.

РАЗРАБОТКА МЕТОДОЛОГИИ ПРОЕКТИРОВАНИЯ ХРАНИЛИЩ ДАННЫХ И ЗНАНИЙ КОМПЬЮТЕРНЫХ СИСТЕМ ДЛЯ ИНТЕЛЛЕКТУАЛЬНОЙ ОБРАБОТКИ ДАННЫХ

Разработана методология проектирования хранилищ данных и знаний для решения задач обработки и анализа данных на основе категориально-онтологических моделей. Методология объединяет проектирование с использованием различных диаграммных методик и языков моделирования. Это позволило устранить недостатки и дополнить достоинства различных подходов к проектированию и получить структуру хранилищ, близкую к оптимальной.

Ключевые слова: компьютерная система, хранилище данных и знаний, категориально-онтологическая модель.

Sahaida Pavlo, *PhD, Associate Professor, Department of Electronic Engineering, Donetsk National Technical University, Pokrovsk, Donetsk region, Ukraine, e-mail: pavlo.sahaida@gmail.com, ORCID: https://orcid.org/0000-0002-4700-8160*