CONSTRUCTING THE ANALYTICAL MODEL FOR SPECIALIZED MODEL-DRIVEN SYSTEM OF SCIENTIFIC DATA CONSOLIDATION

Anna V. Korobko, Alexey A. Korobko
Institute of Computational Modelling SB RAS, Krasnoyarsk, Russia

Abstract. Efficient storage and analytical processing of experimental results is a major part of obtaining new scientific knowledge. Specialized systems for consolidating scientific data allow you to record the results of your scientific research. Model-driven systems contain the control runtime model describing the object of study in detail. The paper proposes an algorithm for the automatic generation of an analytical model based on the control model of consolidating scientific data.

Keywords: analytical model, data consolidation and processing, research data, model-driven development.

Introduction

Of course, the main value of the conducted experimental and full-scale studies are the obtained results, the revealed consistencies and the detected interrelations. In other words, scientists are aimed at obtaining new knowledge about the object under study. In other words, scientists are aimed at obtaining new knowledge about the subject matter. There are several components of efficient processing of experimental data. Effective processing of experimental data is based on the quality of the presentation of the primary results of experiments. We need a detailed description of the subject of the study, the usage of common scales and units of measurement, common thesaurus and thoughtful storage. This is how we can get reliable scientific knowledge.

Scientists often use table editors to store and analyze their data [1]. Their main advantages are accessibility, popularity in the scientific community, standard statistical tools and data visualization. In addition, table editors have several significant drawbacks. As a rule, tools have limitations on the size of rows and columns. There is no input control, consistency and data integrity check. Storing information in files excludes multi-user and cross-platform data access. Data analysis often leads to their change or duplication. Comparative and retrospective data analysis is impossible during complex experiments [2]. These shortcomings make it difficult to use table editors systematically in long-term research.

We have developed a software platform for building systems for collecting and analyzing research data [3]. The development has been accomplished in response to a request from scientists. With the software platform user can create the model-driven systems for consolidating and analyzing. The platform helps the user to intuitively create and develop a system. It does not require any special IT skills from the user.

However, the analysis of the accumulated data of scientific research now entails involving an IT specialist and using a third-party software. It is desirable that the user could also select the data for analysis himself using the platform. Therefore, we must increase the availability of analytical data processing tools within the framework of the platform.

Current trends in the development of data analysis technology lie in reducing the requirements to the qualification of end-user [4]. One of the promising areas of "democratization" of analysis has

become developing means of native formation of analytical queries to data. The main idea is to build a transparent analytical model [5].

Uniqueness of the software platform for building systems for collecting and analyzing research data consists in formatting the control model during the creation of a specialized system. The control model describes a composition of the consolidated data and defines the interface of creating system. The control model is the thematic core of specialized system. It is stored as metadata and can be the basis for building an analytical model.

This article is intended to present an original approach to the formation of the analytical model for a specialized model-oriented system based on the control model. The first section is devoted to the concept of MDD (model-driven development) and formal description of the control model generated in the present software platform on the case of the System of research the state of the soil cover. The second section presents the theory of constructing an analytical model in accordance with the CWM specification. The third section contains an algorithm for formatting the analytical model of the specialized system for collecting research data based on its control model. In conclusion, the results of the article are summarized and the tasks for the continuation of this study are formulated.

The control model for specialized model-driven system of data consolidation. The classical model-driven approach (MDD) of software development involves the construction of a set of models of different levels of abstraction during the design and implementation of software [6]. The development of high-level abstraction models includes the processes of constructing a meta-meta model (M3) and a meta-model (M2). M3 is the model of the modeling language. M2 is the logical model of the subject area of application in the meta-meta model notation. The design of low-level abstraction models consists of the process of building application-level models (M1) and the stage of formation of instances of concepts (M0) defined at the M1 level [7]

The authors of this paper have previously proposed an original implementation of the approach to model-driven development. The main advantages of the proposed implementation in comparison with the classical approach are the formation of the control model (M1 level model) during the construction of the system through the platform and the automatic formation of application models based on the control model. This implementation helps to significantly reduce the requirements for the user's qualification in the field of information technology. This allows the user to focus on research. At the same time, the approach retains the flexibility and versatility of model-driven development. According to the original approach, the control model is included in every system in the form of metadata. This makes it possible to respond quickly to changing requirements for thematic content and permanently develop the systems in accordance with the growing research. The advantages of the proposed implementation of the model-based approach are achieved due to the deliberate narrowing of the platform functionality and through the development of its meta-model (M2 level), corresponding to the purpose of the systems. The functionality of the platform is focused on building systems for collecting and consolidating research data. As a meta-metamodel (level M3) used the notation of Unified Modeling Language (UML).

The metamodel contains three classes of objects: the class "Object" N, the class "Attribute" F and the class "Group" G. Objects of the class "Attribute" are described by the triple F=(A, T, D), where A is the attribute name, T is the name of the specialized attribute type, D is the attribute temporality flag. An example of a control model built in accordance with the proposed metamodel is shown in figure 1.

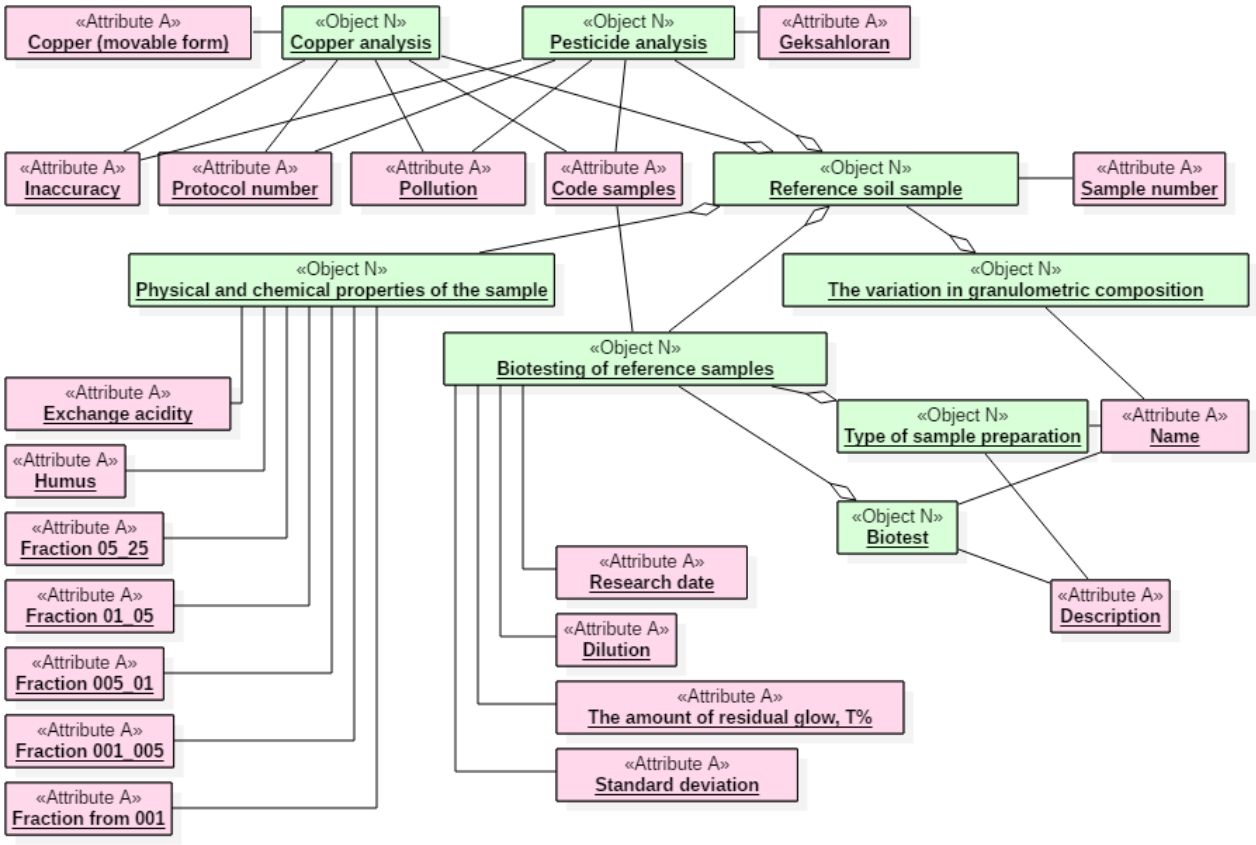

Fig. 1. Fragment of the control model of the System for research the state of the soil cover

The metamodel defines relationships between instances (objects) of model classes. Two types of relations, "Nesting" and "Dependence" are defined between objects. One-multivalued ratio "Nesting" – φ, is given on the set N, $\varphi \subseteq N \times N$ and is intended to set the organizational hierarchy of objects. One-valued relation "Dependence", denote it as χ, is given on the set N, $\chi \subseteq N \times N$. The relationship allows you to link objects to each other, implementing various functional interactions. A multi-valued correspondence between objects and attributes is "Ownership", denoted as θ, where $\theta \subseteq N \times F$. One-valued correspondence between objects and groups is – "Consolidation", denoted as ψ, where $\psi \subseteq N \times G$. The metamodel is described in more detail in [8].

The control model, formed under the proposed metamodel, formally describes both the subject of study and the results obtained in the course of scientific research. In work [3] the order of formation of the control model and requirements to its elements providing consistent storage of data and reliability of results of their analysis is offered.

The analytical model. The concept of model-driven development described above is widely used in many areas of information technology. In particular, the consortium Object Management Group (OMG) in 2003 proposed its own Common Warehouse Metamodel (CWM) to provide information exchange between heterogeneous systems for the purpose of analytical data processing [9].

The CWM specification is a set of M2-level models (according to the MDD approach) that provides a description of relational sources, XML documents, and the multidimensional analytical model (Fig. 2.). The analytical model is described in terms of On-Line Analytical Processing

(OLAP) technology and Data warehouse structure. The Data warehouse is designed to consolidate "operational" data from heterogeneous sources and prepare them for operational processing.

A key requirement of OLAP is presenting data in a multidimensional form that is intuitive to the user. Multidimensionality is the division of data into dimensions and measures. Dimensions are aspects of analysis. Measures are aggregated numerical characteristics of the process under study. The analytical model of the CWM specification is based on the dimensional fact model proposed by Mateo Golfarelli (et al.) in 1998 [10] (Fig. 2). The measures are used to describe a single analyzed fact must be grouped in a "Cube". For example, the measures "Exchange acidity", "Humus" and "Fraction of 0.01" characterize the fact "Physical and chemical properties of the soil sample". Dimensions have the hierarchical structure and combine several levels of analysis (Hierarchy). For example, the "Time" dimension includes hierarchy levels such as "Date", "Day of the week", "Month", "Quarter", "Year", "Season", and so on. Within the same schema, cubes and dimensions are linked by an association relationship ("CubeDimensionAssociation"). This takes into account the different levels of the hierarchy. The dimension values are designated as a separate class ("MemberSelection").

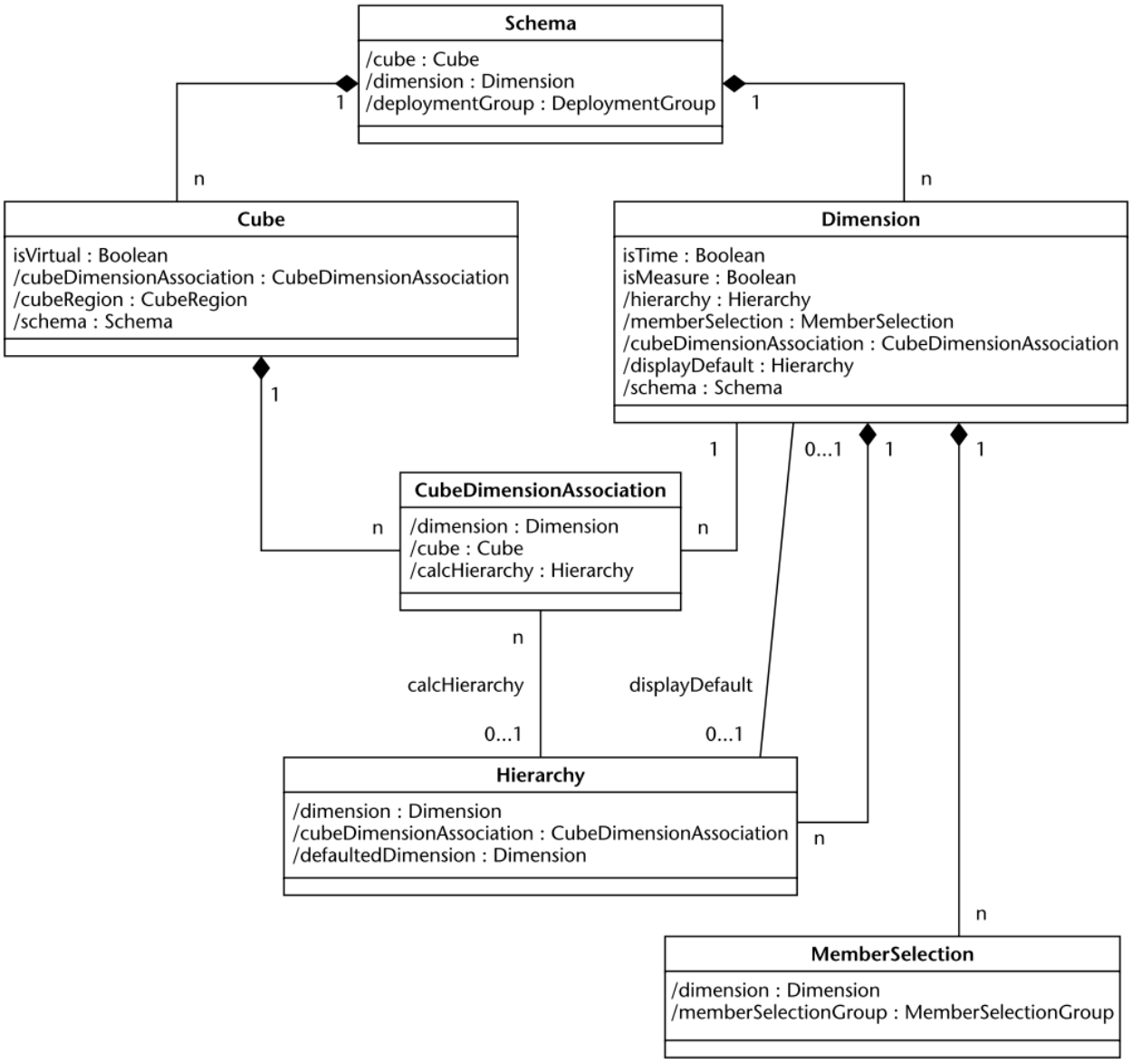

Fig. 2. The metamodel of the analytical model of CWM specification.

The construction of a multidimensional analytical model for a model-driven system provides the ability to process the results of analytical queries in BI-systems (Business Intelligence Systems) that supports information exchange in accordance with the CWM specification. The actual task is to develop an algorithm for formatting the multidimensional model based on the properties of the control model – an original platform metamodel.

Algorithm of formatting the analytical model. To describe the algorithm, we use the above formal description of the metamodel of the control model of the specialized model-driven system in terms of set theory.

The algorithm (Fig. 3) consists in sequential consideration of instances of the class "Object" of the control model ($\forall n \in N$) and their attributes ($\forall f \in F \mid f\theta n$). In the theory of multidimensional modeling, attributes are divided into "non-aggregated" (dimensional) and "aggregated" (non-dimensional). "Non-aggregated" attributes participate in the formation of the "dimension". And "aggregated" attributes of one object (or fields of one table, in the classical case) make up the "cube". Dimensional attributes can also be aggregated because an aggregate function "count" can be applied to them. The exception is dimensions that are functionally independent of other dimensions. In this case, the "isMeasure" property of the analytic model dimension is set to "false". In the algorithm, the standard check of attribute type is supplemented by the attribute name parsing procedure. Despite the numeric type, an attribute is not aggregated if its name contains marker words that indicates its "reference" character. The set of marker words can be expanded by an accumulation of precedents.

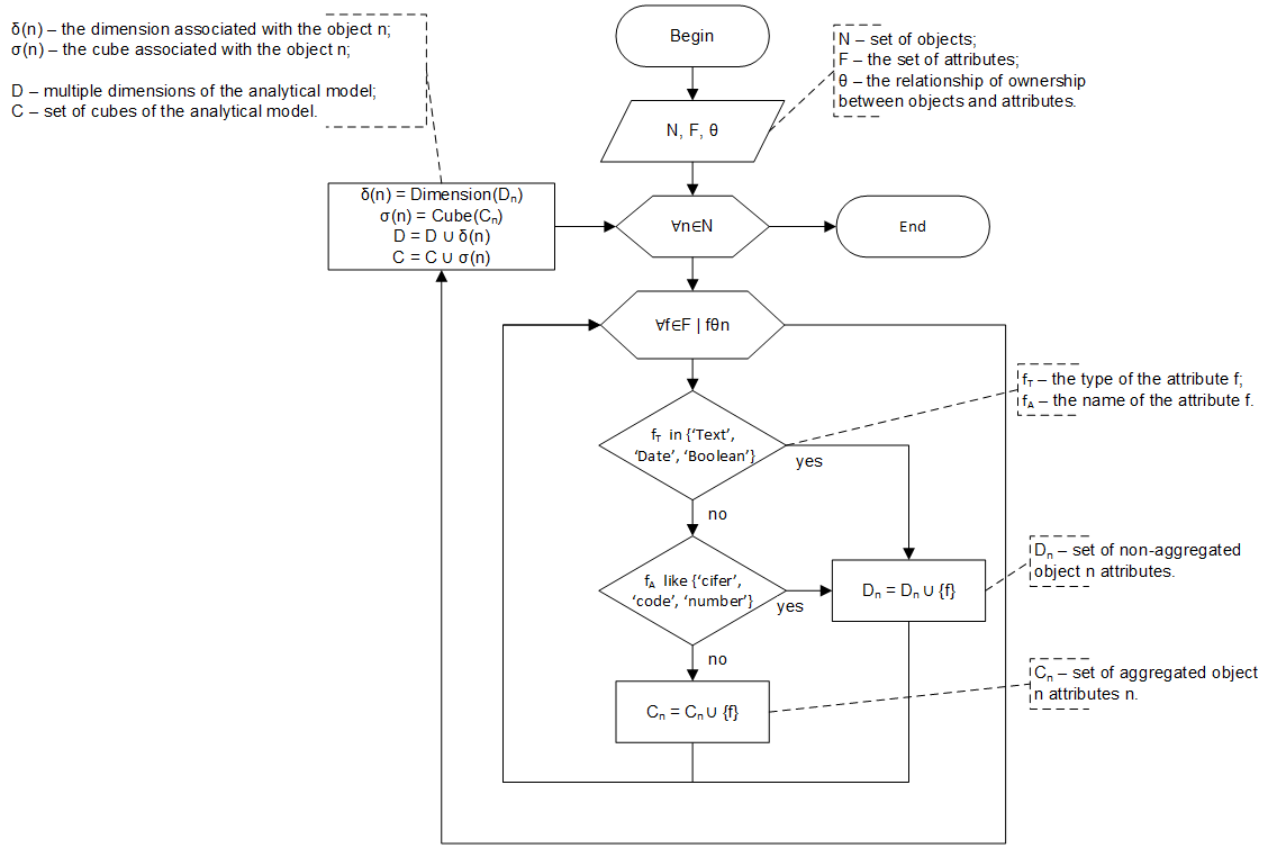

Fig. 3. The algorithm of formation of a set of cubes and dimensions of the analytical model

The "Cube" of each considered object is added to the set of cubes of the analytical model C. An associative relationship is established between the object n and the cube generated by it, written as a relation σ. Similarly, a set of dimensions D of the analytical model is formed and an associative relationship between the object and the dimension - δ is established. These relations are necessary for the operation of the algorithm of forming associations between cubes and dimensions (Fig. 4).

The relation χ describes the relationship between the objects of the control model. The expression nχk, where n,k∈N, means that object n has an attribute associated with object k. For example, the object "Biotesting of reference samples" is associated with the object "Biotest". This means that to set the values of the results of soil sample biotesting, it is necessary to select the specific type of "Biotest". In turn, "Biotest" is an independent object. The user can expand the set of its attributes and add new values if necessary.

When constructing the analytical model, the χ relation is treated as a foreign key or a relation of functional dependence. Associative links between cubes and dimensions are created in two cases. First, a σ(n) cube generated by some object n must be associated with a δ(n) dimension generated by the same object. Second, the σ(n) must be associated with all dimensions generated by objects directly related to the object n. As well as with dimensions generated by objects transitively related by the χ relation to n (by the elements of the set $\chi^*(n)$). The set $\chi^*(n)$ is a transitive closure of the relation χ for object n.

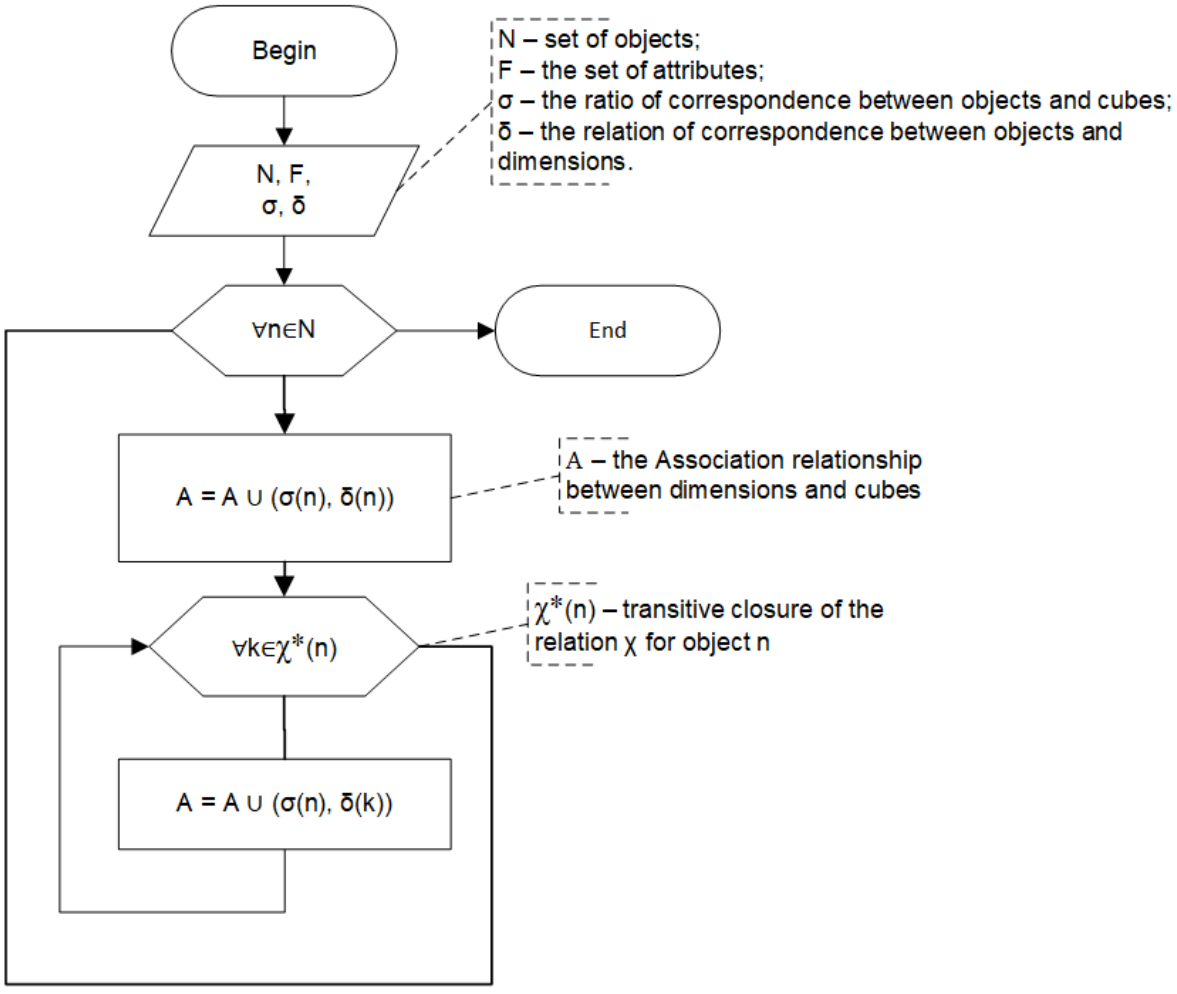

Fig. 4. Algorithm of formation of the relation of Association between dimensions and cubes of the analytical model

Sequential application of two algorithms allows you to create the analytical model consisting of dimensions, cubes and the relationship between them (CubeDimensionAssociation) – the A relation. In this paper, we do not consider the problem of constructing dimension hierarchies, leaving it for future research.

Conclusion. The software platform for building model-driven systems has been developed by the specialists of the ICM SB RAS to support the consolidation, storage and analytical processing of research data. The technological and methodological basis of the platform has allowed to create a System of research of the state of the soil cover by the biophysics-researchers themselves without the involvement of IT-specialists. The platform enables consistent storage of research data, analysis within the framework of long-term multi-stage research projects and comparison of research results carried out by different groups of scientists. In order to increase the availability of analytical processing, in software tools for the analytical querying, the task of developing the algorithm for the formation of the analytical model of the specialized system based on the control model has been set and solved. The presented algorithm is the theoretical basis for the development of analytical reports wizard for the specialized model-driven systems. The next step in creating a native tool for analyzing research data is an algorithm for generating SQL queries to the system database in accordance with user analytical queries in terms of the analytical model.

The study was carried out with the financial support of RFBR and the Government of Krasnoyarsk region, research project №18-47-240005.

LITERATURE

[1] K. W. Broman and K. H. Woo, “Data Organization in Spreadsheets,” Am. Stat., 2018.

[2] R. R. Panko, “What We Know About Spreadsheet Errors,” J. Organ. End User Comput., 2014.

[3] A. Korobko, A. Korobko, and E. Kolosova, “Constructing the model-driven system for scientific researches support on the original software platform for primary data consolidation,” in Surveying Geology & Mining Ecology Management (SGEM), 2018, vol. 18, no. 2.1, pp. 255–262.

[4] P. Alpar and M. Schulz, “Self-Service Business Intelligence,” Bus. Inf. Syst. Eng., vol. 58, no. 2, pp. 151–155, 2016.

[5] D. AnHai, H. Alon, and I. Zachary, Principles of Data Integration. Elsevier, 2012.

[6] E. Seidewitz, “What models mean,” IEEE Softw., vol. 20, no. 5, pp. 26–32, 2003.

[7] C. Atkinson and T. Kühne, “Model-driven development: A metamodeling foundation,” IEEE Softw., vol. 20, no. 5, pp. 36–41, 2003.

[8] A. A. Korobko, “Algorithm of Interface Generation for Model-Driven Data Consolidation System,” in 2018 3rd Russian-Pacific Conference on Computer Technology and Applications (RPC), 2018, pp. 1–4.

[9] L. Peyton, Common Warehouse Metamodel. 2016.

[10] M. Golfarelli, D. Maio, and S. Rizzi, “the Dimensional Fact Model: a Conceptual Model for Data Warehouses,” Int. J. Coop. Inf. Syst., vol. 07, no. 02n03, pp. 215–247, 1998.