A System for Computer-Assisted Design of Inorganic Compounds Based on Computer Training

N. N. Kiselyova[a], A. V. Stolyarenko[a], V. V. Ryazanov[b], O. V. Sen'ko[b], A. A. Dokukin[b], and V. V. Podbel'skii[c]

[a] *Baikov Institute of Metallurgy and Materials Science, Russian Academy of Sciences, Leninskii pr. 49, Moscow, 119991 Russia*

[b] *Dorodnicyn Computing Centre, Russian Academy of Sciences, ul. Vavilova 40, Moscow, 119333 Russia*

[c] *Higher School of Economics State University, ul. Myasnitskaya 20, Moscow, 101000 Russia*

e-mail: kis@imet.ac.ru, rvvccas@mail.ru, vpodbelskiy@hse.ru

Abstract—A system for computer-assisted design of inorganic compounds, with an integrated complex of databases for the properties of inorganic substances and materials, a subsystem for the analysis of data, based on computer training, a knowledge base, a predictions base, and a managing subsystem, has been developed. The methodology of integration of software products for data analysis, built upon different algorithms of computer training, has been devised. In many instances the employment of the developed system makes it possible to predict new inorganic compounds and estimate various properties of those without experimental synthesis.

Keywords: pattern recognition, information–analytical system, design of inorganic compounds.

DOI: 10.1134/S1054661811010081

1. INTRODUCTION

The prediction of the possibility of formation and properties of inorganic compounds, based solely on the information on the parameters of chemical elements composing those, is an important and complicated problem in chemistry. The theoretical methods built upon quantum-mechanical computations developed to date often do not make it possible to solve this problem, especially when the subjects are multicomponent substances in the solid state.

The limitations of conventional physical methods for computation of complex inorganic compounds were the reason for the development of a new approach combining chemistry and modern computer science, namely the computer-assisted design of inorganic substances and materials [1, 2]. The basic hypothesis on which this approach relies holds that *the fundamental properties of multicomponent inorganic substances under varying conditions (including temperature, pressure, relationships among the amounts of components, etc.) are related by periodic dependences to the fundamental properties of chemical elements composing the substances*. That there are such regularities is a consequence of Mendeleev's periodic law.

If a multicomponent chemical substance is regarded as a point in the multidimensional space of the properties of elements, then because of the periodicity of properties of these latter elements, points corresponding to combinations of chemical elements of a similar nature are bound to form compact classes. Let there be some set of chemical substances whose membership in different classes is known (a *training sample*). Here each substance is specified as a set of values of properties of elements. It is required to construct the hyperspace separating the substances in one class from the substances in other classes in the multidimensional space of properties of elements. Such classes can be, for instance, chemical systems with formation or nonformation of compounds of the specified composition or particular type of crystal structure under certain external conditions. It is assumed that, due to the periodicity of properties of elements, the resulting separating surfaces can be used to determine the status of as yet uninvestigated substances. This process of *prediction* demands solely knowledge of the properties of chemical elements composing the unexplored substance. Thus, the problem on the search for substances similar to those already investigated can be reduced to the classical problem in pattern recognition, computer training by precedents.

The instrumental basis for the design of new inorganic compounds is an information–analytical system (IAS) of our development, which integrates databases for the properties of inorganic substances and materials with software products for the analysis of data. The IAS automates the storage and modification of information, preparation of data for analysis, and prediction, as well as visualization and presentation of the results of data analysis.

Received September 27, 2010

2. THE STRUCTURE OF THE INFORMATION–ANALYTICAL SYSTEM

The IAS [3] comprises the integrated system of databases (DBs) for the properties of inorganic substances and materials, the subsystem for the analysis of data, the tasks base (the knowledge base), the predictions base, and the managing subsystem (Fig. 1).

2.1. The System of Databases for the Properties of Inorganic Substances and Materials

The system, developed by the Baikov Institute of Metallurgy and Materials Science, Russian Academy of Sciences (IMET RAS), is the source of information for the computer-aided analysis. At the moment it incorporates the following DBs.

(1) The DB "Phases," for the properties of inorganic compounds [2, 4, 5], which now contains information on more than 46 000 ternary compounds (i.e., compounds formed by three chemical elements) and more than 17 000 quaternary compounds, extracted from more than 25 000 publications.

(2) The DB "Elements," for the properties of chemical elements [5], which contains data for more than 90 parameters.

(3) The DB "Diagrams," for phase diagrams of semiconductor systems [2, 4, 6, 7], which contains information on phase diagrams of semiconductor systems and on the physical and chemical properties of phases forming in those, collected and evaluated by experts. At present this DB comprises detailed information on several tens of the systems that are the most important for semiconductor electronics.

(4) The DB "Crystal," for the properties of acousto-optical, electro-optical, and nonlinear optical substances [2, 4, 7], which now contains information on the parameters of more than 120 materials.

(5) The DB "Bandgap," for the forbidden band width of inorganic substances [4], which currently contains information on more than 3000 substances.

The total size of DBs is about 8 Gbytes. The integrated system of DBs allows specialists to gain aggregate information on the properties of substances and materials from different databases at one time. Authorized users can access to the system of DBs via the Internet (http://imet-db.ru).

2.2. The Subsystem for the Analysis of Data

When selecting the methods of pattern recognition for the analysis of chemical information, we took into account many years of experience in application of these methods to the design of inorganic compounds [2, 8]. As a result, the following methods and software products were adopted.

—A wide range of algorithms from the system "Recognition" developed by the Computing Centre, Russian Academy of Sciences [9]. In addition to the well-known methods of linear machines, Fisher linear discriminant, k nearest neighbors, support vector machine, and neural-network algorithms, this multifunction system for pattern recognition includes algorithms developed by the Computing Centre, Russian Academy of Sciences, that is to say, recognition algorithms based on calculation of estimates, voting algorithms based on deadlock tests, voting algorithms based on logical regularities, weighted statistical voting algorithms, etc.

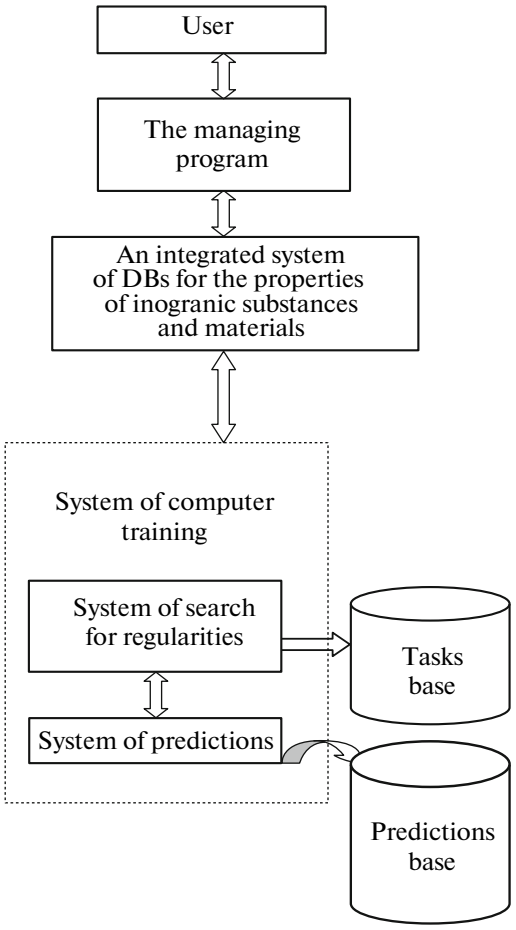


Fig. 1. Diagram of the IAS for the design of inorganic compounds.

—The ConFor system for training a computer in the procedure for formation of concepts, developed by the Institute of Cybernetics, National Academy of Sciences of Ukraine [10]. The system is built upon the arrangement of data in the memory of a computer in the form of growing pyramidal networks.

The software products for data analysis, listed above, were chosen for the following reasons.

• Their universality as regards the dimensions of problems to be solved. The systems make it possible to solve both problems on prediction of rare or unique events, occurrences, or processes, when the initial

(training) information is scarce (tens of precedents), and problems are of great dimensions (tens of thousands of precedents).

• Their universality as regards the data type (numeral, binary, and nominal attributes are allowed).

• Their capability of processing incomplete and partially inconsistent information with unknown or approximate values of some attributes.

The application of a large ensemble of computer training procedures yields several predictions for compounds that have not been synthesized yet. Sometimes the results are not the same, and the question arises of making a definitive decision on the status of the subject of prediction. A promising way to solve this last problem is to use recognition by ensembles of algorithms [9]. In many cases the synthesis of a collective solution affords the compensation of possible recognition errors in individual algorithms by the correct results of other algorithms. Therefore, the procedures implementing diverse strategies of making collective decisions, based on the Bayes method, methods using clustering and selection, decision templates, logical correction, the method of a convex stabilizer, the Woods dynamic method, committee methods, etc., are included in the IAS [9]. In addition, for the first time in the practice of design of inorganic compounds, the IAS embodies the possibility to make collective decisions on different sets of attributes with the use of methods of decision templates and committee methods (majority voting and averaging).

In order to select informative properties of chemical elements, we included into the IAS a procedure built upon minimization of generalized error functionals for convex correcting procedures with respect to ensembles of predictors constructed on the basis of individual attributes [11, 12]. It has been shown that the problem of search for the optimum ensemble reduces to a well-known problem in quadratic programming. The method of optimization is based on the use of necessary and sufficient conditions for the ensembles of predictors to be irreducible.

The selection of the properties of chemical elements, providing the most information for the classification of substances, is of double significance. On the one hand, it enables a drastic reduction of attribute description that includes hundreds of elements' properties for multicomponent substances. On the other hand, the selection of the most important properties of elements in classification of chemical substances affords physical interpretation of the resulting classifying regularities, which improves the confidence in the obtained predictions and makes it possible to find substantial causal relationships among the parameters of subjects and to develop physical and chemical models of the phenomena.

2.3 The Tasks Base

The regularities discovered by an expert as a result of manipulation with the IAS are stored in the tasks base in the intrinsic format of those software products for the analysis of data by whose means they were obtained. Such implementation makes it possible to integrate new software products for the analysis of data into the IAS and resolves the problem associated with the fact that the forms of representation of the resulting regularities in the computer training methods used are substantially different. By a task is meant the procedure of training by the selected methods on a particular training sample. Here it is suggested that not the results of training, as such (like logical expressions or the structure of a trained neural network), but so-called tags for the tasks be stored in the tasks base. The term label is taken to mean the necessary information for the task, which permits identifying it from others. Because different tasks use training information of varying quality (for instance, with different extent of reliability and completeness), a way to support expert judgments of the obtained regularities and predictions has been developed. The following information on the task is stored in the IAS: the unique number of the task; the training sample in standard format; data for the attributes used to form the training sample; the identifier of the software product for the analysis of data by whose means the regularities were obtained; the list of methods employed in training, with their parameters; information on the quantitative and qualitative composition of the compounds used in training; the identifier of the compounds' parameter to be predicted; and expert judgment of the regularity and predictions tasks base.

2.4. The Predictions Base

The predictions base contains the results of previous computer experiments, as well as references to service information stored in the knowledge base. The use of the predictions base made it possible to improve the functionality of the IMET RAS DB for the properties of inorganic substances and materials through providing the user not only with the available information on substances that have already been studied, but also with predictions of inorganic compounds which have not yet been obtained and estimations of the properties of those. At present this base is being filled.

2.5. The Managing Subsystem

The managing subsystem organizes the process of computation and carries out interaction among functional subsystems of the IAS; it also affords access to the system from the Internet. In addition to that, the managing subsystem provides the user with software to prepare data for analysis, to produce reports, to visualize the results, and to utilize other service functions. In particular, a special subsystem has been designed to

retrieve from the DB information that, after its estimation by expert, is used to train the computer. The subsystem allows the expert to edit found information and to form an attribute description of compound, which is a complex description made up of parameters of three or four chemical elements included in its composition. The expert selects the properties of chemical elements to form a training sample, and the subsystem for preparation of training sample retrieves the chosen values of the elements' properties from the DB "Elements," makes up complex attributes as algebraic functions of the initial parameters of elements when needed, and merges the attribute description to produce a table that is thereupon passed to the input of the prediction subsystem. The subsystem for generation of results is intended for the presentation of predictions in the tabular form customary among chemists and material scientists.

In computer training the inorganic substances are represented in the computer memory as sets of values of chemical element properties. Among the difficulties associated with the use of the IAS in inorganic chemistry is the fact that there are gaps in data for the properties of chemical elements. These gaps often impair the accuracy of training and recognition. This effect is especially noticeable for small samples. In the context of the managing subsystem, the following approach to the solution of this problem is proposed. First, properties with more than 25% of missing values are removed from a particular sample, since they are hardly informative and may hinder correct training of the system. Second, the remaining empty values are filled in as follows. Chemical elements which are the nearest, in values of other properties, to the one whose gap is to be filled are sought with regard for the peculiarities of the subject domain and for the way training samples are prepared. Next, the average value of the involved property over the nearest elements is computed. Here, the relative Euclidean distance between elements need be no greater than 10%, and the chemical element is sought only among the elements in the same group of the periodic system. If no appropriate element is found, then either the empty value is replaced by the mean value of the element's property for the substances with equal classifying feature (in the case of the training sample), or this attribute is excluded from the sample (in the case of a sample for recognition), and the system is retrained again without this property, i.e., a sample resulting from the elimination of this parameter of the element from the initial training sample is passed to the input of the IAS.

The training and recognition processes are implemented in the IAS by means of a special asynchronous web service, which makes it possible to accomplish long-term training and prediction tasks on the Internet. The asynchronous web service allows users to initiate long-term execution of resource-consuming instructions, to monitor progress of that, to receive notifications of ready results of computations, and to terminate execution of tasks, saving the intermediate results.

3. PRINCIPLES OF INTEGRATION OF DBS WITH SOFTWARE PRODUCTS FOR THE ANALYSIS OF DATA

The IAS for computer-assisted design of inorganic compounds is built upon the available databases and software products for the analysis of data. This brought up the problem of integration of diverse software and information components. The solution of the problem was complicated by the fact that the source of information in the IAS is a system of databases established at different times and founded on different database managing systems (DBMSs). In addition, there was a need to include in the IAS means of data processing, distinct in concepts and developed by diverse organizations and in different countries.

When developing the principles of integration of databases for the properties of inorganic substances and materials [2, 4–7, 13], we took into account the specific character of the subject domain, that the information is stored in databases with a varying extent of reliability, and different operation systems, data formats, and DBMSs are used. In connection with the above-mentioned specificity, a complex approach to integration, combining data level and user interface integration, is applied [4, 13].

The need for integration of computer training software products is associated with the fact that in order to improve quality of prediction the IAS employs special collective decision-making methods [9] whose operation involves interaction among software products for the analysis of data [9, 10] with distinct principles of operation.

We applied the service-oriented approach (SOA) [14] to solve the problem on integration. The SOA is an applied architecture where all functions are specified as independent services with well-defined interfaces. A certain sequence of requests of these services affords the execution of a particular process. According to the SOA concepts, the interaction between subsystems for the analysis of data, implementing all training and recognition methods, and the managing subsystem is performed by means of programmed adapters which present all necessary functions of the software product for the analysis of data. Integration of a new computer training software product into the IAS requires only a programmed adapter that would incorporate intrinsic data structures of the system to be integrated in the IAS with standard representation of data.

The architectural concept used in service-oriented integration is noteworthy. The case in point is the concept of a service bus. The task is to provide a unified way of transmitting queries and receiving the results of the service, performing necessary conversions of messages and transport protocols, and

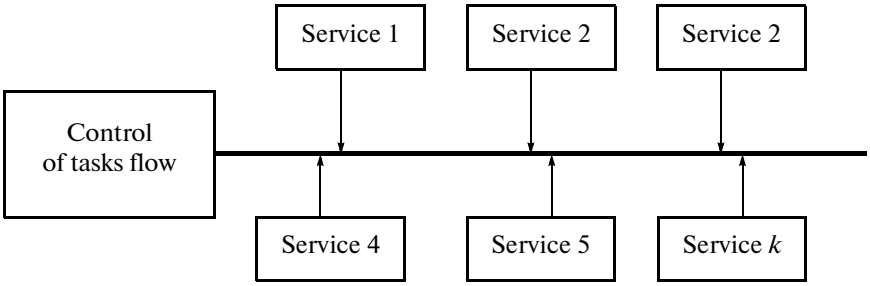


Fig. 2. Model service bus.

most importantly, managing the flow of service requests. Such management makes it simpler to arrange the sequence of calls of services required to execute a process. Looking at the diagram of the bus (Fig. 2), one can see that this approach solves one of the main problems in integration, the problem on minimization of interfaces.

Note that, irrespective of the selected technique of module integration, the development of special programmed adapters which would incorporate intrinsic data structures of the software product to be integrated into the system with the standard representation of data in the system is always required. These adapters operate with intrinsic representation of data in an individual application. In order to make it more simple to include new software products for the analysis of data into the integrated IAS, the requirements on the implementation of programmed adapters for software products to be integrated in the system have been developed and an XML-based format of message exchange between these products and the managing subsystem of the IAS has been proposed.

All computation tasks are performed by the web server, only their results are output to the user to be viewed. Such organization allows the IAS to be easily augmented by integration with new methods for the analysis of data, and makes it possible to add extra functionality with no required update of client applications. Because the IAS is furnished with a web interface, the users can carry out the analysis of data via the Internet.

The proposed means of solving the problem on integration of software products

• make it possible to add software products for the analysis of data to the IAS stage by stage;

• are sufficiently simple to implement, since the development of standardized programmed adapters intended for including new software product for the analysis of data into the IAS on the basis of the proposed procedure is not a difficult problem;

• allow for distinctions in data and information structures used by software products;

• provide complex ways of interaction among software products for the analysis of data.

CONCLUSIONS

The developed information–analytical system is widely applied in design of new inorganic compounds. Its use allows one to predict new inorganic compounds and to estimate various properties of those without experimental synthesis of the compounds. With IAS it proved to be possible to predict hundreds of as yet unobtained intermetallic compounds with a crystal structure type of the Heusler phases [8, 15], which are promising in the search for new materials to be used in magnetic storage devices of great capacity, and compounds with the compositions AB_3X_3 [16] and ABX_2 [17], which are of interest in the development of new semiconductor, nonlinear optical, electro-optical, and acousto-optical materials, as well as to estimate some properties of the predicted compounds (the melting point and the forbidden band width), important for application of those in actual practice [18].

ACKNOWLEDGMENTS

We are grateful to V.A. Dudarev for helpful remarks.

This study was supported by the Russian Foundation for Basic Research, project nos. 06-07-89120, 08-01-90427, 08-07-00437, 05-03-39009, and 09-07-00194.

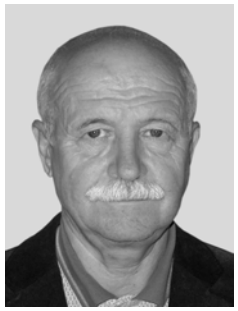
REFERENCES

1. E. M. Savitskii and V. B. Gribulya, *Application of Computer Techniques in the Prediction of Inorganic Compounds* (Nauka, Moscow, 1977; Oxonian Press, New Delhi, 1985).
2. N. N. Kiselyova, *Computer Aided Design of Inorganic Compounds: Application of Databases and Artificial Intelligence Methods* (Nauka, Moscow, 2005) [in Russian].
3. N. Kiselyova, A. Stolyarenko, V. Ryazanov, and V. Podbel'skii, "Information–Analytic System for Design of

New Inorganic Compounds," Int. J. "Inf. Theor. Appl." **2** (4), 345–350 (2008).

4. N. N. Kiselyova, V. A. Dudarev, and V. S. Zemskov, "Computer Information Resources in Inorganic Chemistry and Material Science," Usp. Khim. **79** (2), 162–188 (2010) [Russ. Chem. Rev. **79** (2), 145–166 (2010)].
5. N. Kiselyova, D. Murat, A. Stolyarenko, V. Dudarev, V. Podbel'skii, and V. Zemskov, ""Phases" Data base on Ternary Inorganic Compounds Properties in the Internet," Inf. Resursy Rossii, No. 4, 21–23 (2006).
6. Yu. I. Khristoforov, V. V. Khorbenko, N. N. Kiselyova, V. V. Podbel'skii, I. N. Belokurova, and V. S. Zemskov, "Data Base on Phase Diagrams of the Semiconductor Systems Available for Internet," Izv. Vyssh. Uchebn. Zaved. Mater. Elektron. Tekhn., No. 4, 50–55 (2001).
7. N. N. Kiselyova, I. V. Prokoshev, V. A. Dudarev, V. V. Khorbenko, I. N. Belokurova, V. V. Podbel'skii, and V. S. Zemskov, "Internet-Accessible Electronic Materials Database System," Neorg. Mater. **40** (3), 380–384 (2004) [Inorg. Mater. **40** (3), 321–325 (2004)].
8. G. S. Burkhanov and N.N. Kiselyova, "Prediction of Intermetallic Compounds," Usp. Khim. **78** (6), 615–634 (2009) [Russ. Chem. Rev. **78** (6), 569–587 (2009)].
9. Yu. I. Zhuravlev, V. V. Ryazanov, and O. V. Sen'ko, *RECOGNITION: Mathematical Methods, Software System, Practical Applications* (FAZIS, Moscow, 2006) [in Russian].
10. V. P. Gladun, *Processes of New Knowledge Formation* (SD Pedagog 6, Sofia, 1995).
11. O.V. Senko, "An Optimal Ensemble of Predictors in Convex Correcting Procedures," Pattern Recognition and Image Analysis, MAIK "Nauka/Interperiodica" **19** (3), 465–468 (2009).
12. O.V. Senko, A.V. Kuznetsova, "Methods of regularities searching based on optimal partitioning. Classification, forecasting, Data Mining," Supplement to International Journal "Information Technologies and knowledge" (ITHEA, Sofia, 2009), Vol. 3, pp. 136–141.
13. N. Kiselyova, S. Iwata, V. Dudarev, I. Prokoshev, V. Khorbenko, V. Zemskov, "Integration principles of Russian and Japanese databases on inorganic materials," Int. J. "Information Technologies and Knowledge" **2** (4), 366–372 (2008).
14. N. Bieberstein, R.G. Laird, K. Jones, T. Mitra, "Executing SOA: A Practical Guide for the Service-Oriented Architect," (IBM Press, Boston, 2008).
15. N. N. Kiselyova, V. V. Ryazanov, and O. V. Sen'ko, "Prediction of the Structure of ABX_2 (X = Fe, Co, or Ni) Intermetallics," Metally, No. 6, 98–104 (2009) [Russ. Metallurgy (Metally), No. 6, 538–545 (2009)].
16. N. N. Kiselyova, "Prediction of Occurrence of AB_3X_3 (X = S, Se, Te)," Neorg. Mater. **45** (10), 1157–1160 (2009) [Inorg. Mater. **45** (10), 1077–1080 (2009)].
17. N. N. Kiselyova, V. V. Podbel'skii, V. V. Ryazanov, and A. V. Stolyarenko, "Computer-Aided Design of New Inorganic Compounds with Composition ABX_2 (X = S, Se, or Te)," Materialoved, No. 12, 34–41 (2008) [Inorganic Materials: Applied Research, **1** (1), 9–16 (2010)].
18. N. N. Kiselyova, A. V. Stolyarenko, T. Gu, W. Lu, A. Blansche, V. V. Ryazanov, and O. V. Senko, "Computer-Aided Design of New Inorganic Compounds Promising for Search for Electronic Materials," in *Proc. 6th Int. Conf. on Computer–Aided Design of Discrete Devices (CAD DD'07)* (UIPI NASB, Minsk, 2007), Vol. 1, pp. 236–242.

Nadezhda Nikolaevna Kiselyova. Born 1949. Graduated from the Faculty of Chemistry of Moscow State University in 1971 and completed post-graduate study at the same faculty in 1974. Received candidate's degree in 1975 and doctoral degree in 2004. Scientific interests: computer-assisted design of inorganic compounds, databases for the properties of inorganic substances and materials, and materials used in electronics. Author of more than 130 articles and 2 monographs. Doctor of Chemical Sciences and Head of the Laboratory of Semiconductor Materials in the Baikov Institute of Metallurgy and Materials Science, Russian Academy of Sciences.



Vadim Valerievich Podbel'skii. Born 1937. Graduated from the Moscow Engineering Physics Institute with a degree in "computers." Received candidate's degree in 1973 and doctoral degree in 1989. Scientific interests: algorithms for automation of design and mathematical support and software for automated systems. Author of more than 120 articles and 9 handbooks. Doctor of Technical Sciences and Professor of the Chair of Software Development Control at the Department for Software Engineering of the Higher School of Economics State University.

Andrei Vladislavovich Stolyarenko. Born 1982. Graduated from the Faculty of Applied Mathematics of Moscow State Institute for Electronics and Mathematics in 2005. Received candidate's degree in 2008. Scientific interests: databases and software engineering. Author of more than 20 articles. Candidate of Technical Sciences and Researcher of the Laboratory of Semiconductor Materials at the Baikov Institute of Metallurgy and Materials Science of the Russian Academy of Sciences.

Aleksandr Aleksandrovich Dokukin. Born 1980. Graduated with honors from the Faculty of Computational Mathematics and Cybernetics of Moscow State University in 2002 and completed postgraduate study at the same faculty in 2005. Received candidate's degree in 2008. Scientific interests: algebraic theory of recognition and algorithms for calculation of estimates. Author of more than 30 articles. Since 2000 and to the present day is a Researcher at the Computing Centre, Russian Academy of Sciences.

Vladimir Vasil'evich Ryazanov. Born 1950. Graduated from the Moscow Institute of Physics and Technology in 1973. Received candidates degree in 1979 and doctoral degree in 1994. Academician of the Russian Academy of Natural Sciences, Professor. Since 1976 has been with the Dorodnicyn Computing Center, Russian Academy of Sciences. Currently is Head of the Department of Mathematical Problems of Recognition and Methods of Combinatorial Analysis. Scientific interests: recognition theory, cluster analysis, data analysis, optimization of recognition models, and applied systems of analysis and prediction.

Oleg Valentinovich Sen'ko. Born in 1957. Graduated from the Moscow Institute of Physics and Technology in 1981. Received candidates degree in 1990 and doctoral degree in 2007. Currently is a senior researcher at the Dorodnicyn Computing Center, Russian Academy of Sciences. Scentific interests: data mining, mathematical models of pattern recognition, classification and forecasting, practical applications in medicine and other fields.