PRINCIPLES OF INTEGRATION OF RUSSIAN AND JAPANESE DATABASES ON INORGANIC MATERIALS

Nadezhda Kiselyova, Shuichi Iwata, Victor Dudarev, Ilya Prokoshev, Valentin Khorbenko, Victor Zemskov

Abstract: The methods and software for integration of databases (DBs) on inorganic material and substance properties were developed. The integration of information systems is based on combination of known approaches: EII (Enterprise Information Integration) and EAI (Enterprise Application Integration). The metabase - special database that stores data on contents of integrated DBs is a kernel of integrated system. Proposed methods were applied for integrated system of DBs creation in the field of inorganic chemistry and materials science. Important feature of developed integrated system is ability to include DBs that were created by means of different DBMS using essentially various computer platforms: Sun (DB "Diagram") and Intel (other DBs) and diverse operating systems: Sun Solaris (DB "Diagram") and Microsoft Windows Server (other DBs).

Keywords: Integration of databases, metabase, distributed information system, inorganic substances and materials, EII, EAI.

ACM Classification Keywords: H.2.4 Distributed databases, H.2.8 Scientific databases, J.2 Chemistry.

Introduction

At present rich variety of databases on properties of inorganic substances and materials were developed and maintained in the world [Bale and Eriksson, 1990; Dudarev et al., 2006; Eriguchi and Shimura, 1990; Khristoforov et al., 2001; Kiselyova, 2005; Kiselyova et al., 2004, 2005, 2006; Villars et al., 2004; Xu et al., 2006; Zemskov et al., 1998]. Traditional areas, that DBs cover, are thermodynamic, thermo-chemical, crystallographic and crystal chemical properties. The majority of large industrial corporations support development of DBs that contain the information on physical, technical and technological parameters of materials and substances. The development tendencies of modern DBs on properties of inorganic substances and materials are following:

1. Internet-access to the information.

2. The use of powerful DBMS: Microsoft SQL Server, Oracle, IBM DB2, etc.

3. Great attention has been concentrated on the quality (reliability) of stored information. Highly skilled specialists are engaged in development process of the most "advanced" commercial information systems for data capture and expert estimation of data reliability. So users receive not simply "row" information but recommended values passed filtration for elimination of misprints.

4. Often DBs are supplemented with information analysis tools: from traditional thermodynamic calculations and statistical procedures up to modern means for regularities search in the data allowing predicting behavior of objects and making decisions. In the last case usual DBs, oriented to transaction processing, are often supplemented, for example, with special integrated information systems, that are named in the English literature as *Data Warehouse* [Kimball and Caserta, 2004]. They are intended for data coordination and integration from various information sources and its preparation for the subsequent computer analysis.

5. Integration of DBs on substances and materials. In this case the user can find the most complete cumulative information on properties of a certain substance.

The last problem information integration of resources on inorganic substance and material properties is the most important today. The data about various properties of a certain substance or material are distributed among different heterogeneous DBs. The chemist or material scientist has to look through a great number of DBs in order to find the necessary information. Therefore some superstructure above DBs, that will allow to output some cumulative – the integrated information on all set of properties of a substance stored in different information systems, is required. That is, an integration of DBs is necessary. The solution of this problem is concerned with several difficulties. Databases on inorganic substance and material properties were developed in various organizations and countries and thus they use different database management and operating systems. Taking into consideration differences in data quality, data expertise procedures, data formats, languages and many other troubles it should be stated that full and smooth integration of information resources is practically impossible problem. We developed the approach to DBs integration taking into consideration the peculiarities of DBs on inorganic substance and material properties. The approach can be used for integration of Russian and Japanese DBs in this knowledge domain.

Known Approaches to Database Integration

Principally there are three approaches to database integration [Imhoff, 2005]:

1) Data Warehouse based on ETL (Extract, Transform, Load) paradigm – technology [Kimball and Caserta, 2004].

2) EII (Enterprise Information Integration) – technology [Morgenthal, 2005].

3) EAI (Enterprise Application Integration) – technology [Morgenthal, 2000].

These approaches can be used for solution of wide set of problems: from real-time integration to batch integration and from data integration to applications integration. Fig. 1 illustrates these approaches area of application in relation to the different task types [Imhoff, 2005]. The EII technology is the best approach for real-time data integration. The ETL technology allows the best batch data integration. The EAI technology gives the best results at applications integration in real-time or batch modes.

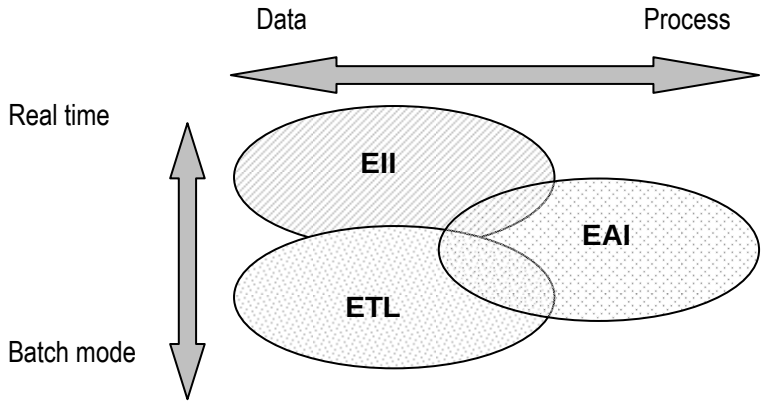

Fig. 1. Modern approaches for information systems integration.

The ETL-technology implies full merging of existing resources (fig. 2). That is the case when database complex is a single information system (*megabase*) for end users, operators and administrators. This approach is also known as Data Warehouse [Imhoff, 2005; Kimball and Caserta, 2004]. So at first information is extracted from DBs to be integrated. Then these data are somehow processed for clearing (that is, check for discrepancies and elimination of obviously false data) and transformations – series of special procedures that allow to get a common

unified format and dimension. Only after these stages cleared and unified data are input into data warehouse or megabase. Database exploitation costs reduction and information duplication reduction can be mentioned among advantages of this integration approach.

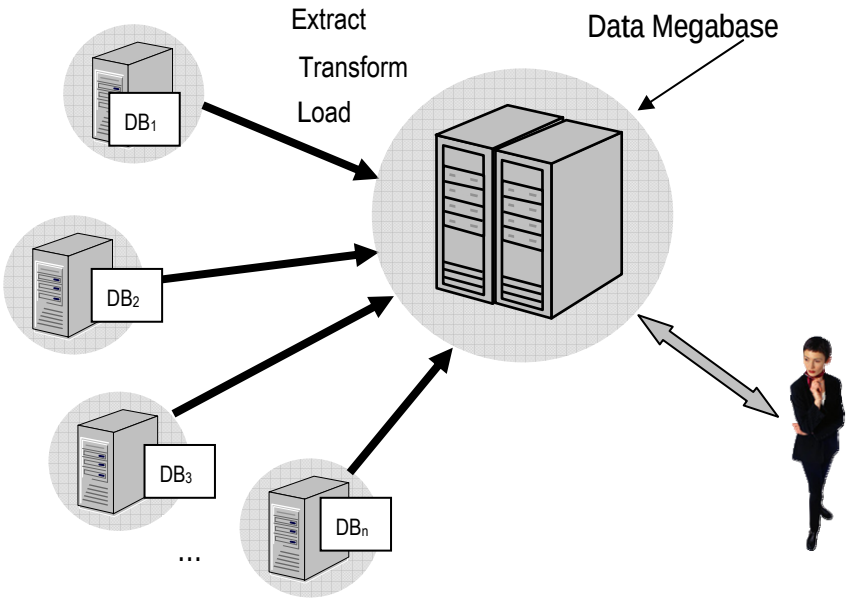

Fig. 2. ETL-approach – full merging of existing DBs.

The second integration approach is based on EII-technology (fig. 3). It is not going to integrate databases themselves [Imhoff, 2005; Morgenthal, 2005]. Integrated data are not transferred into a central megabase but remain in the same information systems, as before. Instead the program interface for data access is developed that allows retrieving required data. EII is data integration means from multiple systems into a unified, consistent and accurate representation format geared toward the data browsing. So the data are aggregated, restructured and relabeled (if it is necessary) and presented to a user. Usually the result of this approach is a virtually integrated heterogeneous distributed information system.

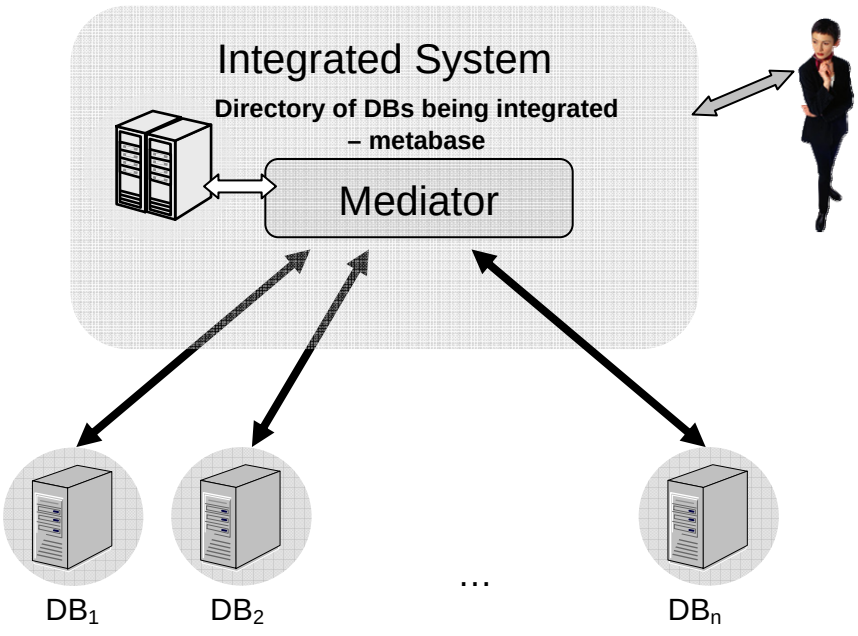

Fig. 3. EII-approach – data integration in real time.

The third approach – EAI – (fig. 4) is aimed for applications integration [Imhoff, 2005; Morgenthal, 2000]. Integration can be carried out in a batch mode or in real-time mode. Combined work of two and more applications can be achieved using this approach. This approach is based on message exchange between several applications. Frequently such information exchange is carried out through some common message exchange infrastructure known as message bus. Applications are connected to this common message bus by means of special adapters.

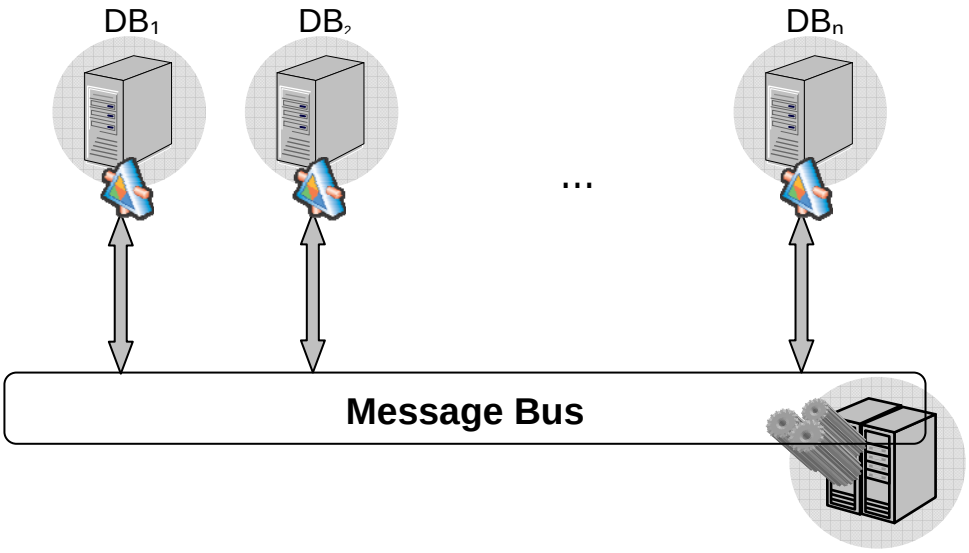

Fig.4. EAI-approach – integration of applications.

The EII and EAI technologies allow not to change every integrated database structure dramatically (and thus established database administration technology). So called “virtual” database integration and heterogeneous distributed information system creation implies an independence in evolution of separate subsystems and at the same time allows to end user to get access to the whole “live” data array on a certain chemical substance or material that are stored in databases of virtually united system.

So EAI technology integrates transactions of two or more applications, ETL technology merges the data of several information sources into a single one, and EII technology carries out virtual data integration from various information sources. It should be mentioned that no one approach can solve all tasks arising when integrating information systems on material and substance properties.

It is necessary to take into consideration that every data center on materials properties is a point of information concentration and data analytical processing bases on different software and hardware. The technology of information accumulating and data processing has been settled down in each organization. So great investments that were made in hardware and software do not allow mechanically transporting all the data into some centralized database. Moreover many DBs on material and substance properties are equipped with ancillary programs for substance parameters calculation. Therefore taking into consideration current development conditions of databases on inorganic substance and material properties the integrated system based on both EAI- and EII-technologies was developed in Baikov Institute [Dudarev et al., 2006; Kornuyshko and Dudarev, 2006] (fig. 5). It allows dynamically integrating a plenty of heterogeneous databases that are supplied with any computational subsystems.

Integration of Russian Databases on Inorganic Material and Substance Properties

From the beginning the proposed approach was used for integration of Russian DBs on inorganic material and substance properties. The successful task decision of integration needs is some coordinating center, which "knows" what information is stored in every integrated DB. Such function can be carried out by *metabase* – a special metadata database that stores information on integrated DBs contents, namely, about chemical systems, substances and its modifications. Chemical systems are identified by a set of, which are included into its composition. Chemical substances are determined by a set of chemical elements (as systems) and their numeric contents in substance. Chemical modifications are defined as chemical substances having special crystal structure of phases. Metabase contains also information on properties, which data are stored in different DBs, and other data. This information is enough to make search for relevant chemical systems and data on properties of substances and materials.

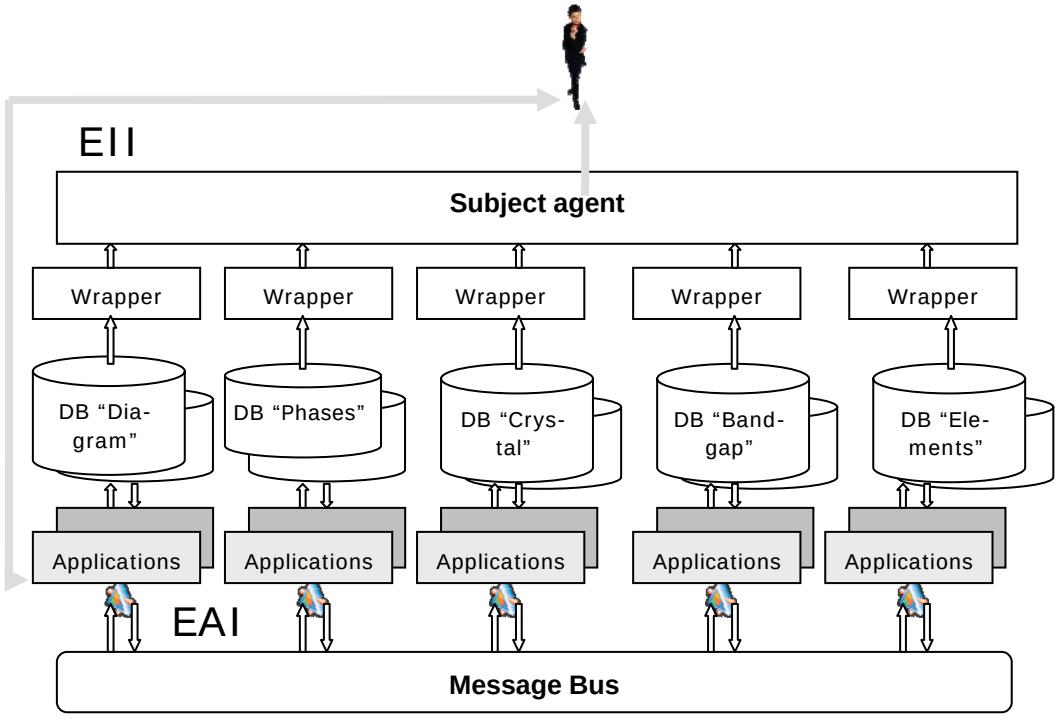

Fig. 5. Structure of integrated DBs system of Baikov Institute.

Now the integrated information system includes five DBs that were developed by Baikov Institute: DB on the properties of inorganic compounds "Phases" [Kiselyova et al., 2006], DB on phase diagrams of semiconducting systems "Diagram" [Khristoforov et al., 2001], DB on substances with significant acousto-optical, electro-optical and nonlinear-optical properties "Crystal" [Kiselyova et al., 2004], DB on width of the forbidden zone of inorganic substances "Bandgap" [Dudarev et al., 2006] and DB on properties of chemical elements "Elements" (fig. 5). One of the most important features of the developed integrated system is that DBs which have been included into integrated system were created with various DBMS using essentially different computer platforms: Sun (DB "Diagram") and Intel (other DBs) and different operational systems: Sun Solaris (DB "Diagram") and Microsoft Windows 2003 Server (other DBs). However the way, offered by us, has appeared successful even in such a difficult case for program realization.

Integration of Russian and Japanese Databases on Inorganic Material and Substance Properties

Next stage is an expansion of integrated system. Information system of Baikov Institute will be integrated with other Russian [Zemskov et al., 1998] and foreign DBs [Villars et al., 2004; Xu et al., 2006] on inorganic materials and substances. Principles of integration are based on the application of metabase and combined approach that was developed in Baikov Institute [Dudarev et al., 2006; Kornuyshko and Dudarev, 2006]. Sometimes small additional tables, that contain information about sets of elements and their contents in substance and crystal structure, will be included into these DBs.

The following structure of metabase can be used for integration of the Web-applications of DBs on properties of inorganic substances and materials (fig. 6).

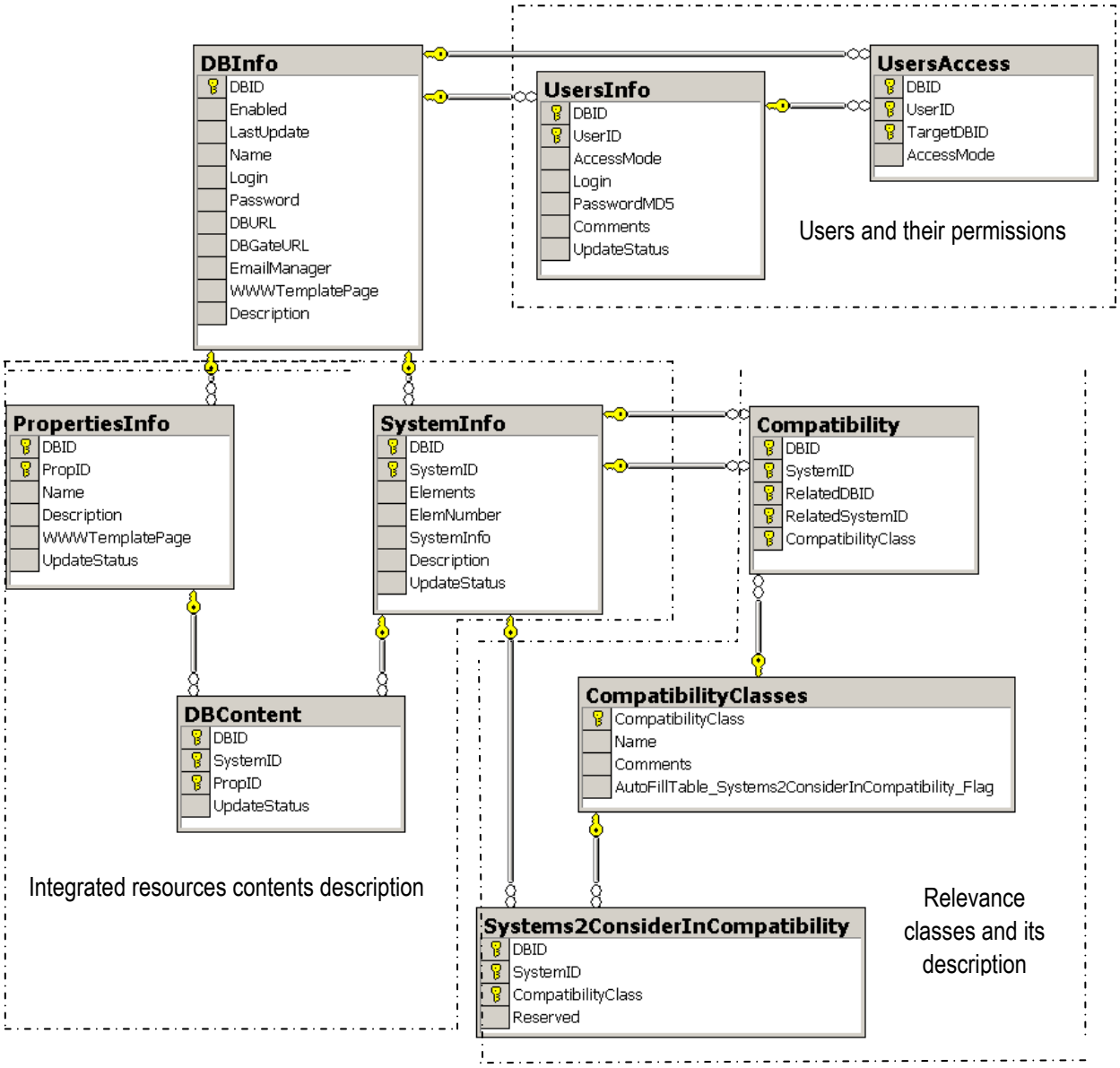

Fig. 6. Metabase structure for DBs Web-applications integration.

Purpose of the tables (fig. 6): DBInfo – main table containing the information about DBs Web-applications to be integrated; UsersInfo, UsersAccess - tables containing the information on the users of integrated system and their access permissions to information; SystemInfo, PropertiesInfo, DBContent – tables that describe contents of resources to be integrated (what information on chemical systems and their properties is stored in what DB); CompatibilityClasses, Compatibility, Systems2ConsiderInCompatibility – tables that contain information on relevance classes and determe relevant chemical systems.

Conclusion

The complex approach to information integration combining the integration at a level of data and user interfaces (EII+EAI) is offered. Within proposed approach access was implemented to all current user interfaces of virtually united information system. Moreover the system allows to users to move freely between different applications (EAI). According to the common developed information schema subject mediator was implemented. It provides rich opportunities on information extraction and aggregation from diverse distributed data sources on material and substance properties (EII).

The tasks of search for the relevant data in integrated information systems and implementation of transparent user transition between DBs Web-applications (taking into account the security issues) were solved at DBs Web-applications integration. For search mechanisms implementation for the relevant information was used metadata database (metabase) – special reference database containing metadata only – the information on information systems to be integrated. Diverse data sources integration is based on conceptual structure of the knowledge domain (inorganic chemistry) and development of the heterogeneity conflicts resolution ways.

The system of databases on inorganic material and substance properties is accessible for registered users of the Internet: http://www.imet-db.ru.

The work is supported by RFBR, grants №06-07-89120 and 05-03-39009.

Bibliography

[Bale and Eriksson, 1990] C.W.Bale and G.Eriksson. Metallurgical thermochemical databases a review. Can.Met.Quart. 1990, v.29.

[Dudarev et al., 2006] V.A.Dudarev, N.N.Kiselyova, V.S.Zemskov. Integrated system of databases on properties of materials for electronics. Perspektivnye Materialy, 2006, N.5 (Russ.).

[Eriguchi and Shimura, 1990] K.Eriguchi and K.Shimura. Factual databases for materials design and manufacturing. ISIJ Int., 1990, v.30.

[Imhoff, 2005] C.Imhoff. Intelligent Solutions: Understanding the Three E's of Integration EAI, EII and ETL. DM Review Magazine, 2005, apr. (http://www.dmreview.com/article_sub.cfm?articleId=1023893).

[Khristoforov et al., 2001] Yu.I.Khristoforov, V.V.Khorbenko, N.N.Kiselyova, et al. Internet-accessible database on phase diagrams of semiconductor systems. Izvestiya VUZov. Materialy elektron.tekhniki, 2001, №4 (Russ.).

[Kimball and Caserta, 2004] R.Kimball and J.Caserta. The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data. John Wiley & Sons, 2004.

[Kiselyova, 2005] N.N.Kiselyova. Computer Design of Inorganic Compounds. Application of Databases and Artificial Intelligence. Nauka, Moscow, 2005 (Russ.).

[Kiselyova et al., 2005] N.N.Kiselyova, V.A.Dudarev, I.V.Prokoshev, et al. The distributed system of databases on properties of inorganic substances and materials. Int.J."Information Theories & Applications", 2005, v.12.

[Kiselyova et al., 2006] N.Kiselyova, D.Murat, A.Stolyarenko, et al. Database on ternary inorganic compound properties "Phases" in Internet. Informazionnye resursy Rossii, 2006, N.4 (Russ.).

[Kiselyova et al., 2004] N.N.Kiselyova, I.V.Prokoshev, V.A.Dudarev, et al. Internet-accessible electronic materials database system. Inorganic materials, 2004, v.42, №3.

[Kornuyshko and Dudarev, 2006] V.Kornuyshko and V.Dudarev. Software Development for Distributed System of Russian Databases on Electronics Materials. Int. J. "Information Theories & Applications", 2006, v.13.

[Morgenthal, 2000] J.P.Morgenthal. Enterprise Applications Integration with XML and Java. Prentice Hall PTR; Bk&CD Rom edition, 2000.

[Morgenthal, 2005] J.P.Morgenthal. Enterprise Information Integration: A Pragmatic Approach. Lulu.com, 2005.

[Villars et al., 2004] P.Villars, M.Berndt, K.Brandenburg, et al. The Pauling File, binaries edition. J.Alloys and Compounds, 2004, v.367.

[Xu et al., 2006] Y.Xu, M.Yamazaki, H.Wang, K.Yagi. Development of an Internet system for composite design and thermo-physical property prediction. Mater.Trans., 2006, v.47.

[Zemskov et al., 1998] V.S.Zemskov, F.A.Kuznetsov, V.B.Ufimtsev. Databanks on semiconducting and other materials for electronics and technology of their productions. Izvestiya VUZov. Materialy elektronnoi tekhniki, 1998, N.3 (Russ.).

Authors' Information

Nadezhda Kiselyova – A.A.Baikov Institute of Metallurgy and Materials Science of Russian Academy of Sciences, P.O.Box: 119991 GSP-1, 49, Leninskii Prospect, Moscow, Russia, e-mail: kis@ultra.imet.ac.ru

Shuichi Iwata – Graduate School of Frontier Sciences, The University of Tokyo, P.O.Box: 113-8656, Room 507A, Building E12 QUEST, 7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan, e-mail: iwata@k.u-tokyo.ac.jp

Victor Dudarev – A.A.Baikov Institute of Metallurgy and Materials Science of Russian Academy of Sciences, P.O.Box: 119991 GSP-1, 49, Leninskii Prospect, Moscow, Russia, e-mail: vic@osg.ru

Ilya Prokoshev – A.A.Baikov Institute of Metallurgy and Materials Science of Russian Academy of Sciences, P.O.Box: 119991 GSP-1, 49, Leninskii Prospect, Moscow, Russia, e-mail: eldream@e-music.ru

Valentin Khorbenko – A.A.Baikov Institute of Metallurgy and Materials Science of Russian Academy of Sciences, P.O.Box: 119991 GSP-1, 49, Leninskii Prospect, Moscow, Russia, e-mail: Khorbenko_v@mail.ru

Victor Zemskov – A.A.Baikov Institute of Metallurgy and Materials Science of Russian Academy of Sciences, P.O.Box: 119991 GSP-1, 49, Leninskii Prospect, Moscow, Russia, e-mail: zemskov@ultra.imet.ac.ru