Creating Inorganic Chemistry Data Infrastructure for Materials Science Specialists

Nadezhda N. Kiselyova[1](✉) and Victor A. Dudarev[1,2]

[1] A. A. Baikov Institute of Metallurgy and Materials Science of Russian Academy of Sciences (IMET RAS), Moscow, Russia
{kis,vic}@imet.ac.ru

[2] National Research University Higher School of Economics (NRU HSE), Moscow, Russia

Abstract. The analysis of the large infrastructure projects of information support of specialists realized in the world in the field of materials science is carried out (MGI, MDF, NoMaD, etc.). The brief summary of the Russian information resources in the field of inorganic chemistry and materials science is given. The project of infrastructure for providing the Russian specialists with data in this area is proposed.

Keywords: Information support in materials science · Materials science database integration · Inorganic substances and materials

1 Introduction

Global market competitive requirements demand permanent improvement and enhancing of consumer products properties. Products quality and novelty are substantially defined by the materials used in its production. As a result, acceleration of search, research and production of new materials with required functional properties is a crucial problem of industry and economic development for all countries in general. At present, according to the American specialists [1], the time frame for incorporating new classes of materials into applications is about 20 years from initial research to first use. It is connected with the fact that very often consumers have no sufficient information even about very promising materials, investigations in a research and technology development for materials preparation and processing are unreasonably duplicated, therefore substances and materials with not the best consumer and other parameters are used that leads to loss of products quality, production cost escalation and, eventually, to loss of the released product market attractiveness.

One of ways to accelerate new materials search, development and deployment is the mature infrastructure for specialists' information support creation. First of all, it is a creation of distributed virtually integrated network of the databases and knowledge bases containing information on properties of substances and materials and technologies of their production and processing, and also, it's development of systems for computer-aided design and modeling of the materials available from the Internet to wide range of

L. Kalinichenko et al. (Eds.): DAMDID/RCDL 2016, CCIS 706, pp. 222–236, 2017.
DOI: 10.1007/978-3-319-57135-5_16

specialists: to scientists, engineers, technologists, businessmen, government employees, students, etc.

In recent years in the developed countries the initiatives aimed to the infrastructure organization for access to experimental and calculated data about materials were announced and supported by the governments. The brief summary of some initiatives was given in some publication earlier [2, 3]. Current review considers set-theory approach to inorganic substances and materials database integration and the project of infrastructure for providing Russian specialists with data in the field of inorganic chemistry and materials science in more details comparing with [2].

2 Materials Genome Initiative (MGI)

In 2011 the USA started a project, called Materials Genome Initiative (MGI) [4]. The MGI aims are to provide accelerated creation of the new materials with a set of predefined properties which is critical for achievement of the high level of competitiveness of the industry in the USA and will promote support to their leading role in many modern materials science and industry areas: from power engineering to electronics, from national defense to health care. In MGI special attention is paid to breakthrough researches support in the theory, materials properties modeling and data mining as means of significant progress achievement in materials science that will lead to cost reduction during new materials research, development, and production. The MGI tasks are in ensuring new materials development and deployment, including research activities coordination and providing access to computable models and tools for materials properties and behavior assessment, and also breakthrough methods of modeling and data analysis usage. The MGI project implementation will allow creation of the mechanisms promoting materials data and knowledge exchange not only between researchers but also between the academic science and the industry. A basis of the MGI is Materials Innovation Infrastructure (MII) which provides integration of modern modeling methods and experimental research. The infrastructure includes interconnected service structures and objects complex (including the objects of megascience) making and/or providing a basis for materials science functioning as a science and in the applied area. Subcommittee on the MGI The National Science and Technology Council (NSTC) includes representatives of the United States Department of Defense, Department of Energy, National Institute of Standards and Technology (NIST), National Science Foundation (NSF), National Aeronautics and Space Administration (NASA), National Institutes of Health (NIH), United States Geological Survey (USGS), Defense Advanced Research Projects Agency (DARPA), etc. [5]. The successful AFLOW system [6, 7] should be highlighted among MGI-supported projects. It contains a database with substances quantum mechanical calculations results and equipped with the computer software package for carrying out such calculations. The new type of the ultra-stable and wearproof glass discovery was also performed by wide use of theoretical calculations [8] within MGI. Another example is the virtual high-throughput experimentation facility with the goal of accelerating the generation of huge data volumes, needed to validate existing materials models for new substances with predefined properties prediction (NIST and the National Renewable Energy Laboratory (NREL)), etc. [9].

3 The Materials Data Facility (MDF)

Considering materials importance for the US industry high level competitiveness achievement, in June, 2014 the National Data Service (NDS) Consortium announced a pilot project for materials science data facilities development: The Materials Data Facility (MDF) [10] supported by NIST. This project is the answer to the MGI of the White House aimed to accelerate modern materials development. The MDF will provide materials scientists with a scalable repository for experimental and calculated data (including prior to their publication) storage, supplied with references to the appropriate bibliographic sources. The MDF will be an instrument of national infrastructure creation for collective information use, including DBs on materials properties developed in the world and information systems for calculation and modeling, and also will promote materials data exchange, including not published data. Materials data and calculation facilities availability is provided by means modern information and telecommunication infrastructure which allows to provide data to materials researchers for multi-purpose use, additional analysis and verification. In addition to NIST, it is necessary to note several main contributors to MDF: The University of Chicago, Argonne National Laboratory, The University of Illinois, Northwestern University, Center for Hierarchical Materials Design, etc. The MDF repository currently includes [10], in addition to a numerous NIST DBs [11], information systems with quantum mechanical calculations results: AFLOW [6, 7], The Open Quantum Materials Database (OQMD) [12], etc.

4 Novel Materials Discovery Laboratory (NoMaD)

The NoMaD project was the European Union answer to the US MGI. The NoMaD project [13, 14] is directed to the European Centers of Excellence creation and contemplates development of DB network (Materials Encyclopedia) on substances and materials properties (mainly on the calculations results), and also a number of facilities for data analysis and substances calculation. The purpose is materials with predefined functional properties development and use acceleration. The program started in November, 2015 within the EU's HORIZON2020 project [14]. Essential disadvantage of the NoMaD is an orientation to the US information resources (mainly, NIST DBs on substances and materials properties) and information systems with calculated data. Nowadays the NoMaD repository [15] contains already synthesized substances quantum mechanical calculations results only. In many respects the NoMaD program correlates with the EU Materials design at the eXascale (MaX) [16] project including infrastructure creation for carrying out quantum mechanical calculations by means of high-performance computer systems. Among the NoMaD partners there are a number of Europe's leading organizations, such as Humboldt University, Fritz-Haber-Institute of the Max Planck Society, King's College London, University of Barcelona, Aalto University, Max Planck Institute for the Structure of Dynamics of Matter, Technical University of

Denmark, Max Planck Computing and Data Facility, Barcelona Supercomputing Centre, etc.

5 Materials Research by Information Integration Initiative (MI^2I)

The MI^2I was offered in 2015 by the Japanese government, which created the Center for Materials Research by Information Integration based on the National Institute for Materials Science (NIMS) [17]. Unlike the European programs the created center is aimed to wide use of not only quantum mechanical calculations, but also to support of DBs on substances and materials properties developed in Japan [18] and to integrate them with foreign information systems and to take advantage of artificial intelligence methods application for new substances design [19, 20].

6 Chinese Materials Genome Initiative

This five-year project started in China in 2016 at support of Ministry of Science and Technology [21]. Previously between 2014 and 2015 several MGI centers were organized: Shanghai Institute of Materials Genome (2014), Beijing Key Laboratory for Materials Genome (2015) and International Institute of MGI in Ningbo (2015), etc. The project goals are similar to the US MGI.

7 Large-Scale Infrastructure Projects for Materials Science Information Support Analysis

There should be noted several general trends in information support systems development in materials science areas:

- Integrated network of DBs on materials and substances properties development;
- development and broad application of computational methods;
- DBs with calculated information on materials development.

The analysis of goals and their achievement methods shows that the projects of the USA and China are most promising. In future, they could allow creation of full-fledged infrastructure for innovative activity information support in new materials development and deployment, having provided the science and the industry with reliable and complete data on substances and materials properties together with various tools (packages for quantum mechanical calculations, data mining, etc.) for substances parameters calculations. Japanese initiative is more limited than American one, because it is based on NIMS databases on materials and substances properties usage, and it also uses already known calculation methods experience (for example, VASP – widely known package for quantum mechanical calculations [22]). Investigations on artificial intelligence methods application were begun [19, 20]. Besides, the Japanese specialists are limited in a research field since they consider materials for electronics only (power supplies, magnetic, thermoelectric and spintronic materials) [17]. The EU projects at their initial stage seems to be the least promising. Orientation to the American DBs on substances properties and together with only quantum mechanical calculations significantly reduces these infrastructure projects potential and opportunities.

Nevertheless, it should be noted that to successfully implement of offered in the USA, the EU, China and Japan initiatives it is required, on the one hand, to get a breakthrough in materials properties calculation methods development, and, on the other hand, to achieve a progress in availability of databases on substances properties developed in recent years in different countries. The overview of available DBs on inorganic chemistry and materials science is given in the article [23] and in the IRIC (Information Resources on Inorganic Chemistry) information system [24]. In spite of the fact that hundreds of millions of dollars are spent for materials science information systems creation and support, their use is economically profitable since they allow reducing costs for new materials development considerably due to researches duplication reduction and reliable online information on substances properties providing to chemists and materials scientists. In turn, calculation methods give us a chance for substances parameters 'a priori' estimation, of prediction of substances, promising for industry applications, and for development of technology for materials production and processing. The consequence of these tasks solution is cost and time reduction for new materials development and deployment.

8 Integrated Information System on Inorganic Substances and Materials Properties Development Experience

Russian investigators undertake attempts to create their own materials infrastructure for different application areas. Premises for materials infrastructure project successful accomplishment in Russia are an experience in available from the Internet databases on inorganic substances and materials properties development and integration, together with expertise in methods and software for new substances and materials computer-aided design based on data mining technologies and, first of all, precedent based pattern recognition methods [23, 25, 26].

It should be noted that interest in data mining methods application to inorganic materials science is connected with objective difficulties arising at quantum mechanical calculations for yet not synthesized multicomponent inorganic substances, especially in a solid phase. For example, to calculate inorganic compound electronic structure by means of VASP package [22], it is necessary to know its crystal structure, i.e. it is necessary to produce and investigate this substance. Using pattern recognition methods for information analysis on already known substances from DBs it is possible to predict yet not synthesized substances and to estimate some of their properties, having only well-known components parameters (chemical elements or simple compounds). The special Information-Analytical System (IAS) (Fig. 1) was developed to solve this task in IMET RAS. The system includes the integrated databases system on inorganic substances and materials properties, the subsystems for regularities search in data and new substances prediction and their properties estimation, the knowledge base, predictions base and other subsystems [23, 26].

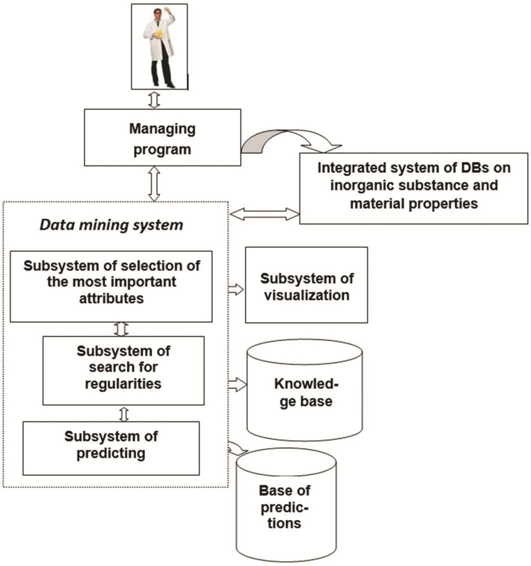
Managing program
Integrated system of DBs on inorganic substance and material properties
Data mining system
Subsystem of selection of the most important attributes
Subsystem of search for regularities
Subsystem of predicting
Subsystem of visualization
Knowled-ge base
Base of predic-tions

Fig. 1. Information-analytical system for inorganic compounds design structure.

8.1 The Integrated System of Databases on Inorganic Substances and Materials Properties

The integrated system of databases on inorganic substances and materials properties currently consolidates several information systems developed in IMET RAS [23, 26]: database on phase diagrams of semiconductor systems (“Diagram”), DB on substances with significant acousto-optical, electro-optical and nonlinear optical properties (“Crystal”), DB on inorganic substances forbidden zone width (“Bandgap”), DB on inorganic compounds properties (“Phases”) and DB on chemical elements properties (“Elements”), and also “AtomWork” - DB on inorganic substances properties, developed in National Institute for Materials Science (NIMS, Japan), and “TKV” - DB on substances thermal constants developed in JIHT RAS and Lomonosov Moscow State University cooperation. This virtually integrated system allows providing access for user to wide variety of materials science information:

- “Phases” DB on inorganic compounds properties [27, 28] currently contains information on properties of more than 52 thousand ternary compounds (i.e. the compounds formed by three chemical elements) and more than 31 thousand quaternary compounds, obtained from more than 32 thousand publications. It includes summary for the most widespread inorganic compounds properties: crystal-chemical (type of crystal structure with indication of temperature and pressure above which the specified structure is formed, a crystal system, space group, number of formula units in unit cell, crystal lattice parameters) and thermophysical (melting type, melting and boiling points and compound dissociation temperature in solid or gas phases at an atmospheric pressure). In addition, the DB contains information on compounds superconducting properties. The “Phases” DB is formed by materials science experts on the basis of data analysis of periodicals, handbooks, monographs, reports, and abstract journals also (more than a half of sources are stored in PDF documents form). Currently “Phases” DB volume exceeds 25 GBytes, and it is available online to registered users from the Internet [28].

- “Elements” DB [23, 29] contains information on 90 most widespread chemical elements properties: thermophysical (melting and boiling points at 1 atmosphere, standard heat conductivity, molar heat capacity, an atomization enthalpy, entropy, etc.), dimensional (ionic, covalent, metal, pseudo-potential radii, atom volume, etc.), other physical properties (a magnetic susceptibility, electrical conductivity, hardness, density, etc.), position in the Periodic table of elements, etc. The DB is available online from the Internet [29].
- “Diagram” DB [30, 31] contains the information collected and estimated by highly skilled experts on phase P-T-x-diagrams of binary and ternary semiconductor systems and about physical and chemical properties of the phases which are formed in them. The DB volume exceeds 2 GBytes. The DB is available online to the registered users from the Internet [31].
- “Bandgap” DB [32, 33] contains information (more than 0.7 GBytes) on the forbidden zone width for more than 3 thousand inorganic substances. The DB is available online to users from the Internet [33].
- “Crystal” DB [34, 35] contains information on properties (piezoelectric (piezoelectric coefficients, elastic constants, etc.), nonlinear optical (nonlinear optical coefficients, Miller tensor components, etc.), crystal-chemical (crystal structure type, a crystal system, space group, number of formula units in unit cell, crystal lattice parameters), optical (refraction indices, transparency band, etc.), thermophysical (melting point, heat capacity, heat conductivity, etc.), etc.) for more than 140 acousto-optical, electro-optical and nonlinear optical substances collected and estimated by highly skilled experts in this subject domain. The DB volume exceeds 4 GBytes. It supports Russian and English languages for user interface and it’s available online to the registered users from the Internet [35].
- “Inorganic Material Database – AtomWork” DB [36, 37] contains information on more than 82 thousand crystal structures, 55 thousand materials properties values and 15 thousand phase diagrams. The DB is available online to users from the Internet [37].
- “TKV” DB on substances thermal constants [38] contains information, available online from the Internet, on about 27 thousand substances formed by all chemical elements.

Taking into account current development trends in materials science databases, the complex integration approach that combines integration at data and user interfaces level is applied to materials database consolidation.

When integrating databases at user interfaces (actually, Web applications) level it’s required to provide facilities for browsing information contained in other databases. This information should be relevant to the data on some chemical system currently being browsed by user. For example, user who browses information on Ga-As system from “Diagram” database should have an opportunity to get information for example on piezoelectric effect or nonlinear optical properties of GaAs substance contained in “Crystal” database. To achieve this type of behavior, it’s required to provide search for relevant information contained in other databases of distributed system. Thus, it’s necessary to have some active data center that should know what information is contained in every integrated database. Obviously, some data store should exist that somehow describes information on chemical entities, contained in integrated database resources. In this manner, metabase concept appears – a special database, that contains some reference information on integrated databases contents [39].

Metabase determines integrated system capabilities. Its structure should be flexible enough to represent metadata information on chemical entities of integrated DBs on inorganic materials properties. Taking into consideration the fact that chemical entities and their corresponding properties description is given at different detail level in different DBs, it’s important to develop metabase structure that would be suitable for description of information residing in different DBs on inorganic substances properties. For example, some integrated DBs contain information on particular crystal modifications properties while others contain properties description at chemical system level. Thus, integrated DBs deal with different chemical entities. Relation between chemical entities can be described by means of chemical entities hierarchy in tree form (Fig. 2).

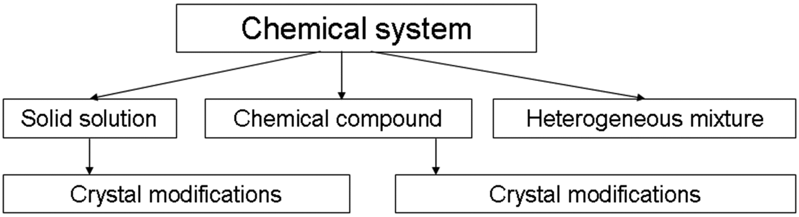

Fig. 2. Chemical entities hierarchy.

Having designated second level entities by the “substance” term, we get three-level chemical entities hierarchy: chemical system, chemical substance and crystal modification [39]. As far as information stored in DBs on inorganic substances properties can be considered at chemical system level, for simplicity in the paper this level will be taken from the top of the entities hierarchy.

The relevant information search problem can be formalized in terms of set-theoretic approach. Hence, metabase should contain information on integrated databases (*D* set), information on chemical substances and systems (*S* set) and information on their properties (*P* set). To describe correlation between elements of *D*, *S* and *P* sets let’s define ternary relation called *W* on set $U = D \times S \times P$. Here *U* is a Cartesian product of *D*, *S* and *P*. Membership of a (d, s, p) triplet to the *W* relation, where $d \in D$, $s \in S$, $p \in P$. can be interpreted in the following way: “Information on property *p* of chemical system *s* is contained in integrated database *d*”. Having defined three basic sets it can be seen that search for information relevant to *s* system can be localized to determination of *R* relationship, that is a subset of Cartesian product $S \times S$ (or in other words, $\mathrm{R} \subseteq \mathrm{S}^2$). Thus, it can be stated about every pair $(s_1, s_2) \in R$ that chemical system s_2 is relevant to the system s_1. So, to solve the task of relevant information search in integrated databases it’s required to define the R relation. It is significant to note that R relation can be created or complemented by means of either of two variants. The first variant is via using predefined rules by a computer. The second one is that experts in chemistry and materials science can be engaged to solve this task.

The second variant is quite clear – experts can form relationship R manually following some multicriterion rules affected by their expert assessments. To consider one of the simplest ways to automatic R relation generation, it's required to implement a couple of rules, based on substances composition [39]:

1. For any chemical systems $s_1 \in S$, $s_2 \in S$ composed from chemical elements e_{ij}, $s_1 = \{e_{11}, e_{12}, \ldots, e_{1n}\}$, $s_2 = \{e_{21}, e_{22}, \ldots, e_{2m}\}$ it is true, that if $s_1 \subseteq s_2$ (i.e. all chemical elements of system s_1 are contained in system s_2), then $(s_1, s_2) \in R$. (1)
2. R relation is symmetric. In other words, for any $s_1 \in S$, $s_2 \in S$, it is true, that if $(s_1, s_2) \in R$, then $(s_2, s_1) \in R$ as well. (2)

These two rules allow determination of a set of chemical systems relevant to the given one. It should be noted that this automatic R relation generation variant is just one of the simplest and most obvious variants of such rules, and in fact more complex mechanisms can be used to get R relation. For example, browsing information on a particular property of a compound in one of integrated databases (in fact, it is information defined by (d_1, s_1, p_1) triplet), it is possible to consider (d_2, s_2, p_2) triplet to be relevant information. (d_2, s_2, p_2) triplet characterizes information on some other property of a system from another integrated database. In this case, more complex relevance relation arises: $R \ni (d_1, s_1, p_1) \times (d_2, s_2, p_2)$, where $d_1, d_2 \in D$, $s_1, s_2 \in S$, $p_1, p_2 \in P$. In fact, it's possible to define even a set of several R relations $(R_1, R_2, \ldots, R_n)$ by applying different rules. Thus, potentially it's possible to perform search for relevant information based on wide variety of R interpretations to provide required flexibility if needed.

To provide secure and transparent user transitions to relevant information DB Web-application it is planned to use special metabase gateways. To consider an example of integrated system functioning let's assume that in one of integrated system a user browses information on some particular chemical system. In other words, the user is in Web application of a particular information system. If it is necessary to get relevant information, the Web application will be capable to send a request to specially developed Web service that serves integrated system users. The request goal is to get information contained in integrated resources that is relevant to the currently browsed data. After the request the Web service sends a response to the Web application in a form of XML document. It describes what relevant information on chemical systems and properties is contained in integrated resources. It's well-known, that data in XML format are properly understood on all major platforms. That information can be output to user for example by means of a XSL-transformation in form of HTML document (XML + XSL = HTML) containing hyperlinks to special gateway [39]. The user could follow from one Web application to another to browse relevant information via this gateway only.

The gateway is a specialized Web application that runs on the metabase Web server. The gateway main purpose is to perform security-dispatching function in distributed system. According to the task stated it is responsible for user authentication and it also checks whether the user has required privileges to address the information requested. If that authentication is successful, i.e. the user is eligible to address the data, then the metabase security gateway will perform redirection to a specialized entry point of desired Web application adding some additional information to create proper security context in target Web application and supplying it with digital signature. It should be stated that the entry point is a specialized page in target Web application that is to perform service functions for integrated system users. At this page target Web application checks digital signature of the metabase security gateway and if everything is fine the page will create special security context for user with given access privileges within target Web application. Finally, the user is automatically redirected to the page with the information required. In spite of redirection process apparent complexity, user transition from one Web application to another is absolutely transparent [39]. Thus, end user can even not note that some complex processing has been done to perform redirection. So, it is an illusion created that having clicked on a hyperlink the user is transferred from one information system to another directly.

8.2 Inorganic Compounds Computer-Aided Design System

The inorganic compounds computer-aided design system background is formed by precedent-based pattern recognition algorithms and programs which are collected within the multipurpose “Recognition” system, developed by Dorodnicyn Computing Centre of RAS [40] and combining in addition to widely known methods of linear machine, linear Fischer discriminant, k-nearest neighbors, support vector machine, neural networks and genetic algorithms, also a number of unique algorithms developed in Dorodnicyn Computing Centre of RAS: the recognition algorithms based on estimates calculations, deadlock tests voting algorithms, logical patterns voting algorithms, statistical weighed voting algorithms, etc. The inorganic compounds computer-aided design system also includes ConFor system, developed at Institute of Cybernetics of National Academy of Sciences of Ukraine [41]. The ConFor is a software for machine learning based on so-called growing pyramidal networks which are special data structures in computer memory that form subject domains concepts.

To select the most important components properties IAS includes several programs based on various algorithms [42–44]. The developed system deployment makes it possible to predict new inorganic compounds and estimate various properties of those without experimental synthesis [25, 26, 32].

9 The Project of Infrastructure for Providing Russian Specialists with Data in the Field of Inorganic Chemistry and Materials Science

IAS is, some kind of, pilot project for creation of information infrastructure for inorganic materials science. According to this project the most known Russian DBs in this area are virtually integrated, and also integration of Russian systems with foreign information systems was begun. The majority of Russian DBs contain references to full publications texts from which information of DBs was extracted. The compound computer-aided design subsystem allows searching for regularities in DB information and applying them to yet not synthesized substances prediction and their properties estimation. It should be noted that at a prediction phase only data on compounds components properties (chemical elements or simple compounds) are used. Obtained predictions are stored in special predictions base that expands traditional databases functionality (user obtains not only well-known experimental data, but also predictions for yet not synthesized compounds and some of their properties estimation).

When developing the Russian infrastructure project for specialists in the field of inorganic materials science information support it is necessary to consider all possible user queries variety. It is natural that academic scientist's queries could significantly differ from queries of design engineers or materials producers. However the general information infrastructure project should necessarily include virtually integrated system of Russian and foreign databases on inorganic substances and materials properties, production and processing technologies, materials producers and consumers, etc., a complex of packages for materials modeling and calculation which are widely used in most cases by academic scientists, and virtually integrated databases system with already calculated values to simplify new materials industry adoption (Fig. 3). It is necessary to emphasize that technologies of processing, storage and search of necessary data require development and usage of the most modern software and powerful data-processing centers (DPC) creation.

Materials DB system should consolidate factual DB on inorganic substances and materials that are the most important for Russian users (Russian DBs: IMET RAS, JIHT RAS, MSU, etc. and foreign DBs: NIMS [18], NIST [11], STN [45], Springer Materials [46], etc.), the leading publishing corporations documentary DBs (Science, Elsevier, Springer, Wiley, American Chemical Society, American Institute of Physics, Science, etc.), and also databases with yet not published information (All-Russian Institute of Scientific and Technical Information, Center of Information Technologies and Systems, etc.), patent databases (Rospatent, Questel, USPTO, etc.), databases on inorganic materials producers and consumers, etc. It is necessary to allocate funds for annual licenses prolongation for foreign materials databases usage and to organize the uniform portal with free access for Russian users (currently such databases are available to limited organizations only). It is necessary to support in every possible way transfer to an electronic form of paper collections of popular Russian scientific journals, which are the most in the world (in addition it will undoubtedly promote them and increase their authority and impact-factors).

To equip research organizations with intellectual calculation systems it's required, first of all, to start with students and graduate students training for use of the most known quantum mechanical, thermodynamic, statistical, etc. calculations packages. It is necessary to develop Russian databases with calculated values and to integrate them with foreign information systems, which are available in the Internet currently (for example, [15]) that will allow partially solution of a problem of qualified calculations on substances properties. Experiment planning should include calculations usage at initial research stage that will allow reduction of time and costs for new materials search and development.

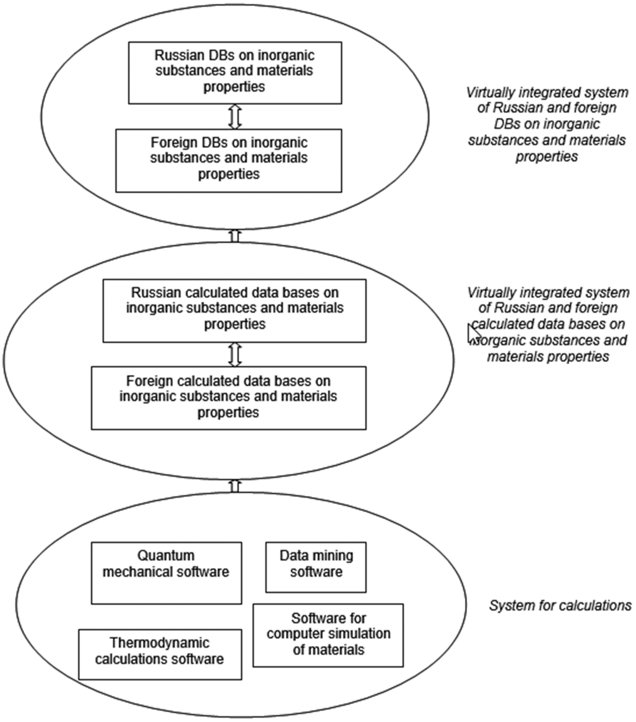
Russian DBs on inorganic substances and materials properties
Foreign DBs on inorganic substances and materials properties
Virtually integrated system of Russian and foreign DBs on inorganic substances and materials properties
Russian calculated data bases on inorganic substances and materials properties
Foreign calculated data bases on inorganic substances and materials properties
Virtually integrated system of Russian and foreign calculated data bases on inorganic substances and materials properties
Quantum mechanical software
Data mining software
Software for computer simulation of materials
Thermodynamic calculations software
System for calculations

Fig. 3. The scheme of infrastructure for providing Russian specialists with data in the field of inorganic chemistry and materials science.

User queries analysis subsystem, especially, for specialists in applied areas should become an important component of the developed infrastructure project. It will allow discovery of special interest materials groups to which research and investigation should be aimed to. The statistics of absence in user requested values of a particular substance parameter can become a signal to initiate an additionally required property investigation.

10 Conclusion

Transition of the Russian economy to an innovative way of development and increase in product competitiveness is defined in many respects by materials quality, novelty and functionality. At the present technologies development stage, new materials search, research and implementation requires mature infrastructure creation including the academic organizations with their new substances theoretical and pilot investigation potential, the organizations conducting applied researches on development and deployment of new materials and their production and processing technologies, the shareable scientific equipment centers with expensive equipment complexes including mega-science objects, etc. In recent years in the developed countries several projects were initiated (MGI, MDF, NoMaD, etc.). These projects are of strategic importance for achieving technological excellence by creating infrastructure for new materials with predefined functional properties set development and deployment acceleration. Special attention in these projects is paid to information support infrastructure. The Russian answer to these strategic initiatives of the USA, the EU, Japan, China in materials information infrastructure can be creation of the federal information center providing specialists with information on substances and materials properties, corresponding production technologies, and calculated properties values also, patent information, etc. In connection with considered subject domain peculiarities the distributed integrated network of Russian and foreign databases and knowledge bases on substances and materials should form a basis of such shared information center. Federal information center creation that integrates materials science information resources will stimulate faster new materials search, development and deployment. In a combination with considerable cost reduction due to researches duplication elimination it also will provide chemists and materials scientists with operational and reliable experimental and calculated information on substances and materials.

Acknowledgements. The authors thank A. V. Stolyarenko, V. V. Ryazanov, O. V. Sen'ko, A. A. Dokukin for their help in an information-analytical system development. Work is partially supported by the Russian Foundation for Basic Research, projects 16-07-01028, 17-0701362 and 15-07-00980.

References

1. Materials Genome Initiative: "Strategic Plan. National Science and Technology Council. Committee on Technology", Subcommittee on the Materials Genome Initiative. https://www.whitehouse.gov/sites/default/files/microsites/ostp/NSTC/mgi_strategic_plan__dec_2014.pdf
2. Kiselyova, N.N., Dudarev, V.A.: Inorganic chemistry and materials science data infrastructure for specialists. In: Selected Papers of the XVIII International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2016), vol. 1752, pp. 121–128. CEUR Workshop Proceedings (2016)
3. Kalinichenko, L.A., Volnova, A.A., Gordov, E.P., Kiselyova, N.N., et al.: Data access challenges for data intensive research in Russia. Informatika i ee Primeneniya – Inf. Appl. **10**(1), 3–23 (2016)
4. Materials Genome Initiative for Global Competitiveness. http://www.whitehouse.gov/sites/default/files/microsites/ostp/materials_genome_initiative-final.pdf
5. Materials Genome Initiative. https://www.mgi.gov/partners
6. Curtarolo, S., Setyawan, W., Wang, S., et al.: AFLOWLIB.ORG: a distributed materials properties repository from high-throughput ab initio calculations. Comput. Mater. Sci. **58**, 227–235 (2012)
7. Taylor, R.H., Rose, F., Toher, C., et al.: RESTful API for exchanging materials data in the AFLOWLIB.org consortium. Comput. Mater. Sci. **93**, 178–192 (2014)
8. University of Chicago: Microscopic animals inspire innovative glass research. http://www.uchicago.edu/features/microscopic_animals_inspire_innovative_glass_research/
9. The First Five Years of the Materials Genome Initiative: Accomplishments and Technical Highlights (2016). https://mgi.nist.gov/sites/default/files/uploads/mgi-accomplishments-at-5-years-august-2016.pdf

10. National Data Service: The Materials Data Facility. https://www.materialsdatafacility.org
11. NIST Data Gateway. NIST Online Databases. http://srdata.nist.gov/gateway/gateway?dblist=0
12. Saal, J.E., Kirklin, S., Aykol, M., et al.: Materials design and discovery with high-throughput density functional theory: the Open Quantum Materials Database (OQMD). JOM **65**(11), 1501–1509 (2013)
13. The Novel Materials Discovery (NOMAD) Laboratory. http://nomad-lab.eu/
14. The Novel Materials Discovery (NOMAD) Laboratory. EINFRA-5-2015 - Centres of Excellence for computing applications. http://cordis.europa.eu/project/rcn/198339_en.html
15. The NoMaD Repository. http://nomad-repository.eu/cms/
16. Materials design at the eXascale. http://cordis.europa.eu/project/rcn/198340_en.html
17. Center for Materials Research by Information Integration. http://www.nims.go.jp/eng/research/MII-I/index.html
18. NIMS Materials Database (MatNavi). http://mits.nims.go.jp/index_en.html
19. Lee, J., Seko, A., Shitara, K., Tanaka, I.: Prediction model of band-gap for AX binary compounds by combination of density functional theory calculations and machine learning techniques. Phys. Rev. B **93**(11), 115104 (2016)
20. Toyoura, K., Hirano, D., Seko, A., et al.: Machine-learning-based selective sampling procedure for identifying the low-energy region in a potential energy surface: a case study on proton conduction in oxides. Phys. Rev. B **93**(5), 054112 (2016)
21. Lu, X.-G.: Remarks on the recent progress of Materials Genome Initiative. Sci. Bull. **60**(22), 1966–1968 (2015)
22. The Vienna Ab initio Simulation Package (VASP). https://www.vasp.at/
23. Kiselyova, N.N., Dudarev, V.A., Zemskov, V.S.: Computer information resources in inorganic chemistry and materials science. Russ. Chem. Rev. **79**(2), 145–166 (2010)
24. IRIC DB (Information Resources on Inorganic Chemistry). http://iric.imet-db.ru/
25. Kiselyova, N.N.: Computer design of inorganic compounds. Application of databases and artificial intelligence. Nauka, Moscow (2005)
26. Kiselyova, N.N., Dudarev, V.A., Stolyarenko, A.V.: Integrated system of databases on the properties of inorganic substances and materials. High Temp. **54**(2), 215–222 (2016)
27. Kiselyova, N., Murat, D., Stolyarenko, A., et al.: Phases database on properties of ternary inorganic compounds on the Internet. Inf. Res. Russ. **4**, 21–23 (2006)
28. “Phases” DB. http://www.phases.imet-db.ru
29. “Elements” DB. http://phases.imet-db.ru/elements
30. Khristoforov, Y.I., Khorbenko, V.V., Kiselyova, N.N., et al.: The database on semiconductor systems phase diagrams with Internet access. Izv. Vyssh. Uchebn. Zaved. Mater. Electron. Tech. **4**, 50–55 (2001)
31. “Diagram” DB. http://diag.imet-db.ru
32. Kiselyova, N.N., Dudarev, V.A., Korzhuyev, M.A.: Database on the bandgap of inorganic substances and materials. Inorg. Mater. Appl. Res. **7**(1), 34–39 (2016)
33. “Bandgap” DB. http://www.bg.imet-db.ru
34. Kiselyova, N.N., Prokoshev, I.V., Dudarev, V.A., et al.: Internet-accessible electronic materials database system. Inorg. Mater. **42**(3), 321–325 (2004)
35. “Crystal” DB. http://crystal.imet-db.ru
36. Xu, Y., Yamazaki, M., Villars, P.: Inorganic materials database for exploring the nature of material. Jpn. J. Appl. Phys. **50**(11), 11RH02/1-5 (2011)
37. “AtomWork” DB. http://crystdb.nims.go.jp/index_en.html
38. “TKV” DB. http://www.chem.msu.su/cgi-bin/tkv.pl?show=welcome.html/welcome.html

39. Dudarev, V.A.: Information systems on inorganic chemistry and materials science integration. Krasand, Moscow. 320 p. (2016)
40. Zhuravlev, Y.I., Ryazanov, V.V., Senko, O.V.: Recognition. Mathematical methods. Program system. Practical applications. FAZIS, Moscow. 176 p. (2006)
41. Gladun, V.P.: Processes of forming of new knowledge. SD “Pedagog-6”, Sofia. 186 p. (1995)
42. Senko, O.V.: An optimal ensemble of predictors in convex correcting procedures. Pattern Recogn. Image Anal. **19**(3), 465–468 (2009)
43. Yuan, G.-X., Ho, C.-H., Lin, C.-J.: An improved GLMNET for L1-regularized logistic regression. J. Mach. Learn. Res. **13**, 1999–2030 (2012)
44. Yang, Y., Zou, H.: A coordinate majorization descent algorithm for L1 penalized learning. J. Stat. Comput. Simul. **84**(1), 1–12 (2014)
45. STN website. http://www.stn-international.de/
46. Springer Materials. http://materials.springer.com/