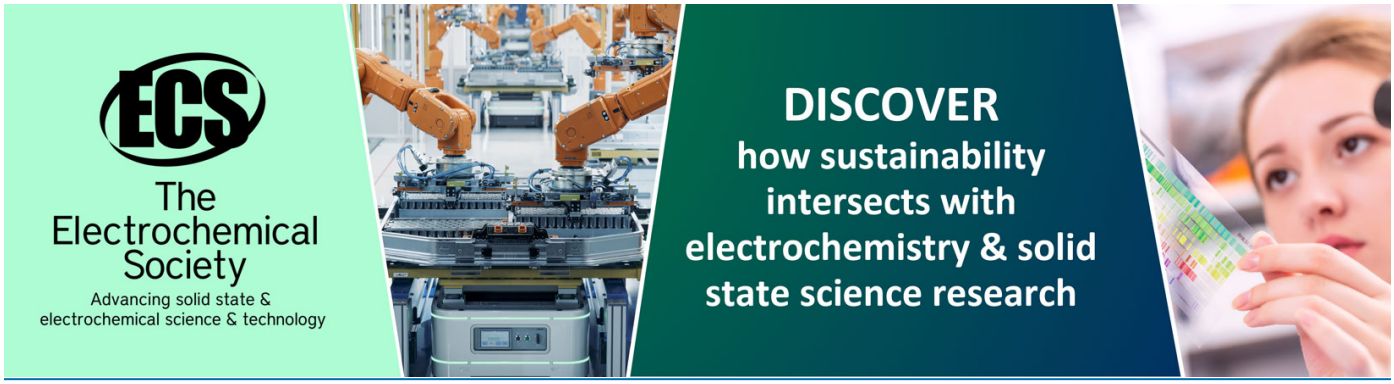
On importance of the multi-level chemical objects description in various problem domain models for information systems integration in inorganic materials science



Journal of Physics: Conference Series **1385** (2019) 012032 doi:10.1088/1742-6596/1385/1/012032

On importance of the multi-level chemical objects description in various problem domain models for information systems integration in inorganic materials science

V A Dudarev, N N Kiselyova

A. A. Baikov Institute of Metallurgy and Materials Science, Moscow, Russia

E-mail: vic@imet.ac.ru

Abstract. The paper proves the urgency of the integration of information systems (IS) on the properties of inorganic substances and materials (PISM). It is shown that consolidation is possible only on the basis of the problem domain formalization. The basic definitions are introduced and the formalization of the IS on PISM content is proposed on the basis of three models: verbal, set-theoretic and ontological.

1. Introduction

Modern researches in many areas of science are characterized by intensive big data accumulation and processing. The development of inorganic chemistry, as a science, has led to a huge number of research works aimed at a comprehensive study of various classes of inorganic substances properties. The results of these studies, as a rule, are drawn up in the form of scientific papers, which, at current stage of information technologies (IT) development, makes it almost impossible to computer-analyze and process existing natural language publications in order to extract knowledge and facts from them.

The development of specialized information systems (IS) on the properties of inorganic substances and materials (PISM) is necessary for the successful development of many modern industry high-tech areas, for example, electronics and mechanical engineering, since it allows a choice the best materials available for solving emerging technical problems. Therefore, in many highly-developed countries, significant investments are being made in the development of IS on PISM and material simulation systems, including those based on machine learning [1]. These systems are, in fact, an infrastructural foundation not only for the innovation industry, but also for the materials science itself.

2. Access problems to information on PISM

It should be noted that there is no single IS on PISM that contains all the data required for the analysis and often information is distributed across several IS on PISM, therefore in practice an access to such information distributed across various sources and its comprehensive analysis is a great problem even for a skilled specialist. The successful problem solution for a specialist is associated with two tasks. Firstly, in order to search for the necessary information,

at a minimum, it is required to know the list of IS on PISM, which may contain the required information. Secondly, if the specialist has access to the target IS on PISM, it will be necessary to search for the information required and to perform its comprehensive analysis.

The solution of the first task of finding the necessary IS on PISM is facilitated by the use of the specialized IS named Information Resources on Inorganic Chemistry (IRIC), which describes information resources on inorganic chemistry and materials science. Essentially, IRIC is an attempt to systematize the most significant IS on PISM [2]. The system is implemented as a Web-application and is 24/7 available at http://iric.imet-db.ru/ in Russian and English.

To solve the second task – to provide access to IS on PISM with the ability to quickly find the required information – it is necessary to integrate IS in this subject domain, which is not only a great organizational but also a technical problem.

3. Verbal model of the problem domain

For the successful consolidation of any ISs, first of all, it is necessary to formalize the problem domain description, which satisfies the ISs to be integrated.

A distinctive feature of many IS on PISM is a narrow orientation within problem domain, determined by the specificity of the field of study. Therefore, such systems store information only about those substances and their properties that relates to the research area. For example, "Diagram" IS can be mentioned as an example of IS on the phase diagrams of systems with semiconductor phases [3] and "Crystal" IS as an example of IS on substances with special acousto-optical, electro-optical and nonlinear-optical properties respectively [4]. These systems are oriented toward experts in the field of semiconductors and dielectrics materials science.

Thus, in different IS various characteristics are presented (we will further call them properties) of various substances and materials (we will further call them entities). The properties values are determined, first of all, by the inorganic substances' composition (a set of chemical elements included and their quantities, i.e. qualitative and quantitative compositions), and often the physical properties depend on the crystalline structure of the solid phase. Since the ISs on PISM are closely related to inorganic chemistry, the entities in the IS on PISM are described using a hierarchy of concepts ("system → substance → modification") in a tree form (figure 1).

If we denote the entities of the second level by the general term "substance", meaning by this term a set of discrete formations with a rest mass (i.e. atoms, molecules, and everything built of them). To describe chemical entities, three levels can be used: system, substance, and crystal (polymorphic) modification (hereinafter – modification). In addition, each subsequent level specifies (refines) information about the described chemical entity in more detail.

We give a brief definition of the basic terms used in the hierarchy of chemical entities:

Chemical system (set of elements that determine the qualitative composition) – a system formed by chemical elements. It can be described as a set of atoms forming a chemical system. More strictly, the chemical system is a combination of micro- and macro-quantities of substances

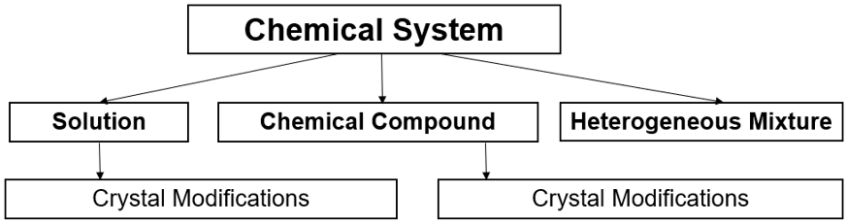


Figure 1. The top of the hierarchy of chemical entities for IS on PISM integration in inorganic chemistry.

capable of being transformed under the influence of external factors (conditions) to form new chemical compounds. For example, the chemical system, which includes the elements: copper, gallium and tellurium is denoted as Cu-Ga-Te.

Chemical compound – a substance formed when two or more atoms of chemical elements are chemically bonded together. On the phase diagram, the compound homogeneity region is separated from the region of the components or solid solutions based on them. The elements in the compound can not be separated by a simple mechanical method, but only by chemical processing, heating, electric current, etc.

Solution – a homogeneous mixture of two or more components, the composition of which under given external conditions can vary continuously within certain limits. The solid solution retains the crystal lattice of the component(s).

Heterogeneous mixture – a mechanical mixture of different components, in which under given conditions there is no any chemical interaction.

Crystalline (polymorphic) modification – a form of space organization of a solid substance.

The above chemical definitions are largely fuzzy. Therefore, it is sometimes difficult to distinguish between, for example, an ordered solid solution and a compound.

It should be noted that the description of entities and their properties in different IS on PISM occurs with varying levels of detail. So, for example, in the "Diagram" IS, the description of the most of chemical entities properties is carried out at the chemical systems level. And in the "Crystal" IS, some properties values are described at the chemical substances level (for example, melting point, solubility, etc.), and some properties values are described at the specific crystal modifications level (for example, nonlinear optical coefficients, Selmeyer's coefficients, etc.).

It is obvious that the properties values specified for chemical entities at the system level apply to all chemical substances and their modifications derived from this system. Similarly, the properties values set at the chemical substances level apply to all chemical modifications of this substance. These remarks are important in the context of formal problem domain modeling.

4. Formal problem domain description in terms of set theory

When consolidating ISs, syntactic and structural conflicts arise due to the fact that ISs use data of different syntax description and structure. In a number of ISs, relational database management systems (DBMS) are used, in another – hierarchical DBMS. Recently, ISs are often built that take advantage of JSON (JavaScript Object Notation), XML (eXtensible Markup Language) or some of its well-known applications, such as RDF for storing information. In ISs, developed a long time ago, it is often possible to find their own proprietary binary formats for storing and processing data. All this diversity of data models and presentation schemes, as well as information processing, lead to the fact that the ISs in the existing form are often incompatible with other software products. It should be noted that, initially, when IS on PISM was designed, no interaction with some external software environment was not planned at all.

It is possible to resolve syntactic and structural conflicts by introducing a general scheme (for representing information and data exchange) constructed according to the problem domain description. As noted above, to describe chemical entities, one can use three levels: system, substance, and crystal modification. The hierarchy of chemical entities, which is considered in the context of the integrated IS, is shown at figure 2.

Let's use the set theory to describe the entities of the considered problem domain, considering that each subsequent level in the hierarchy refines (complements) the object description. We denote the chemical systems set as S, the chemical compounds set as C, and the crystal modifications set as M. Then the chemical system will be denoted as s (where $s \in S$), the chemical substance denoted by c (where $c \in C$), and the chemical modification – as m (where $m \in M$). R^+ is the set of non-negative real numbers, and $R*$ is the set of R^+ extended by the element x.

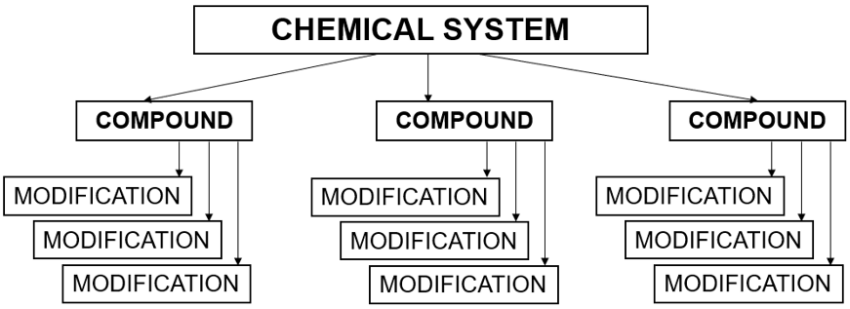


Figure 2. The top of the hierarchy of chemical entities for IS on PISM integration in inorganic chemistry.

The chemical system s can be represented as a set of chemical elements $s = \{e_1, e_2, \ldots, e_n\}$. The chemical substance c is determined not only by the set of atoms of chemical elements, but also by their quantities in the substance, solution or mixture. Therefore, substance c can be represented by a tuple (s, f), where $s \in S$, and f is a mapping of the set of atoms (chemical elements) that forms the substance, onto the set of pairs $R* \times R*$, specifying the homogeneity range of the given element in the substance, solution or mixture c. That is, f: $e_i \to R*_{\min}$, $R*_{\max}$), where $R* = R^+ \cup \{x\}$. The element x is used to denote an unknown number, since in the designation of mixtures, where the quantities of components may vary, it is common to use x to denote an unknown quantity, for example, $Fe_{1-x}Se_x$. $R*_{\min}$ and $R*_{\max}$, respectively, are the minimum and maximum concentration of the chemical element e_i in the substance c. In the case when the concentration of a particular chemical element e_i in substance c is fixed, then $R*_{\min} = R*_{\max}$. Chemical modification m can be represented by a tuple (s, f, mod), where $s \in S$, $f : e_i \to (R*_{\min}, R*_{\max})$, and mod is the string representation of the substance crystal modification adopted in an integrated IS (one of enumeration values (enum) for crystal systems: {*Triclinic, Monoclinic, Orthorhombic, Tetragonal, Trigonal, Hexagonal, Cubic*}).

5. The problem domain ontology

With the development of information technology, many ways have emerged to describe subject areas using diagrams and formal languages [5–8]. UML (Unified Modeling Language) is widely used for building diagrams (e.g., sequence or use case diagrams). As a rule, OWL (Web Ontology Language) or RDF (Resource Description Framework) are used for the problem domain formal description. We formalized the above-mentioned problem domain by means of a description in OWL language using the Protege software package.

Thus, the chemical concepts hierarchy ("system → substance → modification") is reduced to the class hierarchy: ChemicalSystem → ChemicalSubstance → ChemicalModification, as shown at figure 3. To model atoms (chemical elements), an auxiliary class named ChemicalElement is introduced, and to indicate the number of atoms in a substance, the QuantitativeElement class inherited from ChemicalElement is introduced. Basically, QuantitativeElement expands the base class definition by adding the number of atoms. To designate the crystal lattice types for solid phases, CrystalSystemType class is introduced. To provide the use of the ontology, instances of all classes are created. It is given an example of the lithium niobate ($LiNbO_3$) with hexagonal and rhombohedral crystal lattices described at all hierarchy levels. This ontology is intended primarily to illustration of the integration principles of information systems on inorganic substances properties and is subject to further development and extension. The current ontology version defined on OWL is available at http://meta.imet-db.ru/meta.owl.

In proposed ontology, there is QuantitativeElement class, which is nor a problem domain

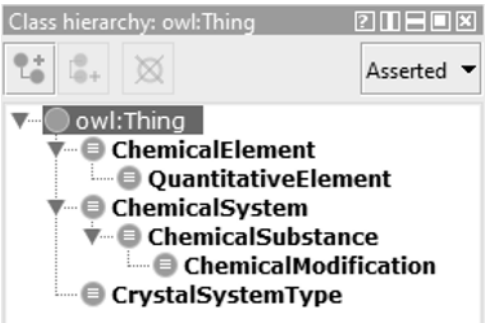


Figure 3. The class hierarchy in ontology for integrated IS on PISM.

concept not a subconcept of a chemical element. This class implementation is caused by need to specify quantitative composition of substance, i.e. the quantity of each type of atoms in the substance. We should note such QuantitativeElement definition causes some redundancy at the substance level, because the ChemicalSubstance class contains both the ChemicalElement set (through the "has_Elements" property) inherited from the ChemicalSystem class, and the QuantitativeElement set (through the "has_QuantitativeElement" property) directly defined in ChemicalSubstance class. Although it is obvious that QuantitativeElement includes ChemicalElement. It is necessary to mention the possibility of specifying incorrect (inconsistent) data in such class configuration.

It would more appropriate, instead of the QuantitativeElement class implementation, to add a list of quantities of atoms defined one level above (in the ChemicalSystem class) to the ChemicalSubstance class definition. However, the OWL does not allow operating with ordered lists; only sets (unordered sequences of unique elements) manipulation is supported, which makes it impossible to implement, for example, a list with quantitative definitions for atoms defined at a higher level, i.e. in the base class. Moreover, it is not possible to provide the same number of elements when describing the list of atoms (in ChemicalSystem class instance) and describing their number as part of a substance in ChemicalSubstance class instance in OWL.

So that was the reason of the QuantitativeElement class introduction in the attempt to guarantee the correct description of the quantitative composition of a substance, if one does not take into account the inherited set of ChemicalElement instances from ChemicalSystem class. Thus, we conclude that the OWL expressiveness may not be enough for a correct description of the problem domain proposed in section 4 of this paper.

6. The formal problem domain description using an object-oriented language

Note that if we use an object-oriented language, then there are no problems when describing the domain formalisms proposed in section 5. As a confirmation of the thesis, we consider formalization using the C# version 7.0 language (freely available at https://github.com/vicdudarev/ChemicalHierarchy).

Without considering the proposed C# implementation in detail, we briefly discuss the moment that caused difficulties in the ontological description – the transition from system to substance level – the addition of information about the quantitative description to qualitative substance composition. In the proposed implementation, the chemical system (ChemicalSystem class) is described as a one-dimensional array of type "ChemicalElement []", where ChemicalElement is the class for representing the chemical element (contains the element designation and its atomic number). At the quantitative composition description level, the ChemicalSubstance class is introduced with the ChemicalSystem base class. The

ChemicalSubstance class expands the description with the quantitative composition represented in the form of a one-dimensional array of the type "Quantity[]", where Quantity is the simplest class containing a pair of Min and Max floating-point values. Note that all necessary checks for the specified values correctness are performed in class constructors. For example, in the ChemicalSubstance class objects constructor, it is verified that the length of the quantitative description array coincides with the corresponding length of the qualitative description array inherited from ChemicalSystem. Thus, object-oriented languages advanced capabilities allow the correct implementation of the formalization proposed in section 4.

7. Entity properties representation

Having considered the formalization of chemical entities description, it is possible to proceed to a brief summary of the proposed representation of chemical entities properties. As noted, ISs to be integrated contain information on properties of chemical entities (for example, density, solubility, thermal conductivity, forbidden zone width, etc). At the same time, there are often several records to describe property values of particular chemical object in the database (DB). This is due to the different circumstances. Firstly, the information contained in the IS database can be taken from various sources, and the data differ in the most cases. This can be explained by different measurement methods, measuring equipment accuracy, substances impurity and so on. Thus, as a rule several values are stored for every substance property in different ISs on PISM. Secondly, property values for substances depend on the conditions under which the measurements were taken. For example, solubility and bandgap depend on temperature. Different properties should have different data structures which consider its nature. Moreover, the same properties in different ISs on PISM are actually functions of different number of arguments, and therefore it is impossible to offer a universal format for representing a given property for all ISs on PISM. This can be explained by the fact that with more detailed study of a particular property nature, the number of such functional dependencies on external parameters may increase. Consequently, if such a property is considered in detail in some IS on PISM, which is not yet consolidated with the integrated information system, then at the time it is included eventually in the integrated information system, there will be a problem of harmonizing the presentation formats of the specified property. Thus, it is impossible to foresee all dependencies in advance and take them into account in some common data presentation format for even a single specific property.

Considering the above-mentioned properties values representation difficulties, some mechanism is required that allows a flexible representation of property values within the integrated information system. Currently, there are a number of widely used languages describing arbitrary data formats, the most common being JSON and XML. Using these languages, it is convenient to describe various data structures because XML and JSON are cross-platform formats which are supported by most languages and libraries [9]. Nowadays information representation by means of these formats is the foundation for ensuring the interoperability between various software and hardware platforms. Thus, an increasing amount of information in modern industrial systems is represented in JSON and XML formats. The use of these formats is also appropriate in integrated systems because they are applied as the basis for Web services functioning.

To resolve semantic and structural conflicts, it is necessary to standardize the presentation formats for described chemical entities and their properties within the integrated information system in XML and JSON formats. I.e. it is necessary to develop document formats for chemical entities and their properties representation. This will allow an information exchange between nodes of the integrated information system.

8. Conclusion

The problem of information systems integration in general and ISs on PISM, in particular, is extremely urgent, since access to data sets on substances allows considering such a consolidated information source as a subject for comprehensive analysis and new knowledge discovery in chemistry.

At the first stage the integration attempts in the inorganic materials science based on the problem domain specificity are the most realistic. The proposed formal description of the problem domain — inorganic materials science – does not claim to be deep enough to satisfy the materials scientist. In each of the many materials science areas, there are a number of important peculiarities, which can be more or less taken into account when constructing ontologies based on complex taxonomies.

It is important to understand that the integrated IS implementation complexity depends on the complexity of formal problem domain description. In this sense, the proposed formal model ("system → substance → modification"), in our opinion, is an acceptable compromise between the complexity of the integrated information system implementation and the detail level of the chemical information description presented in heterogeneous IS on PISM.

Acknowledgments

The work was carried out with the partial financial support of the Russian Foundation for Basic Research, projects 17–07–01362, 18–07–00080. The work was carried out according to the government task No. 007–00129–18–00.

References

[1] Kiselyova N 2005 *Kompyuternoye konstruirovaniye neorganicheskikh soyedineniy. Ispolzovaniye baz dannykh i metodov iskusstvennogo intellekta* (Moscow: Nauka)
[2] Dudarev V 2016 *Integratsiya informatsionnykh sistem v oblasti neorganicheskoy khimii i materialovedeniya* (Moscow: URSS)
[3] Khristoforov Y, Khorbenko V, Kiselyova N *et al* 2001 *Izv. VUZov. Materialy elektronnoy tekhniki* **4** 50–5
[4] Kiselyova N, Prokoshev I, Dudarev V *et al* 2004 *Neorgan. materialy* **42(3)** 380–4
[5] Ashino T 2010 *Data Science J.* **9** 54–61
[6] Erkimbaev A, Zitserman V, Kobzev G *et al* 2012 Ontology-based problem domain modelling for data on substances and materials properties integration *Proc of XV All-Russian Joint Conference "Internet and modern society" (IMS-2012)* pp 38–47
[7] Zhao S and Qian Q 2017 *AIP Adv.* **7** 105325/1–18
[8] Connelly N, Hartshorn R *et al* 2005 *The Red Book: Nomenclature of Inorganic Chemistry* (Cambridge: RSCPublishing)
[9] Koffina I, Serfiotis G, Christophides V, Tannen V and Deutsch A 2005 Integrating xml data sources using rdf/s schemas: The ics-forth semantic web integration middleware (swim) *Semantic Interoperability and Integration* (*Dagstuhl Seminar Proceedings* no 04391) ed Kalfoglou Y, Schorlemmer M, Sheth A, Staab S and Uschold M (Dagstuhl, Germany: Internationales Begegnungs- und Forschungszentrum für Informatik (IBFI), Schloss Dagstuhl, Germany) URL `http://drops.dagstuhl.de/opus/volltexte/2005/34`