Development of the system of reactor thermophysical data on the basis of ontological modelling

I A Chusov¹, P L Kirillov¹, G P Bogoslovskaya¹, L K Yunusov¹, N A Obysov², G E Novikov², V G Pronyaev², A O Erkimbaev², V Yu Zitserman³, G A Kobzev⁴, M S Trachtengerts⁴, and L R Fokin⁴

¹Institute of Atomic Energetics «NRNU MEPhi», Moscow, Russia
²State Corporation «Rosatom», Moscow, Russia
³Private Institution «Atomstandart», Moscow, Russia
⁴Joint Institute for High Temperatures of the Russian Academy of Sciences, Moscow, Russia

Abstract. Compilation and processing of the thermophysical data was always an important task for the nuclear industry. The difficulties of the present stage of this activity are explained by sharp increase of the data volume and the number of new materials, as well as by the increased requirements to the reliability of the data used in the nuclear industry. General trend in the fields with predominantly orientation at the work with data (material science, chemistry and others) consists in the transition to a common infrastructure with integration of separate databases, Web-portals and other resources. This infrastructure provides the interoperability, the procedures of the data exchange, storage and dissemination. Key elements of this infrastructure is a domain-specific ontology, which provides a single information model and dictionary for semantic definitions. Formalizing the subject area, the ontology adapts the definitions for the different database schemes and provides the integration of heterogeneous data. The important property to be inherent for ontologies is a possibility of permanent expanding of new definitions, e.g. list of materials and properties. The expansion of the thermophysical data ontology at the reactor materials includes the creation of taxonomic dictionaries for thermophysical properties; the models for data presentation and their uncertainties; the inclusion along with the parameters of the state, some additional factors, such as the material porosity, the burnup rate, the irradiation rate and others; axiomatics of the properties applicable to the given class of materials.

1. Thermal properties of materials for nuclear engineering - the requirements for data, their structure and formats

Thermophysics belongs to the category of disciplines, in which the management of large sets of numerical data came to the forefront. It is characterized by increased attention to the raw data on the substance properties, their uncertainty and reproducibility, data handling and correlation. Recent trends in the dominance of science data management, including their storage, aggregation and analysis, led to the formation of a series of so-called X-informatics, where the X identifies a subject area for which the infrastructure is created: geoinformatics, bioinformatics, material informatics etc. The data infrastructure for the nuclear industry, by the nature of the tasks being solved, is closest to the latter, because within the framework of material informatics, there are means for storing and processing data for materials of any class with the possibility of adequately representing their properties, states, technologies, and so on. In this article, we propose an ontological approach to the
problems of representation material science data, which was developed in detail for thermophysical data for nuclear technology.

Specification of requirements for reactor materials, as well as other objects, presupposes, first of all, the choice of the nomenclature of materials and properties. Table 1 contains the minimal set of properties that can be expanded at necessity, if there is information about the phase composition [1]. In the presence of polymorphic transitions in the solid phase, data are required for each of the crystalline phases, including temperature range, crystal structure, cell dimension, heat of phase transition.

Table 1. The minimum set of materials and thermophysical properties peculiar to nuclear engineering.

| Reactor materials [main classes] | Thermal properties [minimal set] | Properties inherent in the liquid phase |
|---------------------------------|-----------------------------------|---------------------------------------|
| Nuclear fuels                   | Density, $\rho$                   | Melting point, $T_m$                  |
| Coolants                        | Heat capacity, $c_p$               | Boiling point, $T_b$                  |
| Moderators                      | Thermal conductivity, $\lambda$   | Heat of fusion, $\Delta H_m$          |
| Absorbing materials             | Thermal diffusivity, $a$           | Volumetric expansion coefficient, $\beta$ |
| Structural materials            | Emissivity, $\varepsilon$         | Dynamic viscosity, $\mu$              |
|                                 | Surface tension, $\sigma$         |                                       |

The difference between alloys and individual compounds requires taking into account data on the phase diagrams with a set of phases and phase boundaries. Expansion of the nomenclature of properties can also be related to the individual characteristics of a particular substance or material. As an example, one can point to the partial pressures of nitrogen and uranium, that are relevant for processes with nitride fuel, or the values of solidus and liquidus temperatures, which qualitatively distinguish MOX fuel ($U, Pu)O_2$ or from pure oxides $UO_2, PuO_2$.

In addition to the physical factors that determine the differences in the property nomenclature, such factors as the relevance of certain characteristics for thermal calculations, as well as the availability and completeness of the data presented in the literature, also affect the data collections creation (databases, digital libraries etc). For example, for mononitride fuel there is no data for the liquid phase, for $UC$ fuel these data are limited by density and saturation pressure, while for $UO_2, PuO_2$ oxide fuels in the liquid phase there are sufficiently detailed tables determining their temperature dependence [1]. In a large group of liquid metal coolants only for four alkali metals ($Li, Na, K, Cs$), data were available for saturated vapor properties, namely, density, heat capacity, thermal conductivity, viscosity, and sound velocity. On the other hand, for such coolants as $H_2O, D_2O$, data are available for the entire complex of thermodynamic and transport properties in both phases, including the saturation line. Quite often information is limited only by data at ambient temperatures (such as for absorbing materials $B, B_4C$), if the temperature dependence of such properties as heat capacity or thermal conductivity is ignored.

Already this brief summary allows us to draw a conclusion about the specifics of thermal data for nuclear engineering - the extreme irregularity of the data structure, determined by the differences in physical properties, as well as the different data availability. Such data is usually classified as semi-structured in contrast to structured (table-oriented data) presented in relational databases or unstructured data such as text, image etc. [2]. However, the differences in the set of characteristics for each substance are not the only possible manifestation of semi-structured data (SSD). The second feature of the irregular data structure is related to the presence of various factors (so-called extra factors), supplementing the usual data on the state, such as phase, temperature, pressure etc. Among them, there are the characteristics of the internal structure and surface, information on manufacturing technology, the influence of the environment and so on. The most significant for reactor materials is the deviation from the stoichiometry of the oxide fuel ($x=O/M-2$, where the oxygen and metal content per stoichiometric unit), the porosity of the sample, the degree of irradiation.
(fluence), the burning out of the fuel. The effect of sample preparation on thermal properties is illustrated by the information given in [1] for Be, BeO moderators. So for beryllium four characteristic states are distinguished: well-annealed polycrystalline beryllium of high purity; hot-pressed beryllium after exposure of 1000 hours at 1300 K; hot-pressed beryllium; cold-pressed beryllium. In this case, the values of the properties in states 1 and 4 can differ approximately by 1.5-2 times. Similarly, depending on the purity of the BeO sample, the properties can vary by an order of magnitude, and the characteristic parameter determining the state of the sample is the density at room temperature, which varies from 1890 to 3010 kg/m³.

For many materials, primarily structural, integral characteristic, summarizing technological factors, is the trade name. In particular, the graphite data is so strongly dependent on the production techniques, which deprives them of meaning without reference to the manufacturer and the specific grade. For example, the free database MatWeb [3] includes 378 trade names of different graphite types, the structure and properties of which vary considerably. In this case, even the indication of the grade does not fully determine the properties - it is necessary to take into account the type of raw materials, the graphitization of the initial graphite, the porosity of the material obtained, and a number of other factors.

Finally, the third indication of SSD is the frequent variation of the data type, for example the transition from a scalar to an array of values, from a constant to a function, and so on. A good illustration is the graphite discussed above, where pressing during its production results in anisotropy of physical properties. As a consequence, instead of a scalar value, it requires a pair of values (array) to determine its values in the direction parallel to the applying axis, and to the perpendicular direction.

Similarly, for pure plutonium, it becomes necessary to take into account the differences in thermal expansion along the crystallographic axes. Of 6 phases, this feature is manifested for γ-phase (rhombic face-centered at 479-592 K), where data are required for three axes and δ-phase (face-centered tetragonal phase at 724-749 K), where data for two axes are required; for the remaining 4 phases, these differences do not appear. The transition from scalar values to arrays is also required when identifying the composition of steels and alloys, where available experimental data on the concentration are specified as intervals of values (xmin ÷ xmax) or half-intervals of the type (x ≤ xmax). In the first case, it is required to specify a pair of values, in the second one a single value, but indicating that this is the upper concentration limit. A similar type of data is required when using data on the average value of α, the coefficient of linear thermal expansion, which is accompanied by an indication of the boundaries of Tmin,max, the temperature interval.

Finally, the change in the type of data accompanies the transition from liquid-metal coolants to gases (air, He) or to H2O, D2O, which may be in the liquid and vapor phases. Data for these substances are presented as functions of two variables, temperature and pressure, while for the vast majority of reactor materials (fuels, coolants, etc.), the influence of pressure may be neglected, so that any of the thermal properties can be considered as temperature dependent function or constant.

As was first noted in [4], the whole practice of working with data on the properties of substances shows the impossibility of unifying the table format for a wide range of substances without conflicting with the objective need to convey their specifics and standards of description in different sources. Data storage requires in this case the rejection of the relational database schema and the attraction of the SSD concept with the use of appropriate tools and technologies [5].

The most relevant is the appeal to such concepts in the transition from pure substances to industrial materials, where the presence of impurities, the technology of production, the conditions for sample preparation play a primary role in the presentation of data on physical and performance properties. The recent advances in the material informatics way, especially those initiated by the US program Materials Genome initiative [6], led to the selection of two main principles: the design of extensive infrastructures integrating autonomous resources, and the flexibility of the overall data structure, allowing adaptation to the domain and specific material features. In the following sections, we will look at the potential of ontologies that can provide sufficient flexibility of the data structure with characteristic variability of the properties nomenclature and data types.
2. Ontological modeling of the subject field

Last years the concept of ontology, originally emerged as a philosophical category, has moved into the field of informatics, where it defines the formal specification of concepts and relationships in a particular field of knowledge. It is believed that the simplest and most precise definition of ontology was given by Gruber [7]: "ontology is a specification of conceptualization", that is, a standardized and coded representation of knowledge in a specific subject area.

The main purpose of the ontology is that it enriches the data by adding semantics (meaning, correct interpretation) and logical relationships to them, which jointly describes "knowledge" with the possibility of its machine interpretation. Thus, ontologies can provide the necessary top-level layer, which can integrate disparate data.

The rich potential inherent in ontologies for storing and distributing data can be evaluated by comparing it with the capabilities of databases [8]. The conceptual scheme of the database, defining all concepts and data structure, serves only for those purposes that the specific database implements. At the same time, the dissemination of information through ontology is carried out in an agreed manner, that is, the transmitted data structure is publicly available and is equally interpreted in a particular community. All members of the community can use the ontology and have access to information.

With respect to the subject "properties of substances and materials," a special role is played by the possibility, on the level of ontologies, to support data schema evolution associated with the expansion of the object set, their states, factors of influence, and so on. As noted in our papers [2, 4], it is impossible to maintain their static scheme for a wide range of substances because of their specificity and differences in descriptions adopted by different teams. In general, traditional databases and ontologies proved to be complementary in their capabilities for categorizing data.

Databases are capable of structuring and storing large amounts of data with high search efficiency. Ontologies allow eliminating their main defects – the lack of common semantics and also a rigid structure, incapable of evolution.

Awareness of the enormous possibilities of ontologies for the categorization and dissemination of scientific knowledge stimulated active development of subject-oriented ontologies for the formalization of individual segments of many scientific disciplines: biomedical, chemistry, earth sciences, etc. In particular, the authors [9] developed a preliminary version of the thermophysical data ontology, which is limited to the representation of the properties of individual substances.

Below, for the domain of "thermophysics of nuclear materials" a prototype of an ontology is proposed that extends the boundaries of the previous version [9] by incorporating new objects and concepts specific to reactor materials. Among them, all the factors that fundamentally differentiate the reactor material from ordinary materials: the state of the specimen, the environmental effect, the degree of fuel burnout, and so on. For the same purposes, a whole set of identifiers is provided that can uniquely determine the composition of a complex system, as well as concepts such as a phase diagram and the crystal structure.

2.1. Prototype of the ontology “NuclThermo”.

The general methodology for building ontologies [10] adopted here assumes: 1) the active involvement of classes from previously created ontologies presented in the Web; 2) reservation of positions to include, if necessary, new objects or physical factors. Here, a large part of the conceptual framework is imported from two upper-level scientific ontologies located in the Ontobee [11], web-based data server specifically designed for ontology terms. As up-level ontologies used: Semanticscience Integrated Ontology (SIO) и Chemical Information Ontology (ChemInf). First of them provides a simple description of scientific activity. The second ontology covers common concepts of chemical and material sciences such as molecule, solution, substance, crystal structure etc. The set of concepts attracted from external ontologies has been enlarged due to new concepts reflecting the features peculiar to the reactor materials (their names in Fig. 1 are underlined).
Figure 1. (a) Top-level classes of the NuclThermo ontology
(b) Child classes with respect to Material_by_application
(c) Child classes with respect to Quality

To implement the specification of conceptualization (according to Guber [7]), we use the Protégé, open-source ontology editor, which allows one to write it in OWL (Web Ontology Language) and
visualize class hierarchy. The class hierarchy in abbreviated form for the NuclThermo project is shown in Fig. 1. In Fig. 1 (a) shows the higher-level classes reflecting the general structure of most science disciplines. First of all, the Continuant classes are allocated for objects that retain their identity in time and Occurent, for entities that have time parts and evolve with time. In turn, the class Continuant has as its heirs Independent Continuant and Dependent Continuant, covering objects related to the subject area, and their attributes. As for the class Dependent Continuant, it also has two child classes: Specifically Dependent Continuant (summarizing all qualitative characteristics of objects - size, density, etc.) and Generically Dependent Continuant (summarizing all kinds of data and the concepts required for this). The Specifically Dependent Continuant class has two child classes: Quality (the whole set of properties of objects, processes and documents), and State, a class that defines objects in terms of their physical states (gas, liquid, etc.). Thus, a sufficiently strict and universal structure is formed, adapted to the required domain. On the same Figure 1 (a) for example, the child classes illustrating the "descent" to the lower levels occupied by concepts characteristic of reactor thermophysics are shown. In particular, typical objects Material by Application (fuel, coolant, etc.), Chemical Entity (arbitrary kind of substance, solution, particles) and finally, Sample (a specific specimen, with details of its state and preparation conditions).

In Fig. 1a the child classes of Occurrent are not shown. They correspond to processes of various types: those occurring with the material under operating conditions (Burn-up, Irradiation), stages of production (Manufacturing, Testing) and research (Measurement procedure, Simulation). In accordance with the accepted hierarchy of the top level, class Continuant has been categorized into three groups, representing objects, their attributes (quality, state) and so-called information entities (document, description, Identifier etc.).

The hierarchy of materials relevant to reactor thermophysics (Material_Entity, Figure 1 (a)) is based on three classes: Material by Application, combining those materials that can be assigned to one of the five main categories (Fig. 1(b)); Chemical Entity, that is, all other materials that can not be assigned to these categories; Sample, i.e. a limited portion of a substance to be used for analysis or investigation. Only classes that inherit the Material_by_application class are shown in Fig. 1(b).

The Chemical_entity class corresponds to the concept of the upper level, not equivalent to the concept of substance or material. This category includes any atoms, molecules, ions, radicals, condensed matter with known stoichiometry (for example $U_3O_8$ or $Gd_2O_3$) or multicomponent systems such as $U$-$O$ or $U$-$Zr$-$Fe$-$O$. As for the Sample class, its child classes correspond to possible variation of a Sample, for example Pellet, Powder etc.

The taxonomy shown in Figure 1(b) clearly illustrates the possibility of its expansion with inclusion of new objects. For example, one can enter an additional class of Liquid_Fuel at the same level as Metallic_Fuel and Ceramic_Fuel, or introduce in the class of Ceramic_Fuel a child class MOX_Fuel. Another important point is the availability of a special class Sample, which allows to specify the characteristics of the sample, along with the general properties of the material. In another block (Fig. 1(c)), the Sample_Quality class is also presented, which combines the qualitative and quantitative characteristics of the sample.

The hierarchy in Fig. 1(c) includes all classes related to the characteristic of the basic entities. The Quality class has three child classes that define the properties of objects, processes, and so-called information entities. Quality (or physical properties) of objects includes thermal properties, data on phase diagrams and crystal structures (lattice type, cell dimensions, etc.). Each of the classes allows including additional child classes, for example, by expanding the list of thermal properties.

Of particular value is the Information_content_quality class, which allows performing a multifaceted assessment of the reliability and quality of information. First of all, it is the uncertainty associated with the error of an experiment, the method of processing or appraisal, the authenticity of a model, etc. The characteristics used in the appraisal of uncertainty are rather diverse. Following to the IUPAC Project ThermoML [12], it is justifiable to distinguish between standard uncertainty (i.e. standard deviation), expanded uncertainty defining interval for a given confidence level, a combined uncertainty when the measurement result takes into account the contribution of other variables, for
example, state parameters. In addition, in practice, estimates such as reproducibility, root-mean-square deviation from a fitted curve, accuracy of measurement may be used.

Class **Status_descriptor** allows you to distinguish different types of data: experimental, computational, simulation results, critical evaluated, review data, reference etc. Finally, the **Quality_descriptor** class allows a qualitative assessment of the data, summarizing the set of factors: uncertainty, completeness of the sample description, reliability of the measurement method, degree of reproducibility, etc.

According to Fig. 1(c), class **Quality**, without exhausting all objects and processes, leaves some of them for inclusion in another class, namely class **State**. Provision is made in the ontology project of eight possible states of matter, namely: gas, liquid, solid, three lines of phase equilibrium, two-phase region and critical state. Other possible states of matter (powder, colloid, ultra disperse/nanoscale) in this ontology are referred to the class **Sample**.

The hierarchy of classes that inherit the **Information_content_entity** class (child class of **Generically Dependent Continuant** – see Fig. 1(a)) reflects information objects, required for reporting data in reports, publications or databases. Collected for this purpose, the concepts perform one of three roles: the identification of objects and information entities, the presentation of documents and the formation of sets of numerical data. Most of this hierarchy is subordinate to the **Identifier** class, performing an appropriate role for different objects and by different means. Among the identifiers are the keywords, names, IDs from controlled dictionaries (**ChemSpider** and **ChEBI**), etc.

All concepts needed to represent numerical data are collected in classes that are children of the **Mathematical_Entity**. Among them are typical concepts: **number**, **quantity**, **variable** (with the separation of independent and dependent), etc. The **Unit_of_Measurement** class provides links to the units of measurement used in documents and databases. Its detailing, that is, the inclusion of units corresponding to different physical quantities (energy, temperature, density, etc.), is provided by importing terms from external ontologies, for example from **UO** on the **Ontobee** server [11].

The hierarchy of classes described above provides only a controlled vocabulary of concepts, that is, objects, properties, processes, and so on. The rich possibilities of ontology as a structure of knowledge are associated with the use of so-called object properties or associative relations, for example, between object and its identifier or object and its part. As a rule, such relationships are not too tightly connected to the specifics of the domain. Due to this, most object properties can be imported from external ontologies. When developing the **NuclThermo** ontology the most convenient source was **Semanticscience Integrated Ontology** [11]. It includes a set of 207 relationships capable of representing adequately almost all conceivable links between objects and their attributes. The table 2 shows a limited group of the most commonly used relationships from this set. Each of them is defined by means of the ID included in the unique URI for this object property. An example of using the property has_part (ID = SIO_000028) gives an assertion <molecule has part some atom>; another statement <density is quality of uranium dioxide> gives an example of using the property is_quality_of (ID = SIO_000218).

**Table 2. Object Properties set imported from SIO**

| №№ | ObjectctProperties | ID        | №№ | ObjectctProperties | ID        |
|----|--------------------|-----------|----|--------------------|-----------|
| 1  | Is related to<BR>      | SIO 000001 | 12 | Has role<BR>            | SIO 000228 |
| 2  | Denotes<BR>            | SIO 000020 | 13 | Is component part of<BR> | SIO 000313 |
| 3  | Has part<BR>           | SIO 000028 | 14 | Is covalently connected to<BR> | SIO 000334 |
| 4  | Is denoted by<BR>      | SIO 000060 | 15 | Is weakly interacting with<BR> | SIO 000335 |
| 5  | Is part of<BR>         | SIO 000068 | 16 | Has component part<BR>  | SIO 000362 |
| 6  | Is contained in<BR>    | SIO 000128 | 17 | Is described by<BR>      | SIO 000557 |
| 7  | Contains<BR>           | SIO 000202 | 18 | Describes<BR>            | SIO 000563 |
| 8  | Is connected to<BR>    | SIO 000203 | 19 | Has identifier<BR>       | SIO 000671 |
A projected ontology defines a common vocabulary of terms and some specification of their meaning and use. As a result, we have a common basis for understanding the domain and its formal model, which allows us to search for concepts and derive logical inferences. Since Prototype of the Ontology is not static and can grow with needs, it can provide an adequate representation of the structure of the evolving domain. At the same time, an expandable list of possible processes allows us to foresee the influence of a number of extra-factors as the manufacturing process, measurement method, influence of burn-up of the nuclear fuel on the value of the measured property and its uncertainty. Likewise, the concepts covered by the Sample and Sample_Quality classes provide for the features of a particular sample - configuration, impurities, size, porosity, and so on. Finally, a rich set of textual and mathematical entities provides the required adjustment of the data type, for example, to replace, if necessary, a single number with an array or interval of values.

3. Ontology-based approach to the storage and sharing of nuclear materials data

The main goal of an ontology is the so-called semantic integration, where a variety of different data sources can be aggregated with identical reference terms. On this basis the concept of ontological access to the database was born. It is believed that the databases exists independently of the conceptual layer represented by the ontology, and some mechanism provides the mapping of relational data from many sources to ontology elements. Thus, the ontology plays the role of a unique access point for the interaction between the users and the multiple data sources, that exist independently from the ontology.

When analyzing the possibilities of ontologies, the researchers considered their main strategy not to "crowd out" the database, but to create a deep connection between them, through the so-called the database-to-ontology mapping. The goal of this strategy is to use both tools (database and ontology), combining semantics with high performance when working with data. An analysis of the proposed concepts and products is given in a number of reviews [13, 14].

A chemical database ChEBI [15] gives a vivid example of the advantages introduced by linking a resource to an ontology. For each entity (substance, atom, molecule, etc.), not only chemical data is presented there, but also a taxonomy fragment with the possibility of navigation. For example, an entry for the U atom allows all actinoid and f-block element atoms to be studied at the upper level, and the uranium molecular entity (i.e. uranium compounds) on the lower level. The taxonomy of substances is related to the sub-ontology of the Role, in which entities are classified on the basis of their possible application (fuel, reagent, refrigerant etc) and chemical role, so that, along with the data for the substance, ChEBI provides a "knowledge" fragment of the subject area.

Many approaches to implementing database-to-ontology mapping can be found in the literature, including designing a new relational database by binding its tables and attributes to the underlying concepts of the ontology. With this approach, the ontology can be easily rearranged in connection with the evolution of the subject domain. However, significant re-engineering of the database structure and interface is required in order to reflect ontological changes in the conceptual scheme.

Here in order for the linking data to ontology an alternative approach was adopted, previously used in the framework of the large-scale program Materials Genome Initiative [6]. The program as a whole is focused on the categorization and analysis of Big Data accompanying the development of new materials. As an effective way of storing heterogeneous data, the so-called JSON-format (JavaScript Object Notation) was used to structure information inside a text file. The advantage of a text document is the ability to easily read and edit, accessibility for human perception, convenient form of storage and data exchange. Previously, structured text based on the XML format was proposed as a means of storing thermophysical data in the ThermoML project [12] and data on structural materials in the MatML project [16]. The JSON format, unlike XML, is less overloaded with details, simplifying the
presentation of the data structure, reducing their size and processing time. Among other advantages of the format, one can note a simple syntax, as well as the ability to write hierarchical, that is, unlimited nested structures of the "key-value" type.

The originators of the Citrination platform [17] proposed a hierarchical MIF (material information file) scheme with a description of the document structure detailing the object, properties, manufacturing technologies, data sources, etc. The top level in this hierarchy is the Chemical System object whose fields include three data groups that explain what an object is (name, ID), how it was created/synthesized and what its properties are. To every field in this object there corresponds separate JSON files that unveils its contents. For example, several JSON files correspond to the Properties field, one for each of the properties. In turn, the file with the name Property contains references to the measurement method and conditions, units of measure, the types of data used, and so on. For most of these links, separate JSON files are also prepared. Thus, a hierarchical data structure is formed that allows for the detailing of each of the concepts and/or the introduction of a new concept, not envisaged at the design stage.

However, the proposed hierarchical scheme itself, as also are the relational database, does not provide the semantics of data necessary for the implementation of the machine search and logical conclusions. Moreover, due to the arbitrary depth of the hierarchy, this problem is exacerbated - at different hierarchy levels, the fields provided by the scheme can have different meanings. Overcoming this problem, as in the case of the database, lies in linking the fields with ontology concepts.

With this aim in mind an ontology-based data manager (OBDM) must be created, which performs search and editing terms/classes of ontologies, and to associate them with JSON-documents. The role of ontologies is to introduce semantics (shared interpretation of meaning) into documents, as well as the ability to adjust the structure of JSON document data, by editing the ontology. Linking documents with ontologies allows to execute semantic (meaningful) data search (more precisely, metadata search) using SPARQL queries, which makes it possible to reveal the information of the upper and lower levels (super and sub-classes), and side connections (related terms), not knowing the data schema inside the JSON-document. Thus, the user can view and retrieve information without getting acquainted with the conceptual schema outlined in the JSON-document.

The prototype of the NuclThermo ontology contains most of the terms necessary for linking to fields (that is, to metadata) provided in the MIF scheme, see table 3. Missing concepts can be found by searching for a term across multiple ontologies from server Ontobee [11]. In other files that are subordinate to the top-level object Chemical system, references to narrow-specialized ontologies or dictionaries are possible. As an example, one can specify the ontology UO for units of measurement or database ChEBI on chemicals [11].

| Table 3. Conformity the fields of the Chemical System document with ontological terms |
|-----------------------------|-----------------------------|-----------------------------|
| Fields from MIF             | Ontology and unique code    | Class/term                  |
| uid                         | SIO_000728                  | Chemical identifier         |
| chemicalFormula             | NuclThermo                  | Molecular_formula           |
| composition                 | NuclThermo                  | Chemical_composition        |
| names                       | NuclThermo                  | Title                       |
| ids                         | NuclThermo                  | Identifier                  |
| classifications             | NCIT C 25161                | Classification              |
| source                      | NCIT C 25392                | Manufacturer                |
| quantity                    | NuclThermo                  | Quantity                    |
| properties                  | NuclThermo                  | Quality                     |
| preparation                 | NuclThermo                  | Sample_history              |
| subSystems                  | NCIT C 45313                | Part                        |
| references                  | NuclThermo                  | Reference citation          |
| contacts                    | SIO_000498                  | Person                      |
The data processing technology presented by the MIF scheme is described in detail in the article [17], devoted to the Citrination project. In general, there are several ways to efficiently store and access data. One of them is the storage of JSON-documents in the file system, like those that are part of the Hadoop ecosystem, which unites libraries and utilities of Big Data technology [18]. In this case, the search for data can be implemented by means of a search engine, for example, ElasticSearch. Complex queries against large sets of files with semi-structured data can be performed using a search engine. Another way of storing is related to the capabilities of the so-called No SQL databases [19]. Unlike relational, structure of a No SQL database is not regulated, i.e. one can add an arbitrary field in a separate line or document, without first declaring the structure of the entire table. Special class of No SQL, the document-oriented databases designed for storing and managing hierarchical data are particularly appealing for use as JSON store. Such databases allow even for a rather complex structure to find the path to the requested data. Finally, it is possible to store and access the JSON-documents collection using Big Data platform Apache Spark [20]. The included Spark SQL module provides an opportunity for SQL-queries to the content of structured JSON documents, which is the main advantage of the proposed technology. Another feature that determines the efficiency of Spark when storing and processing data is the ability to support interaction with a variety of storage types, from HFDS (Hadoop Distributed Files System) to traditional databases on local computers.

As a result, the technologies that have emerged in connection with the common Big Data problem turn out to be the most effective for managing the materials data. In so doing the great diversity of types and sources of data peculiar to material science plays a decisive role. On the other hand, the use of ontology (as native NuclThermo so previously created ontologies and dictionaries) allows us to clearly define the meaning of a vast set of terms. Besides it the ontology provides the taxonomic and associative relationships between concepts, so that a data warehouse acquires the distinctive properties of a Knowledge Base.

4. Conclusion
The concept of a system for storing and processing heterogeneous data on the thermophysical properties of nuclear power materials is formulated. The concept is based on the use of subject-oriented ontology, which provides the principles and technology of data warehouse design. It is developed the prototype of the NuclThermo ontology, which includes the basic domain concepts: materials, their properties, factors of influence, types of data. To store and disseminate data, it is suggested to use text documents in JSON-format, suitable for supporting hierarchical data with a flexible (irregular) structure. Linking documents with ontologies allows to execute a semantic (meaningful) search for data in the conditions of permanent occurrence of new materials and concepts.

References
[1] 2006 Thermophysical Properties of Materials for Nuclear Engineering: Tutorial for students of specialty «Nuclear Power Plants». Second revised and augmented edition, edited by Kirillov P L (Obninsk, Institute for heat and mass transfer in nuclear power plants), 182 PP
[2] Yerkimbaev A, Zitserman V and Kobzev G 2008 The Role of Metadata in the Creation and Application of Information Resources on the Properties of Substances and Materials. Scientific and Technical Information Processing 35 247
[3] MatWeb. Material Property Data http://www.matweb.com
[4] Erkimbaev A, Zitserman V, Kobzev G and Fokin L 2008 The logical structure of physicochemical data: Problems of numerical data standardization and exchange. Russian Journal of Physical Chemistry A 82 15
[5] Abiteboul S, Buneman P and Suciu S 2000 Data on the Web: From Relations to Semistructured Data and XML (Burlington, Massachusetts: Morgan Kaufman Publishers)

[6] Anubhav J, Persson K and Ceder G 2016 Research Update: The materials genome initiative: Data sharing and the impact of collaborative ab initio databases. APL Mater 4 053102

[7] Gruber T 1993 A translational approach to portable ontology specification Knowledge Acquisition 5 199

[8] Uschold M 2015 Ontologies and Database Schema: What’s the Difference? Applied Ontology 10 243

[9] Erkimbaev A, Zitserman V, Kobzev G and Kosinov A 2015 Linking ontologies to databases on the properties of substances and materials Nauchno-tehnicheskaya informatsiya. Seriya 2. Informatsionnyie protsessyi i sistemyi 12 1

[10] Uschold M, Gruninger M 1996 Ontologies: Principles, methods and applications Knowledge Engineering Review 11 93

[11] Ontobee: A linked data server designed for ontologies http://www.ontobee.org

[12] Frenkel M 2009 Global Information Systems in Science: Application to the Field of Thermodynamics J. Chem. Eng. Data 54 2411

[13] Laallam F, Kherfi M and Benslimane S 2014 A survey on the complementarity between database and ontologies: principles and research areas Int. J. Computer Applications in Technology 49 166

[14] Martinez-Cruz C, Blanco I and Vila M 2012 Ontologies versus relational databases: are they so different? A comparison Artificial Intelligence Review 38 271

[15] Chemical Entities of Biological Interest (ChEBI), a freely available dictionary of molecular entities focused on ‘small’ chemical compounds http://www.ebi.ac.uk/chebi

[16] Kaufman J and Begley E 2003 MatML. A DataInterchange Markup Language ADVANCED MATERIALS & PROCESSES/NOVEMBER, 35-36.

[17] Michel K and Meredig B 2016 Beyond bulk single crystals: A data format for all materials structure–property–processing relationships Materials Research Society Bulletin 41 617

[18] Apache Hadoop https://hadoop.apache.org/

[19] Sugam S, Tim U, Wong J, Gadia S and Sharma S 2014 A Brief review on leading Big Data models. Data Science Journal 13, 138

[20] Apache Spark http://spark.apache.org/docs