Ontology-based methods of thermophysical data integration

A V Kosinov, A O Erkimbaev, V Yu Zitserman, G A Kobzev
Joint Institute for High Temperatures of the Russian Academy of Sciences, Izhorskaya 13
Bldg 2, Moscow, 125412, Russia
E-mail: vz1941@mail.ru

Abstract. The article describes new technologies and tools used for drastic upgrading the database THERMAL, one of the oldest thermophysical databases. The database upgrade project provides for a significant expansion of its volume and service capabilities, as well as the possibility of permanently updating the conceptual scheme with the advent of new objects and notions. The developed technology ensures the conversion of old and newly loaded documents to a common semantically integrated infrastructure. In accordance with the project, the whole process of creating data and their further support includes three stages: conversion of all data, regardess of format, to a structured JSON document; creation of the ontology repository and the binding of terms from a JSON documents with the classes of ontologies for semantic integration; tuning of the high-performance computing platform to manage an extensive collection of JSON files and perform complex searches.

The general trend of the last decades, typical for many natural sciences, is the maximum orientation of researchers to work with digital data, including the procedures for their collection, storage, systematization and processing. All such disciplines (for example, Earth sciences or medicine) are now generally referred to as data-intensive (data-dominated, data-centric) sciences. There was even a special term, fourth paradigm [1,2], by which it means work with data, while the previous three covered experiment, theory and simulation. Thermophysics (first of all, bearing in mind the acquisition and processing the substance properties data) should also be assigned to this category, however, taking into account certain specificities. With regard to thermophysics (as opposed to the Earth sciences or medicine), the source of data is the increasing publication flow, reflecting the synthesis of new substances, their laboratory study and modeling. In doing so, the data volume is determined not so much by the number of objects studied, as by the unlimited variety of conditions for synthesis, measurement, morphological and microstructural features, and so on. Therefore, of the three defining properties of Big Data, “3V-Volume, Velocity, Variety” [3], is the last one, that is, the set of data types and sources, plays a decisive role.

With regard to data on the substance properties, the concept of their diversity (or Variety) is associated with differences in types of sources, data formats, as well as frequent variations in the nomenclature of characteristics for different categories: pure substances, solutions, composite materials, nanostructures, and so on. The need for new concepts may also be associated with the inclusion in the analysis of new properties with the expansion of the range of parameters, accompanied by chemical reaction, phase separation, etc.

The need to work with the growing volume of publications, network documents, experimental
and simulation data determines the main direction in the work with data—the transition from stand-alone resources to a unified infrastructure capable of data integration. Integrating different sources means their virtual connection, with a common interface and a single query system. The purpose of such an infrastructure is to aggregate, store and distribute data of various formats, both structured (from relational databases) and unstructured, such as texts or images. The specific problem to be solved during the creation of each infrastructure includes the standardization of formats and types of data, metadata, criteria for entering and archiving data, protocols necessary for the free exchange of data.

Awareness of this issue has stimulated large-scale research and development in order to restructure the entire data infrastructure, focusing on the growth trends of its data with a variety of structures and models [4–6]. The project of the corresponding infrastructure as applied to substance thermophysical properties was proposed by authors [7]. It was based on using the Big Data platform Apache Spark [8] in combination with ontology-based data management.

The Apache Spark platform is responsible for storing heterogeneous data, accessing and analyzing it. The role of ontology in the integration process is to support a single, recognized in the professional community vocabulary of all concepts, expanded by logical connections and axioms. In encoded form, the ontology, describing knowledge with the possibility of its machine interpretation, becomes a control superstructure capable of ensuring semantic integration of heterogeneous data.

In this article, the previously proposed infrastructure [7] is used to radically update and expand the thermophysical database THERMAL, one of the oldest databases created at the Joint Institute for High Temperatures (https://www.jiht.ru). The database THERMAL appeared in the mid-70s of the last century during the formation of a special Institute department, called the Thermophysical Center (TPC) IVTAN [9]. The subject categories of THERMAL were quite broad, with coverage:

- almost all types of inorganic substances;
- thermophysical (thermodynamic and transport) properties, with simultaneous invoking related characteristics (electrical, optical, mechanical, etc.), as well as some parameters at the atomic-molecular level;
- abstracts from an extensive archive of printed documents.

As a result, THERMAL served as a documentary database, referring to the data themselves to an extensive archive of printed copies, later replaced by files with full texts of publications. Such breadth of substances and properties range necessitated the imposition of a number of subject and service limitations:

- data were limited to the properties of individual substances, without the use of mixtures and solutions, with rare exceptions specifically specified in the guidance documents (e.g., alkali metal alloys);
- organic substances data were attracted only at a small number of carbon atoms (up to 2), for example, C2H6, CaC2, C2N2, etc.);
- all data reflecting the effect on the properties of manufacturing technology (doping, irradiation, heat treatment), sample configuration, surface condition, environment, and similar factors were excluded;
- only bibliographic data plus information about the substance (formula and compound class), names of properties and phases were stored in the database according to the detailed thesaurus.

As a result, THERMAL served as a documentary database containing only links to data per se in an extensive archive of hard copies, later replaced by full-text file collection.
An analysis of the 40-year period of maintenance of the database when compared with the technological level now achieved in data-intensive disciplines revealed an urgent need for its radical restructuring with the expansion of the topics and functions. In accordance with the previously adopted scenario [7], such a restructuring involves the solution of two key tasks:

- Transition from an autonomous (isolated) database to an infrastructure capable of working with resources of different formats and content: tables, texts, codes, images, Web-documents, etc. This alone allows you to dramatically expand the amount of warehouse by combining heterogeneous data: structured and unstructured, experimental and calculated, “raw” and recommended.
- Creation of system configuration tools for the subject area, that is, a certain type of substances with their characteristic nomenclature of properties and identifiers. It is this solution that will eliminate the burdensome restrictions imposed on the content and functions of the original version of the THERMAL.

The solution of the second task (adaptation to the frequently changing data structure) presents the greatest difficulties for an adequate representation of physical properties. In addition to the obvious need to take into account the peculiarities of technology and the sample for the characteristics of industrial materials, the need for variation in the data structure also arises for simple substances. Suffice it to mention H$_2$O, where in the supercooled state there is a stratification into two types of liquids with the presence of a second critical point, or N$_2$, capable of forming a polymer phase at high pressures. The problem of complex and frequently changing data structure on the properties of a substance has been discussed in the literature for a long time [10–12]. But only now there appeared adequate means and technologies for its solution [6,7,13].

The solutions used here are [7] are based on the joint application of the Big Data tools and ontological models for integration and operational work with data of different types and changing structure. The first step of the THERMAL update project is the conversion of source documents into structured text in JSON format, one of the most convenient for data and metadata exchange [14]. The advantage of a text document is the availability for human perception and editing, a convenient form for storing and exchanging arbitrary structured information. It is also important that the JSON format is a working object unit of some platforms, in particular Apache Spark, allowing you to organize the exchange, storage and queries for distributed data. A structured text that provides storage of numerical data has already been used with success in the ThermoML project [5] for the global exchange of thermal and thermochemical data. In our project, the bibliographic records of the old version of the database, tabular data (in CSV or XLS formats), texts of publications in PDF format, hyperlinks to thematically related web resources, etc. are used as source documents to be integrated.

The second stage of the project involves the building of an ontological repository, including ontologies that cover the terminology and logical connections of certain subject areas: thermophysics, structural materials, nanostructures, etc. For them, as a rule, suitable versions are already available on the Internet, on the basis of which one can create and maintain your own domain-oriented ontology [7].

It is also mandatory to include in the repository some upper-level ontologies containing basic types and relations for describing scientific activity, for example Semanticscience Integrated Ontology, presented on the portal Ontobee (www.ontobee.org).

The role of ontology is to introduce semantics into documents, as well as the possibility to correct the data structure by editing them. The ontology is based on a dictionary of concepts in the form of a taxonomy, supplemented by logical relations that determine their connections and possible limitations. Its main purpose is to ensure the interoperability of several resources, that is, consistent within the subject area the use of identical vocabulary and relationships, facilitating
the exchange of information. When a user searches for data, it opens up the possibility of expanding or narrowing the query, switching to related concepts and navigating through all levels of the hierarchy of classes. Linking documents with ontologies allows you to conduct a semantic search, revealing information of the upper and lower levels (parent and child classes) and side links (related terms), without knowing the source data schemes. The essence of the proposed technology [7] is to link attributes (keys), as well as possible values (objects) with the concepts of ontologies, see figure 1. An important point is that this operation is reflected both in the content of the literal data of the ontological model, and in changing names of the attributes/keys of the data sets in JSON format. Such a procedure provides the required integration and the possibility of modifying the data structure as new objects or concepts arise.

The proposed structure suggests the presence of two related data storage systems:

- repository of ontologies and metadata models stored in a RDF store of type Virtuoso [15];
- data store as JSON files and other formats (PDF, JPG, XML, CSV, XLS etc).

This provides a solution to both problems that make up the content of the THERMAL restructuring project: the integration of heterogeneous resources and the constant adjustment of the data structure.

Finally, the third phase of the project is to move to using the Big Data platform Apache Spark. Its advantage under the circumstances is associated not so much with high performance in parallel computing, as with great attention to handling data, including its storage, processing and analysis in a distributed environment. Use the high performance computational platform Apache Spark provides the ability of productive handling the JSON-format files. Among other features, it is distinguished by the presence of built-in libraries for analytical processing, including the Spark SQL library for organizing SQL (Structured Query Language) queries, through which one can access the content of structured JSON documents. It is the ability to organize SQL queries to large data sets plays a key role in the task of their integration. In so doing browsing the ontology repository (search for a term, navigation through classes, etc.) is performed using the query language SPARQL (recursive acronym from English SPARQL Protocol and RDF...
Query Language) whereas SQL queries are made to the content of documents in the Spark SQL environment. The effectiveness of Apache Spark in storing and processing data is also determined by its ability to interact with many types of storage up to including traditional databases on local computers.

All the functional management of the infrastructure inheriting the old version of the THERMAL is based on software packages and programming languages (JAVA, Python, Scale, PHP) in the Web environment. Currently, the selected solutions are being checked, the modules of the future system are being developed and tested. Figure 2 shows the scheme of management and operation of the platform.

![Figure 2. The scheme of uploading data from the THERMAL into a JSON file and managing it in an Apache SPARK environment.](image)

The proposed technology and data management scenario provide a number of significant advantages of the new version of database THERMAL. First of all, a collection of JSON documents using ontological terms gives an unambiguous interpretation of the content and type of data, both by users and program agents because the ontology term associated with a key or value in the body of a JSON file has an common definition and properties. For example, it is justified to describe references to various file types (text, graphics, spreadsheets, etc.) through key terms from ontologies to describe data formats. Second advantage ability to work with data sources when active access is denied or physically restricted. In this case, uploading data to an external JSON file provides separate storage of data and their full description through the use of an ontological models. These technologies, provided by Apache Spark, open unlimited possibilities for performance and variety of work with complex data types on the substance properties.

References

[1] Agrawal A and Choudhary A 2016 Perspective: Materials informatics and big data: Realization of the “fourth paradigm” of science in materials science APL Materials 4 053208
[2] Erkimbaev A O, Zitserman V Yu and Kobzev G A 2017 The intensive use of digital data in modern natural science Automatic Documentation and Mathematical Linguistics 51 201-13
[3] Definition from Techopedia URL www.techopedia.com/definition/29176/3vs-big-data
[4] Thanos C 2013 A vision for global research data infrastructures Data Science Journal 12 71–90
[5] Frenkel M 2009 Global information systems in science: Application to the field of thermodynamics J. Chem. Eng. Data 54 2411–28
[6] O’Mara J, Meredig B and Michel K 2016 Materials data infrastructure: A case study of the citration platform to examine data import, storage, and access JOM – The Journal of The Minerals, Metals & Materials Society (TMS) 68 2031
[7] Erkimbaev A, Zitserman V, Kobzev G and Kosinov A 2018 Properties using big data technologies and domain-specific ontologies Data Analytics and Management in Data Intensive Domains. DAMDID/RCDL 2017 (Communications in Computer and Information Science vol 822) ed Kalinichenko L et al (Cham: Springer)
[8] Apache Spark project web site URL http://spark.apache.org/docs/
[9] Gorgoraki E A, Kraevskiy S L, Trahtengerts M S et al 1977 Automated information retrieval system of the Thermophysical center IVTAN (in Russian) Reviews on thermophysical properties of substances #4, pp. 3–129 (Moscow, Russia: IVTAN)
[10] Abiteboul S and Buneman Pand Suciu S 2000 Data on the Web: From Relations to Semistructured Data and XML (Burlington, Massachusetts: Morgan Kaufman Publishers)
[11] 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 15–25
[12] Eletskii A V, Erkimbaev A O, Kobzev G A et al 2012 Properties of nanostructures: Data acquisition, categorization, and evaluation Data Science Journal 11 16–39
[13] Knoblock C and Szekely P 2015 Exploiting semantics for big data integration Ai Magazine 36 25–38
[14] The JSON Data Interchange Format. Standard ECMA-404. 1th edition URL http://www.ecma-international.org/publications/files/ECMA-ST/ECMA-404.pdf
[15] Open link Virtuoso official website, URL https://virtuoso.openlinksw.com/