

Data Curation Approach to Management of Research Data. Use Cases for a Upgrade of the Thermophysical Database THERMAL

Andrey Kosinov¹, Adilbek Erkimbaev¹, Geirgy Kobzev¹, and Vladimir Zitserman¹

¹ Joint Institute for High Temperatures, Russian Academy of Sciences, Russia
vz1941@mail.ru

Abstract. Procedures are considered to support extensive archives of digital data called “data storage”. Particular attention is paid to the support of scientific data. It is shown that the activities aimed at updating the thermophysical THERMAL database correspond to the approaches provided by the “data curation”. A communication system for metadata with external ontology is proposed. The new version of metadata provides the possibility of multilateral assessment of the origin, quality and status of scientific data. It is shown that the use of new metadata provides a significant increase in the value of these studies.

Keywords: Research Data, Data Curation, Data Quality, Thermophysical Database

1 Introduction

Digital data uploaded to repositories or databases require permanent procedures that guarantee safety, quality top-level and enduring access to data. The set of such procedures is the subject of particular activity of managing the use of data called **data curation**. This term is rarely used in the Russian literature, although all the necessary actions for data integrity and management are of course performed to support digital repositories or databases. The meaning and content of the concept of curation of data can be revealed by referring to the history of its appearance. This concept originates from the Museum's practice, which is traditionally based on the curator's work on preservation, renovation and description of exhibits.

As for the term “data curation”, apparently, it first appeared in the article by Diana Zorich [1], which pointed to the common problems facing libraries, museums and research centers involved in supporting digital collections. According to [1], digital archives, as well as their supporting tools (vocabularies, thesauruses, metadata) should be regularly monitored and updated for data consistency, maintaining its quality, availability, etc., and activities in this direction is the essence of curating process.

Oddly enough, digital data is subject to erosion, as are physical artifacts, manuscripts or museum exhibits. It can be related to the use of outdated metadata, terminology, dictionaries, formats, software, as well as the absence of references to more

Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

relevant documents or external resources. By analogy with engineering, it can be considered as technological obsolescence (or deterioration) of the data structure, file formats, software, etc. An unrecoverable failure of storage media (bit rot in IT slang) during data storage is possible in parallel with content obsolescence.

History of the “Digital Curation” concept and its gradual adoption in data manager’s community is considered by [2]. In particular, the concept of “data curation” here is clearly separated from the narrower and service-oriented concepts: “archiving” and “preserving”. Unlike the latter, curation involves not only the preservation and maintenance of digital storage, but its indispensable enrichment by expanding the functions and content. For example, a common and effective way of enriching content is to place it in a wider context by linking the data set to thematically related resources, so called **Contextualizing**.

Among the main objectives of data curation, as a rule, such are mentioned as their storage, description, safety measures, the so-called cleaning, that is, monitoring and restoring quality, as well as a number of other measures. The expanded definition of the Digital Curation Center [3] covers all activities related to data management, starting with data creation, digitization, documentation, and accessibility and future reuse. The detailing of these processes, carried out in [4], allowed us to identify about 50 curation practices, most of which also fall into such categories as data preservation, data cleaning, and finally, description in terms of the complex structure of metadata.

1 Curation of Research Data

The purpose of this report is to discuss the specific recipes foreseen in the framework of digital curation in the implementation of the project for updating the Thermophysical Database THERMAL. The project considered in the report at the previous conference [5] includes: a significant expansion of the database volume due to the rejection of the restrictions adopted at the stage of creation in the 70s of the last century; creation of tools for flexible variation of the data structure, reflecting the uniqueness of objects and their characteristics; transition to a new platform that allows you to store and process data of various structures and formats. A significant part of the activities performed during the project, in fact, refers specifically to the data curation, as it comes down to checking and correcting old documents in accordance with the newly adopted format, vocabulary and requirements for data completeness and quality.

In general, curation refers to digital objects of arbitrary origin and kind. Therefore, numerous measures for data preservation such as regular back up, defect detection at the bit level, overcoming technological obsolescence of hardware or file formats are applicable in all cases regardless of the content. In solving scientific problems, the curation process provides not just conservation, but confirmation and reliable expansion of previous data of the experiment or simulation. On the contrary, the absence or poor quality of the curation process inevitably leads to the loss, distortion or misinterpretation of data.

A brief list of features and capabilities of “data curation” as applied to e-Science (or Data-intensive science) was given by the Digital Curation Centre [6]. In this list, the specificity of scientific data, to a certain extent, is taken into account at all stages

of the curation process. For example, long-term data storage may require replacement of obsolescence of storage devices, which has already been encountered in astronomy. At the stage of data cleaning (that is, data correction and updating), it is important to establish links between different versions of evolving datasets or between primary and secondary data. However, the most noticeable specificity of data curation is manifested in their description, that is, in the composition and structure of metadata.

2 Metadata (Update and Extension)

In general terms, metadata document the context and record information about how the data was obtained and what processing and verification procedures were performed during the retention period. There is an extensive literature on scientific metadata and their use in various disciplines [7–9]. Metadata, accompanying subject information, allow you to: identify a dataset with its position in the repository; define access rules; describe the logical structure and data formats; to ensure the operation of various data analysis tools. Metadata standards for different disciplines and types of documents are collected in the catalog (rd-alliance.github.io/metadata-directory/) and the “Disciplinary metadata” section of the Digital Curation Center (www.dcc.ac.uk/resources/metadata-standards). Both mentioned sources contain also references to domain-agnostic standards for formal description of digital resources (e.g. the **Dublin Core metadata set**), or for the identification and citation of digital resources (e.g. **DataCite Metadata Store**). Metadata for thermophysical properties (ThermoML), characteristics of ordinary materials (MatML) and nanomaterials are described in detail in [7, 10–12].

Regardless of the subject area, scientific metadata must satisfy a number of requirements that guarantee sufficient completeness and accuracy in the presentation of each data set. Relevant elements should provide: unambiguous identification of the object of study; presentation of information about the source and data acquisition method (research method, equipment, program code etc); uncertainty and data quality information; linking with controlled vocabularies or ontologies; the possibility of flexible adjustment to the features of the object and its characteristics. The expansion of metadata carried out in the updating the database THERMAL, provides for the implementation of each of these requirements, see Table 1.

First of all, in the progress of curation, the possibilities of identifying objects are expanded, which can now include, along with inorganic substances, complex organics, natural and industrial materials, and so on. The “**Identification**” metadata element provides, along with a stoichiometric formula, the use of several common names (synonyms), as well as links to publicly available databases. The pointer to the database and the corresponding identifier uniquely identify the object, providing, in addition, access to information that complements the information stored in THERMAL. As an example in Fig. 1 the identification of the compound called **epoxyethane** (oxirane, ethylene oxide) in the old and new versions of the database is shown.

Table 1. A comparison between old and updated metadata versions

Old metadata set	New metadata structure		
Unique record ID	Unique record ID		
	Data type [bibl, full-text, factual]		
	Data status [experiment or simulation, predicted, critical evaluated, recommended, stale]		
	Research type [experimental, theory, simulation, review]		
Source	Provenance	Source [bibl, database, external agency]	
		Data origin [method, equipment, software]	
Abstract			
Stoichiometric formula	Identification [common names, stoichiometric formula, public database ID]		
Substance class	Linking to ontology classes [Linking to sub-classes of the Chemical_entity]		
Properties	Linking to ontology classes [Linking to sub-classes of the Quality]		
Properties type			
Phase	Linking to ontology classes [Linking to sub-classes of the State_of_matter]		
Phase transition	Linking to ontology classes [Linking to sub-classes of the Transition]		
	Data Quality	Uncertainty [type, value]	
		Data quality attributes (timeliness, reliability, currency, completeness etc)	
	Data Features [SubstanceFeatures, Sample, Influence Factors]		
	References	Full-text	
		Tables or equations	
		External documents (from Web or Server)	

Entry in the old database version

The compound identification in the new database version

Common names	epoxyethane; oxirane;ethylene oxide
Formula	C2H4O
DBASE ID	ChEBI: 27651; CAS_RN:75-21-8

Fig. 1. Identification of the substance “ethylene oxide” in the old (top) and new (bottom) versions of the database.

Another example is the natural mineral **mullite**, where variations of the elemental composition allow the use of several stoichiometric formulas (for instance, $\text{Al}_6\text{Si}_2\text{O}_{13}$ и Al_4SiO_8), and the exact identification is provided by the record (URL: www.webmineral.com/data/Mullite.shtml#.WaMcG_hJaUk) in the mineralogical database WEBMINERAL.

A more complete identification is provided by linking metadata with ontology classes, which includes entities that reflect the types of objects, their states and properties, Fig. 2.

In particular, **chemical_entity** identifies types of substances or materials, focusing on the systematics adopted in chemistry (elements, oxides, acids ...), as well as on categories determined by properties or by application (polymer, solution, mixture, refrigerant, fuel ...). A pointer to subclasses in relation to the class **mixture** allows to identify binary and multicomponent alloys and solutions, for example, such relevant objects in thermophysics as air, humid air, combustion products, etc. The **State_of_matter** class allows you to detail the phase and type of the crystal lattice based on an extensive hierarchy of child classes.

Similarly, linking to classes that inherit the **Quality** and **Transition** classes reflects the rich variety of physical properties inherent in an object and the phase transitions that occur in it. It is essential that linking of metadata to ontology during the curation provides unambiguous interpretation of terms and concept, and through editing of ontology, the possibility of flexible adjustment in connection with the emergence of new objects and concepts. For example, concepts such as **second critical point** [13],

topological insulator [14] or previously unknown allotropic forms of carbon (metallic carbon, T-carbon) have recently been included.

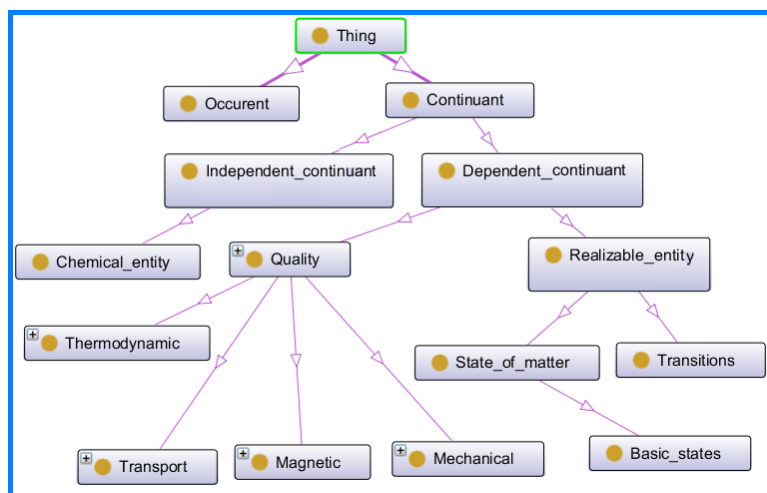


Fig. 2. Fragment of the ontology – top level classes

3 Data Quality and Features

New curatorial perspectives have emerged with concepts such as “**Data Quality**” and “**Data Features**” in the metadata set, see Table 1. Alternative choices for evaluation the scientific data quality are discussed in detail in article [12]. It was shown that the best way to assess the research data quality was to combine the traditional uncertainty assessment with a certification procedure based on several quality attributes, every of them represents some aspect of quality. It is necessary to select a specific metric for the multidimensional certification of the dataset and the data compliance indicator for each of the quality attributes. In many cases, it is useful to use a domain-agnostic metrics reflecting quality factors that are important to data consumers, for example accuracy, timeliness, reliability, completeness, relevancy, interpretability etc. Such an approach to quality certification is most justified in interdisciplinary projects, for example, when integrating thermal data with the performance characteristics of structural materials.

However, when you update the thermophysical database THERMAL, you can reduce the number of quality attributes. Data certification proposed by the authors [15] in relation to physico-chemical properties is based on the use of three attributes: completeness of information about the state and preparing the sample in the experiment; completeness of the method and measuring instruments description or codes and data processing in the case of simulation; the consistency of the numerical data with ground rules and regularities as well as with previous fairly reliable measurements. Combined with uncertainty assessment, certification identifies three main aspects of

data quality: accuracy, completeness, and consistency. The technique provides a generalized evaluation of the data set, by assigning each of the attributes of the quality level (high, medium and low), focusing on compliance with the requirements of completeness and consistency. The expert gets the opportunity, based on this data curation technique [15] select the data set of top-quality by assigning them the special **Recommended** status (for what to use the **Data Status** element). The information needed to assess the data quality allows it to be carried out only for two types of data (Full-text and Factual) specified in the **Data Type** element, Table 1. At the same time, for unstructured data in the form of the text of an article (full-text data) it is justified to conduct only certification with indication of quality attributes, but without its own assessment of uncertainty.

Table 2. Examples of the non-standard data sets

Data Set Title	Data Feature
Amorphous polymeric nitrogen-toward an equation of state	SubstanceFeatures [amorphous, polymeric]
Melting point of high-purity germanium stable isotopes	SubstanceFeatures [stable isotopes]
Relationship between changes in the crystal lattice strain and thermal conductivity of high burnup UO ₂ pellets	Sample [pellet] Influence Factors [high burnup]
Study of near-critical states of liquid - vapor phase transition of metals by isentropic expansion method of shock-compressed porous sample	Sample [porous]
Thermophysical properties of liquid Co measured by electromagnetic levitation technique in a static magnetic field	Influence factors [field]
Phase diagram of water under an applied electric field	Influence factors [field]
Shock compression of preheated molybdenum	Influence factors [prehistory]

The concept of **Data Features** in a metadata set is based on other data evaluation criterion. It allows you to select those data sets where there is any deviation from the standard (i.e. an anomaly) in the characterization of the object or its properties. The well-defined specificity (features) that distinguishes one dataset from another allows you to overcome the inevitable contradiction between structured data and poorly formalized information hidden in context. As can be seen from the Table 1, “**Data Features**” element includes three groups of the features: SubstanceFeatures, Sample,

Influence Factor. The first allows you to extend the traditional substance identification, indicating the isomeric form, nonstoichiometry, isotopic composition, etc. The indication **Sample** includes pointers to the features of the sample: shape, size, surface condition, prehistory, etc. Finally, the **Influence factor** sign includes pointers to external factors that determine the experiment and properties of the substance: external field, mechanical load, environment, radiation, etc. Some examples of non-standard data sets from Table 2 illustrate the signs defining the specifics of a substance and its properties. In so doing the specificity can be attributed to any data set of the three types indicated in the **Data Type** element (Table 1), in contrast to the quality assessment, depending on the type of data.

4 Data Cleaning and Preservation

The update procedure of the THERMAL database, includes (along with the volume expansion) revision of old records based on the new metadata system. As a result, the data curator needs to check the completed conversion. This activity, called “**Data Cleaning**” (or cleansing), involves the detection and correction of “dirty”, that is distorted or incomplete data [16]. Data pollution occurs for various reasons, among which may be distortions of old records (input errors and duplication, incorrect data distribution by fields) and errors when using new metadata to determine the object and properties, see Table 1.

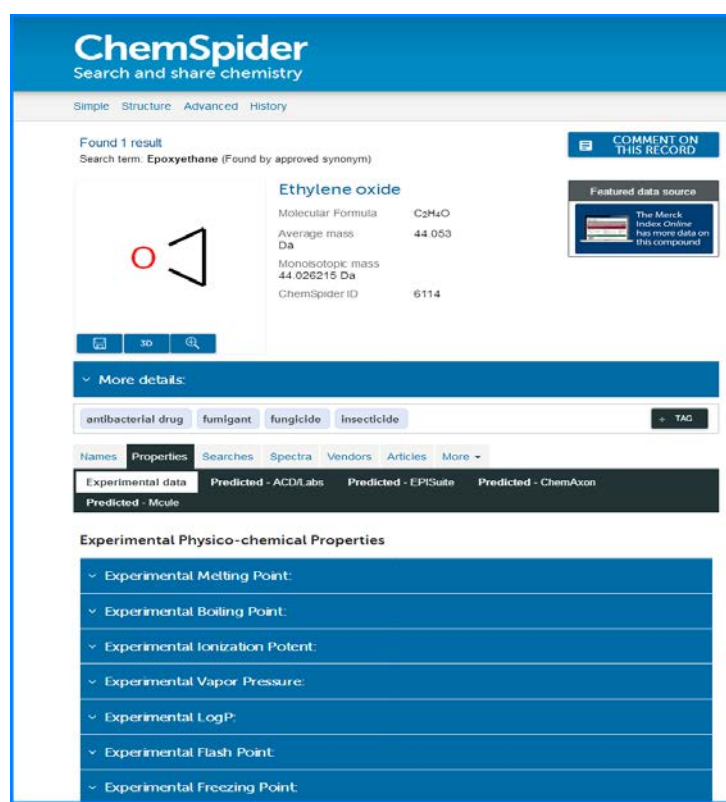
Previous data were entered without the use of controlled dictionaries, so the most important task of cleaning is to eliminate ambiguity in terms and concepts, subjecting them to ontology classes. For example, earlier a whole set of lexical elements (including English and Russian terms) was associated with the concept of dynamic compressibility, namely: **Hugoniot**, **Hugoniot data**, **Hugoniot adiabat**, **shock Hugoniot**, **shock adiabat**, **shock compression**, **release isentrope** etc. Linking this whole set of terms with a single ontology class (**Dynamic Compressibility**) eliminates synonymy and dramatically facilitates semantic search. The same procedure requires that the names of substances used in different records now appear in each record as a set of common names, which eliminates search losses.

In addition to linking and unifying terms, data cleaning also provides for the correction of content, first of all, the correction (or fixation) of obviously obsolete numerical data. At the same time, old records with correct filling of fields have historical value, even with outdated data. Therefore, measures are proposed to clean the data, excluding the physical removal of the record. One of them is to add to the old record an indication of the unreliability of the data by linking it with later (or network resources), including reliable data. Another measure is to assign the obsolete data to the sign “**low quality**” using the attribute “**consistency**” (see above “**Data Quality**”). Finally, you can assign the status to “**Stale**” or “**Recommended**”, which allows you to immediately separate high-quality data from clearly obsolete data during the search.

Regular data cleaning inevitably requires entering them into a new context, explaining concepts, offering an introduction to the available handbooks, databases, manuals, and so on. In the list of data curation practices such activity was named as

Contextualizing, i.e. “Use metadata to link the data set to related publications, dissertations, and/or projects that provide added context for how the data were generated and why”. A new set of metadata (Table 1) allows linking with external resources through two elements, “**Identification**” and “**References**”. The first uses a link to Public Databases to accurately identify a substance with access to additional data.

For example, by including the reference CSID: 6114 (ID from Database ChemSpider) in the “**Identification**” field, we can select “**Ethylene oxide**” from the group of substances with the same formula C_2H_4O , gaining access to the reference data, Fig. 3.



The screenshot displays the ChemSpider interface for Ethylene oxide (CSID 6114). At the top, the ChemSpider logo and navigation tabs (Simple, Structure, Advanced, History) are visible. The search results show 'Found 1 result' for the search term 'Epoxyethane'. The main entry for Ethylene oxide includes its chemical structure, molecular formula (C_2H_4O), average mass (44.053 Da), and monoisotopic mass (44.026215 Da). Below this, there are tabs for 'More details' (antibacterial drug, fumigant, fungicide, insecticide) and 'Names' (Properties, Searches, Spectra, Vendors, Articles, More). A section for 'Experimental Physico-chemical Properties' lists various properties such as Melting Point, Boiling Point, Ionization Potent, Vapor Pressure, LogP, Flash Point, and Freezing Point, each with a dropdown arrow.

Fig. 3. Typical entry for ethylene oxide from ChemSpider (CSID 6114)

The **References** element (Table 1) provides a link to thematically pertaining resources, but without requiring exact identification of the object. An example is the linking with the Sacada database [**Samara Carbon Allotrope Database**, <http://sacada.sctms.ru/>], which expands the set of information on carbon allotropes presented in the THERMAL database.

Obviously, contextualization as a data curation practice requires the participation of human experts, but not programmers. Therefore, the term “clearing” within the

framework of data curation means not only the rejection of dirty data, but also data analysis and decision making. Thus, contextualization as an element of data curation completely corresponds to the expression “added value to digital research data throughout its lifecycle”, in response to the question “What is digital curation?”

4.1 Preservation

Long-term storage, to a large extent, a purely technological problem related to service life of memory devices (a maximum of 100 years), the solution of which requires significant funding. The required measures include regular backup and failure detection at the bit level. Particular attention should be paid to the physical protection of the data storage, as the frequency of bit rot in data significantly increases due to pollution, thermal and radiation exposure and other external agents. Some of the protection activities are most adequate for research data. Among them are format migration (i.e. consistent change in line with technological changes) and emulation recreating outdated hardware and software on a modern platform.

When transferring the THERMAL database to the Big Data platform, the obsolete ISO-2709 format, adopted in the 60s of the last century [17] as ISO standard for bibliographic description, is discarded. In the new version, documents in ISO format are converted to structured text in JSON format, one of the most convenient for exchanging data and metadata [5]. The advantage of a text document is the possibility of simple reading and editing, accessibility for human perception, convenient form of storage and exchange of arbitrary structured information. The JSON-format (unlike ISO) is convenient for storing factual information in the form of tables and nested structures, as well as numerous links to files of different formats (images, presentation files, Web-pages, etc.), which is especially important when expanding the functions of the THERMAL database. It is also important that the JSON format is a working object for some platforms, in particular for **Apache Spark**, allowing for the exchange, storage and queries for distributed data. There is already an experience of using structured text as a means of thermal properties data interchange [12].

Along with the obsolescence of formats, outdated software that is incompatible with more modern platforms affects long-term storage. This fully applies to the database THERMAL, built on the basis of the documentary Database Management System (DBMS) CDS/ISIS [17] with a fairly limited scope, mainly for the storage of catalogue cards. The transition to the Apache Spark platform (<http://spark.apache.org/docs/>) in combination with ontology-based data management opens up much greater opportunities in storing and integrating data of arbitrary format, converted into JSON-format. In turn, ontology supports a single vocabulary of all concepts, expanded by logical connections and axioms. An ontology encoded as an OWL file becomes a control superstructure capable of semantic integration of heterogeneous data.

The arsenal of tools proposed in the project (JSON, Apache Spark, SPARQL) meets modern standards for storing and processing heterogeneous data and is capable of supporting interaction with many types of storage. Thereby, long-term data storage

will be provided with the possibility of painless migration to subsequent versions of the software.

References

1. Zorich, D.M.: Data management: Managing electronic information: Data curation in museums. *Museum Management and Curatorship* **14** (4), 431 (1995).
2. Beagrie, N.: Digital curation for science, digital libraries, and individuals. *The International Journal of Digital Curation* **1** (1), 3–16 (2006).
3. Abbott, D.: "What is Digital Curation?". DCC Briefing Papers: Introduction to Curation. Edinburgh: Digital Curation Centre. Handle: 1842/3362 (2008). Available online: <http://www.dcc.ac.uk/resources/briefing-papers/introduction-curation>
4. Johnson, L.R., et al.: How important are data curation activities to researchers? Gaps and opportunities for academic libraries. *Journal of Librarianship and Scholarly Communication*, 6(General Issue), eP2198 (2018). Available online: <https://doi.org/10.7710/2162-3309.2198>
5. Kosinov, A.V., Erkimbaev, A.O., Zitserman, V.Yu., Kobzev G.A.: Ontology-based methods of thermophysical data integration. In: XV Russian Conference (with international participation) on Thermophysical Properties of Substances (RCTP-15), 103–104. Book of Abstracts. Moscow, Russia, (2018).
6. Pennock, M.: Curating e-Science Data. DCC Briefing Papers: Introduction to Curation. Edinburgh: Digital Curation Centre. Handle: 1842/3330 (2006). Available online: <http://www.dcc.ac.uk/resources/briefing-papers/introduction-curation>
7. Yerkimbaev, A.O., Zitserman, V.Yu., and Kobzev, G.A.: The role of metadata in the creation and application of information resources on the properties of substances and materials. *Sci. Tech. Information Process* **35** (6), 47–255 (2008).
8. Davenhall, C.: "Scientific Metadata", DCC Digital Curation Manual, J. Davidson, S. Ross, M. Day (eds), (2011). Available online: <http://www.dcc.ac.uk/resources/curation-reference-manual/scientific-metadata>
9. Willis, C., Greenberg, J., and White, H.: Analysis and synthesis of metadata goals for scientific data. *J. American Soc. for Information Science and Technology* **63** (8), 1505–1520 (2012).
10. Erkimbaev, A.O., Zitserman, V.Yu., Kobzev, G.A., and Fokin L.R.: The logical structure of physicochemical data: problems of numerical data standardization and exchange. *Russian Journal of Physical Chemistry A*. **82** (1), 15–25 (2008).
11. Erkimbaev, A.O., Zitserman, V.Yu., Kobzev, G.A., and Trakhtenhers, M.S.: A universal metadata system for the characterization of nanomaterials. *Sci. Tech. Inf. Process* **42** (4), 211–222 (2015).
12. Erkimbaev, A.O., Zitserman, V.Yu., and Kobzev, G.A.: The intensive use of digital data in modern natural science. *Automatic Documentation and Mathematical Linguistics* **51** (5), 201–213 (2017).
13. Water Structure and Science. P7. Supercooled water has two phases and a second critical point. Available online: http://www1.lsbu.ac.uk/water/phase_anomalies.html
14. Kane, C. and Moore, J.: Topological insulators. *Physics World* 32–36 (2011).
15. Eletsii, A.V., Erkimbaev, A.O., Kobzev, G.A., Trachtengerts, M.S., and Zitserman V.Y.: Properties of nanostructures: data acquisition, categorization, and evaluation. *Data Science Journal* 11, 126–139 (2012). Available online: <https://www.jstage.jst.go.jp/browse/dsj>
16. Rahm, E. and Do, H.H.: Data cleaning: problems and current approaches. *IEEE Data Eng. Bull.* **23** (4), 3–13 (2000).
17. CDS/ISIS for Windows: Reference Manual (Version 1.31). Paris: UNESCO (1998).