ISSN 0147-6882, Scientific and Technical Information Processing, 2008, Vol. 35, No. 6, pp. 247–255. © Allerton Press, Inc., 2008.
Original Russian Text © A.O. Yerkimbaev, V.Yu. Zitserman, G.A. Kobzev, 2008, published in Nauchno-Technicheskaya Informatsiya, Seriya 1, 2008, No. 11, pp. 13–19.

The Role of Metadata in the Creation and Application of Information Resources on the Properties of Substances and Materials

A. O. Yerkimbaev, V. Yu. Zitserman, and G. A. Kobzev

Received July 14, 2008

Abstract—This paper examines building and using metadata concerning electronic resources with digital information about the properties of substances and materials. Transformation of the metadata concept as a description of the structure and content of basic information is studied with respect to storage and dissemination of digital data on properties. The manner in which the system of metadata can be used in determination of the logical structure of a subject domain, as determined by the choice of compounds and nomenclature of properties, is shown. As examples, specialized formats are given allowing adequate reproduction of the multidimensional and probabilistic nature of physical properties.

Key words: information technologies, information resources, metadata, digital data on properties.
DOI: 10.3103/S0147688208060038

1. INTRODUCTION. GENERAL INFORMATION ABOUT METADATA

Any information resource (a database, web-portal, or an electronic library), independent of the type, structure and medium of representation, includes, along with the subject information, some volume of attending metadata (MD). In the literal meaning, *MD are data about data*, i.e., formalized descriptions of the internal structure and behavior of information systems. According to the definition given by the Meta Data Coalition (www.mdcinfo.com), "MD is descriptive information about the structure and meaning of data, as well as of appendices and processes which manipulate the data." As such, MD does not carry information, and only describes the attributes of data containing information.

MD can be considered to be an outgrowth of the traditional means of library cataloging (catalog cards). However, the tasks solved by MD are far broader than those to be solved in the practice of libraries or archives. By accompanying the subject information, MD makes it possible to:

(i) identify a resource, including its address in the network;

(ii) determine rules of access and conditions of use;

(iii) describe the structure and schemes of data storage in a database (DB);

(iv) carry out automated analysis of the contents of the resource, along with the construction of search indices;

(v) provide integration of data from various kinds of sources, each with its own structure.

As noted by Antoppolskii [1], MD is an integral concept with respect to previous concepts from the sphere of informatics, i.e., to the formats of data representation and the languages of data description and linguistic support. The main difference from these concepts is their more general character, allowing them to be used in dealing with various tasks on data structuring and processing. A large literature has appeared that analyses the role of MD in the information resources (IR) of different typology and thematic areas [1–6]. Standards have been developed for different subject domains (geography, medicine, archiving, etc.) and types of documents [4]: bibliographic format MARC, Dublin nucleus, format Z39.87 for description of images, or format CDF for multidimensional arrays. Standardization provides homogeneity of representation of meta information and facilitates its dissemination and universal access to the subject information.

The purpose of this work is to analyze the role and form of MD concerning a special kind of scientific information resources relating to information about the properties of substances and materials. The specificity of these resources, which previously was noted by the authors of [7] include: integration of divergent data; the multiplicity of the forms of representation (table, graphic, analytical, etc.); the complicated character of the logical structure determined by a physical model; and the variability of the structure of data depending on the class of a substance and source of data.

The difficulty and diversity of the structures of information resources create serious barriers in the development of means of storing, and, in particular, in the integration of data from different media. As is

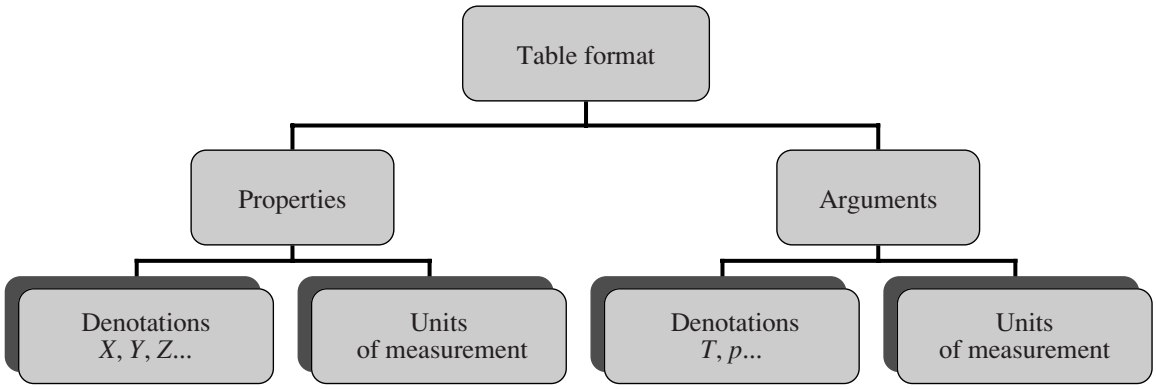

Fig. 1. Common format of a table for the representation of the features of a substance.

shown in [8, 9], the key to this problem is the detailed development of the logical structure of data, which allows adaptation to the subject domain determined by the choice of connections and the nomenclature of their characteristics. The basis of a logical structure is a deeply detailed system of MD. For every fragment of subject information they yield a formalized description of the content, structure, format, dimensionality, and all other characteristics. Traditionally MD performs the function of a guide explaining the kind of information, its source, degree of relevance, or measure of confidence, while in respect to information on properties it is possible with its help to form the basis of a conceptual model of the subject domain represented as a concrete scheme of a DB.

We will show that in a DB on the properties of substances, the scheme plays a role not less important than the information content by providing not only the correctness and unambiguity of its interpretation, but also the general idea of the subject domain, including dictionaries of concepts, nomenclature of characteristics, structure and formats of data, etc. Such a scheme is needed by systems administrators, developers and end users, although the main "consumers" of MD are program-realized in information resources. A mandatory condition, therefore, for the functioning of the MD system is formalization provided by a set of dictionaries and rigid description of formats.

In general, the delineation of data and MD in any subject domain is rather conventional, since the description of an element of data also carries in itself an information content. Thus, bibliographic data, depending on its purposes, can refer both to main and to meta information. In an earlier project on the standardization of thermodynamic data [10], the MD included information on the parameters of the state, phase, method for measurement, and the uncertainty of the result. With a high volume of MD, the concept of higher level, meta-meta information, i.e., a description of MD group, is justified.

In this paper, a relatively universal system of data on properties is proposed, in which it is possible to deepen the structure while tuning it to the subject domain. This system, at the same time, treats the choice between content and meta information unambiguously.

2. METADATA IN THE SYSTEM OF REFERENCE DATA ON PROPERTIES OF SUBSTANCES AND MATERIALS

The role of MD as the basis for formalization of a subject domain is vividly illustrated by accumulated work with digital data about the properties of substances and materials that has been accumulated in physical–chemical and engineering sciences [7–9]. Long before the appearance of present-day computer DBs, the reference literature had actually established standards of representation of reference data, mostly as model tables.

The format of tables, unlike their content, was actually just a harbinger of the present concepts about MD. It included a list of characteristics (attributes), designation of each of them and measurement units, with the same information for arguments (measurements) determining the values of each property. The scheme in Fig. 1 presents no more than a minimal set. In reality, the standard format for presentation of data about properties often includes optional elements, i.e., references to the sources of data, information about errors, methods of measurement, temperature scale, etc.

Two characteristic points that later determined the MD structure for properties can be found in the formats of reference tables. The first is identification of the subject, i.e., its substance, material, and composition. In the simplest case the stoichiometric formula or trivial name serves as the identifier. But for the case of organic substances the need appears for a set of identifiers, for which the structural formula, list of synonyms, number by Chemical Abstracts classification and a number of other identifiers are used. With reference to the solid phase, there is a need to detail the polymorphic modifi-

cations, porosity, and other features in the preparation of the sample, etc.

The second point is the *permanent* departure of the authors of any of the reference books from a rigid format predetermined by the choice of columns and arguments. As was noted in [8, 9], all the work with digital data shows the impossibility of unifying the format of a table for a broad totality of substances without being in contradiction with the objective need to represent their specificity, as well as standards of description by different authors. The correction of the format can be associated with the appearance of new characteristics, a change of the physical model, manifestation of new parameters of state, etc. In a text document (a handbook or paper) the "breaking" of the table structure is not a serious problem, being explained by notes, comments, etc.

At the same time, in this case the computer form of storage has required abandoning the relation scheme of a DB and the use of a new concept, so-called *semistructured data* (SSD) with appropriate instruments and technologies [11, 12], see Section 3. Here it is important to note that the basis of the technology is a varied system of MD, which forms the model of the subject domain, taking the typology of different subjects into account. A de facto prototype of such a system has been established already in the reference literature. For example, the multi-volume journal Thermal Constants of Substances [13] and its electronic analogue on the server of the Department of Chemistry of the University of Moscow (www.chem.msu.su) are built on the basis of a type table, including a set of thermodynamic constants for each of the elements or compounds. Depending on the aggregate state, crystalline modification, presence of a solvent, kind of molecule (radicals or ions), deviations from stoichiometry, the format of the table constantly changes (see Fig. 2 where the formats of tables for O and O_2 are compared)

Now we will show how experience obtained in the editing of reference literature has naturally been used in information technology, where MD yields the required features for scientific data, i.e., the flexibility of its logical structure and the multivariate and probabilistic character of the data.

3. THE UNIFIED SCHEME OF SEMISTRUCTURED DATA (SSD) ABOUT PROPERTIES OF SUBSTANCES AND MATERIALS

3.1. Specificity of SSD

The term "semistructured data" relates to data with a varied structure, which changes (can change) from entry to entry [11, 12]. With all its diversity, the SSD have one (or several) of the following features:

• irregularity in logical structure: differences in the number of attributes, parameters of the state, form of representation, etc.;

(i) an a priori table-oriented scheme accepted for structured data is replaced by an *a posteriori* one formed after the addition of the last set of data;

(ii) permanent evolution of the scheme simultaneously with the data renewal;

(iii) a large volume of MD fixing the logical scheme commensurable with the volume of digital data.

This type of data has turned out to be fully adequate for those standards which appear in the creation of DBs using the properties of substances, with account for variability of the nomenclature of characteristics and form of representation [9]. Two points are important, i.e., (1) at all stages of systematization and dissemination of data, the scheme of their presentation plays a role not lower than the data themselves; (2) the scheme is always an *a posteriori* one, evolving with the accumulation of data. The latter signifies the *principal impossibility of predicting the nomenclature of concepts used for identification of a substance and a list of possible attributes for representation of its properties in advance*.

Efficient possibilities for the creation of a varied structure are laid in the DB instruments in PostgreSQL [14, 15], which are seen as an alternative to commercial products in the storage of large arrays of natural scientific data, for example, astronomic catalogs. This is possible to create, in particular, by developing one's own composite types of data, COMPOSITE, consisting of elements which may have different tolerable types in a DB.

Working with cells of a composite type for some entry of a DB table it is possible using a database management system (DBMS) to change or expand the content of the cell, including additional attributes (which generates new tables without their formal announcement). This opens the possibility in principle of connecting the table cell with the *hierarchical structures* the complicated and varied character of which are required for adequate communication of the essence of the real world. The task for MD attending all nodes of the hierarchical scheme formed on its basis is the full description of the type of data inserted during use.

Along with its hierarchical structure, the model of the data includes a set of *dictionaries–classifiers* for the nomenclature of concepts used for the description of objects and their properties. By this term, a numbered list of names to denote classes of substances, properties, units of measurement, methods of representation, etc., are understood. The classifier formalizes the used system of concepts, excluding the possibility of different interpretations, and provides the basis for a machine search. Each dictionary is able to receive new concepts, the need of which arises with new entries. The names of dictionaries, like their contents, also relate to the MD category, by which it is possible to structure and search

[Database of thermal qualities of substance (TQS)] [Parameters and definitions] [Search and Browsing of Information]

Substance: O_2 — Units of measurement: calories / Pass to Joules

State	Thermal parameters	
r	$DH°f_0$, kkal/mol	0
r	$DH°f_{298.15}$, kkal/mol	0
r	$DG°f_{298.15}$, kkal/mol	0
r	D_0, kkal/mol	117.97 ± 0.06
r	$H°_{298.15} - H°_0$, kkal/mol	2075 ± 1
r	$S°_{298.15}$, kkal/(mol K)	49.005 ± 0.008
r	$C°p_{298.15}$, kkal/(mol K)	7.015 ± 0.002
p-p; nH2O	$DH°f_{298.15}$, kkal/mol	−2.9 ± 0.05
p-p; 0.3 Ar	$DH°f_{84}$, kkal/mol	0.0137
p-p; 0.4 Ar	$DH°f_{84}$, kkal/mol	0.0169
p-p; 0.5 Ar	$DH°f_{84}$, kkal/mol	0.0198
p-p; 0.6 Ar	$DH°f_{84}$, kkal/mol	0.0220
p-p; 0.7 Ar	$DH°f_{84}$, kkal/mol	0.0242
p-p; 0.8 Ar	$DH°f_{84}$, kkal/mol	0.0259
p-p; 0.9 Ar	$DH°f_{84}$, kkal/mol	0.0274
p-p; 1.0 Ar	$DH°f_{84}$, kkal/mol	0.0286
p-p; 1.2 Ar	$DH°f_{84}$, kkal/mol	0.0307
p-p; 1.4 Ar	$DH°f_{84}$, kkal/mol	0.0324
p-p; 1.6 Ar	$DH°f_{84}$, kkal/mol	0.0338
p-p; 1.8 Ar	$DH°f_{84}$, kkal/mol	0.0351
p-p; 2.0 Ar	$DH°f_{84}$, kkal/mol	0.0364
p-p; 2.2 Ar	$DH°f_{84}$, kkal/mol	0.0375
p-p; 2.4 Ar	$DH°f_{84}$, kkal/mol	0.0387
p-p; 2.6 Ar	$DH°f_{84}$, kkal/mol	0.0392
p-p; 2.8 Ar	$DH°f_{84}$, kkal/mol	0.0407

Ionization potential

ev	cm^{-1}	kkal
12.075 ± 0.01		278.49 ± 0.23

Values of experimental values in this section are associated with hyper-references with a bibliography of publications, by which these values have been chosen

Crystallographic data

State	Syngony	Symmetry class	Spatial group	Structural type
kI(gamma)	cu	m3m	$Fm3m(O_h^3)$	Cu
kII(beta)	hex	−3m	$R\text{-}3m(D_{3d}^3)$	own
kIII(alpha)	rhomb			

Parameters of phase change

type of change *p*, atm	T.Q.	*t*, °C	*DH*, kkal/mol	*DS*, kkal/(mol K)
kIII → kII	23.85 ± 0.04	−249.30	0.02242 ± 0.0001	0.94
kII → kI	43.77 ± 0.05	−229.38	0.1776 ± 0.0005	4.058

[Database of thermal qualities of substance (TQS)] [Parameters and definitions] [Search and Browsing of Information]

Substance: O — Units of measurement: calories / Pass to Joules

State	Thermal parameters	
r	$DH°f_0$, kkal/mol	58.985 ± 0.03
r	$DH°f_{298.15}$, kkal/mol	59.556
r	$DG°f_{298.15}$, kkal/mol	55.393
r	$H°f_{298.15} - H°_0$, kkal/mol	1608 ± 1
r	$S°_{298.15}$, kkal/(mol K)	38.467 ± 0.005
r	$C°p_{298.15}$, kkal/(mol K)	5.237 ± 0.002

Values of experimental values in this section are associated with hyper-references with a bibliography of publications, by which these values have been chosen

Ionization potential

ev	cm^{-1}	kkal
13.6177 ± 0.0018	109836.7 ± 1	314.037 ± 0.036

Fig. 2. Comparison of the structure of table for representation of thermodynamic data for oxygen (O and O_2).

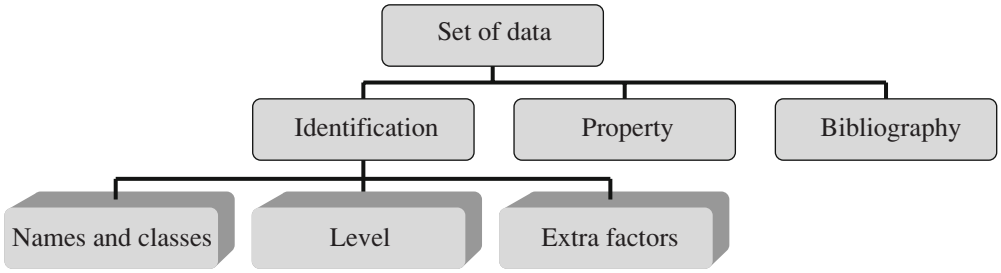

Fig. 3. General structure of the set of data on substance properties.

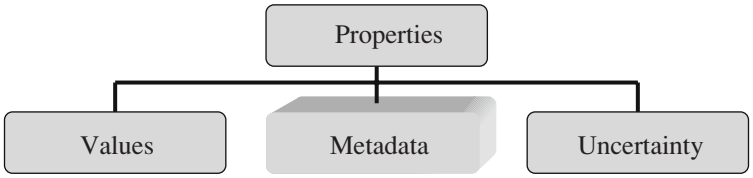

Fig. 4. Data block "Properties".

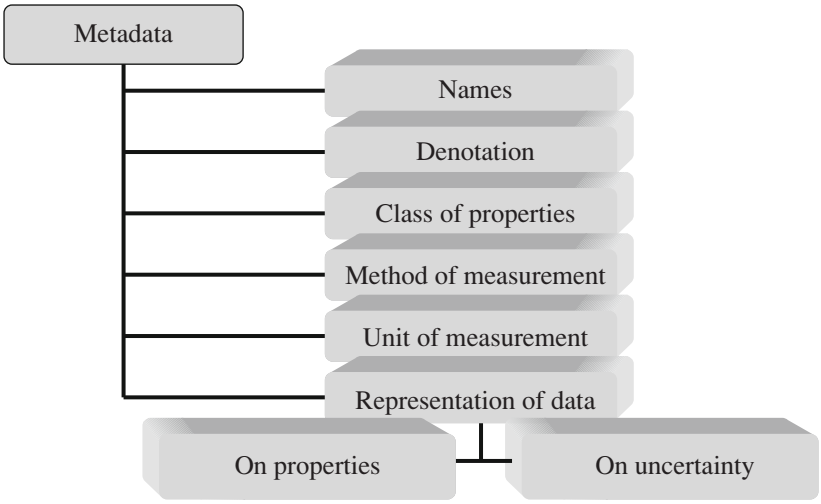

Fig. 5. Metadata representing information on property.

the information content, including tables of numbers and texts, graphics, mathematical constructions, etc. The content of the data is opened by the reference to the dictionary, the name of the dictionary and the number of the item. Thus, the MD system forms a model of the subject domain reflecting variations of the structure appearing with new data.

3.2. Scheme of Data for Representation of the Properties of Substances

The above scheme implies the possibility of arbitrary versions of the structure for each entry. At the same time, experience in systematization of reference data on properties [9] has allowed us to propose a model with a relatively "unstable body" (Figs. 3–5), the

versions of which concern only particular elements. The set of data includes three blocks of the upper level, i.e., "identification," "properties" and "bibliography" (see Fig. 3). The block of the "identification" integrates signs by which it is possible to unambiguously distinguish the subject distributed among the blocks "names and classes," "level," and "extra factors." The first has as an identifier one or several names used with respect to inorganic and/or organic compounds (chemical formula, systematic or trivial name *CAS Number*, etc.), and directions to relate the substance to one or several classes, say, metals, oxides, minerals, etc. The lists of classes, in accordance with the effective nomenclature adapted to the subject domain, are included in the previously prepared dictionaries.

The second block, the "level," separates objects relating to the macro, micro and meso levels. At the macro level, phases and interphase areas or points are identified. At the micro level, molecular forms, i.e., molecules, radicals, or ions, are distinguished. The *meso* level contains objects that are intermediate in size and properties between a macroscopic object and an individual molecule. Such objects can be clusters, fullerenes, nanotubes, two-dimensional and one-dimensional structures (films, nanowires, etc.). The identifiers of the block "level" make it possible to unambiguously divide data relating to a substance as a whole from data relating to certain molecular forms, phases, interphase areas, etc.

The last block, conventionally called "extra factors," has those identifiers which can supplement chemical identifiers with the specifications of data about the structure and state of the sample. These factors can include data about the porosity of the sample, degree of dispersion of a powder, the geometry and thermal background of the sample, etc.

3.3. Block of Data PROPERTIES

The second block of the upper level (Fig. 4) includes a set of attributes inherent in the subject identified by identifiers from a certain set. For any property, it is also possible to propose a relatively unified scheme (see Figs. 4, 5), which can be adapted to the specificity of the subject and to the form of data representation. Information on the property is distributed among three blocks of the lower level, i.e., two of them yield values of the property itself and information on uncertainty, the third one presents a set of MD explaining the content, structure, and the form and formats of representation, providing unambiguous interpretation of data by both the user and the computer.

In explaining the content, MD can specify the names, denotations, units of measurement, method of measurement (or estimation) and, most significantly, the *form of representation* of the physical value and its uncertainty. *Representation of data* (the lower block, see Fig. 5) is a fairly capacious and polysemic concept covering details in the definition of a property, the types of data used, a set of measurements for multivariate data, along with identification of their types and meanings. Detailing is necessary, since in many disciplines the definition of properties is associated with the context, i.e., the method of measuring, modeling, the area of utilization, etc. A characteristic example is the hardness of material determined by its method of measurement (Brinnel figure, Rockwell hardness, Vickers hardness, etc.). Multiplicity in the definition of a particular property is found especially often in physical chemistry, when, setting the value of a property, one has to fix the standard state, start of measurement, temperature scale, or scale of relation in recording reduced values (for example, parameters of the potential or critical constants).

Possible versions in the representation of data are fixed in appropriate dictionaries that are included in the MD system. Information about uncertainty also "frames" the metadata, thus securing correct and unambiguous interpretation (more details are in Section 5). Apart from the content, MD fix the formal aspects of the used data, i.e., its format (numbers, text, graphics, mathematical statements) and the type of data (symbolic, digital, arrays, composite types, etc.). The latter integrate elementary types, allowing the user to manipulate complicated objects.

4. TYPES OF DATA FOR REPRESENTATION OF MULTIVARIATE PROPERTIES

The limited abilities of relational DBs for the solution of natural scientific tasks, which are found in the literature [7, 16] are connected not only with the variation of structures but with the need to support large multivariate arrays. This has been most vividly shown for the Earth or Space sciences, where an element of data is indexed, in addition to its physical characteristics, by spatial coordinates and the time of measurement. A number of open formats for work with multivariate data of scientific direction [4] have been developed, for example, Common Data Format (CDF, NASA development) or Hierarchical Data Format (HDF5, National Center of Supercomputer Applications of the United States).

The problem of dimensionality is also typical of tasks in data storage using properties. At a minimum, two dimensions are required in order to fix the state of a single-component substance. For solutions, this number increases up to $N + 1$, wher N is the number of components. In many cases there is also a need for additional parameters. Thus, for radiation strength, in addition to parameters of state, the wavelength and angle to the normal are set.

For effective support of huge DBs, highly effective multivariate index structures are required. In the general case, the organization of a multivariate array $A(I, J, \ldots, K)$ implies the presence of a whole series of

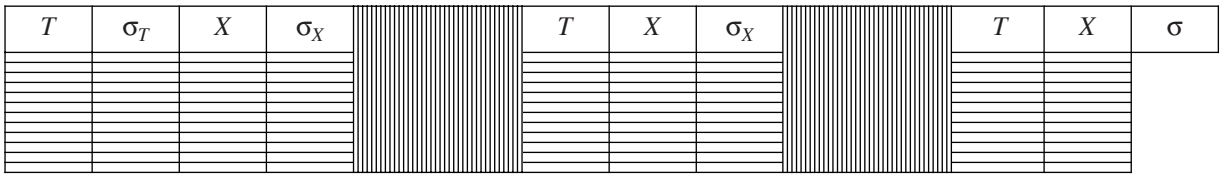

Fig. 6. Schematic representation of data uncertainty.

agreements. For example, in creation of the ASAR system (Array Streaming and Processing) [16] for any of the indices $I, J, \ldots, K$, one of the primitive types of data (an integer or floating number, line, etc.) was set, a regular or irregular step was introduced, and a regular sequence of values was supported. The very element of massive A for any set of values $I, J, \ldots, K$ was considered as a list of values of attributes, each having a name containing meaning of a primitive type, and it could also include an indefinite value (NULL). In essence, the element of the array was no more than a *relational table*. This makes it possible to store and manipulate an entire table of related properties (for example, thermal capacity, enthalpy, entropy, etc.) within one structure, which meets the established practice of work with physical–chemical data. The preliminary work on representation of the table structure as a multivariate array is done to specify the values of metadata, determining the type of indices, steps, and attributes of elements $A(I, J, \ldots, K)$.

5. INFORMATION ON UNCERTAINTY

Another invariable characteristic of data on properties is the *uncertainty* associated with the error of an experiment, method of processing or appraisal, authenticity of a model, etc. The characteristics used in appraisal of uncertainty are rather diverse. The simplest method is to "double" data on the value of property X at every point of the value of the mean square error σ, although the probability that the value of the property will be found in the range $[X - \sigma, X + \sigma]$ remains unknown. A more informative method is the confidence interval σ_L, where the probability of being in the interval, i.e., the level of significance $L(\%)$, is set.

In the ASAP system, developed as a prototype of a DBMS for the management of scientific data, three methods for the representation of uncertainty have been proposed, i.e., described, as the set of pairs (X_1, p_1), (X_2, p_2), …, (X_N, p_N) where p_i is the probability that the value of the property is the number X_i: the standard one, as the pair of meanings (X, σ); the mean and error and; finally, as the fixation of the lower and upper borders and probabilities of overrunning the borders of the interval $p_1 = \text{Prob}(X > UB)$, $p_2 = \text{Prob}(X < LB)$ where UB, LB are the upper and lower borders, respectively. Additional versions are associated with the concept that the error can be ascribed both to the property and to the independent variables (state parameters), for both cases, the above parameters are applicable.

In the ThermoML project [10] proposed for the standardization of processes of storage and exchange of data about thermodynamic properties, three types of appraisals have been identified, i.e., standard σ, expanded σ_L and combined. The first two can be associated with both dependent and independent variables and the latter type characterizes the uncertainty of the property, combining the errors of independent and dependent variables. The version of appraisal of uncertainty which is reasonable to store in a DB is associated with the completeness of the "primary" data and with requirements for the quality of the data from applications. In many cases one appraisal for the whole set is enough (in absolute or relative values). The statistical appraisals in project [10] are ascribed to each point from the set of data, and for the set as a whole a group of rough appraisals of authenticity, i.e., repeatability, reproducibility, deviation from the approximating curve, has been suggested.

The choice from the set of choices for the appraisal of uncertainty is realized by the DB system combining the dictionaries of concepts, types of data and dimensionality of arrays. The set of tables in Fig. 6 illustrates the versions of appraisal: for the error for the parameter of state σ_T and property σ_X in each point (on the left), the error only for the property in each point (in the center), and one error per entire set of data (on the right). The entire set of different interpretations of the concept *uncertainty* should be presented in the dictionary. For example, uncertainty σ_X can correspond to the mean square error, the tolerance interval at a certain level of significance, relative error, etc.

6. CONCLUSIONS

It is shown here how significant the role of metadata is in the creation of means of storage of data on the properties of substances and materials. In addition to the usual functions developing old forms of library cataloging, for the resources of this type, MD provides for the formation of varied logical structures of data. In its volume and significance for the user, the scheme of the description of data formed by MD is quite commensurate with the main (content) information. It ensures realization of all basic functions, i.e., search, conversion of formats, logical and mathematical processing

and, what is most important, integration of heterogeneous sources with different structures.

The formula "information = data + MD" suggested by Bill Inmon, the father of the concept of the present-day storage of data [3], is realized in an ideal way. If the statistical theory created for problems of communication and coding identified the concepts information and data [17, 18], in modern technologies the accent is transferred from the bit chain to the interpretation of messages and data. The basic difference of both interpretations in informatics has been emphasized in the lectures of Gilyarevskii [19].

The change from data to information becomes a tendency in information technologies, being expressed in the appearance of new architectures, concepts, languages, etc. The best known example is the language of augmented tracking, XML, structuring documents by MD [3, 11, 12]. *Content* and *service* directed approaches to the integration of data, including program agents for the identification of MD, have appeared, which are used for access to data. In [3] the program solutions created by the Infostoria company are described, which have radically improved the possibilities of *enterprise content integration* (ECI) that work only with information rather than data. These extract MD from "rough" data, placing it in catalogs, automatically or with human participation. Later, the MD serves to integrate the *content*. The created information resources contain, in addition to a rigidly programmed body, an easily tuned part managed by MD, which ensures its potential for development and integration with other resources.

As was noted in [7], new information technologies are created preeminently in the sphere of business, slowly penetrating the academic community, which is also true for technologies directed toward the use of MD. On the other hand, it has been shown that MD play as important a role in storing data on properties as the content information, ensuring the unambiguity of interpretation, description of the subject domain, supporting varied structures, etc. Earlier methods, the basic ones of which are relational DBs, lead unavoidably to serious losses (unjustified narrowing the circle of objects, inadequate transmission of individual features, or problems with integration of data from different media). All these losses are the manifestation of the inability of traditional methods in the communication of *information* rather than data.

The technologies designed to work using the formula "information = data + MD" slowly begin to penetrate the instruments of scientists as well. As an example, we have cited the ThermoML standard [10] for presentation of and exchange with thermodynamic data, specialized formats for multidimensional and probabilistic data (Sections 4 and 5). In [8, 9] a model with a relatively "stable body" was proposed, allowing variation of the structure depending on the specifics of the subject domain, i.e., the circle of substances, nomenclature and typology of characteristics. The model uses the DB instruments in PostgreSQL [14, 15], providing the ability to determine its own types of data with a fairly complicated structure.

The introduction of methods and technologies directed to the communication of information by MD is becoming especially powerful in the creation of DBs using the properties of nanodimensional objects. Occupying an intermediate position between a single molecule and a macroscopic substance, these objects perceivably differ in their properties from their analogues on the micro and macro scales. The difficulty of identification and dependence of properties on many factors (sizes, geometry, conditions of determination) practically exclude the possibility of data storage without a deeply ramified system of MD, an example of which is suggested in [9].

REFERENCES

1. Antoppolskii, A.B., *Sistemy metadannykh v elektronnykh bibliotekakh* (Systems of Metadata in Electronic Libraries), http://www.gpntb.ru/win/inter-pevents/crimea2001/tom/sec4/Doc5.NTML
2. Grishchenko, A., Makarenko, I., Systems Based on Metadescriptions, *Otkrytye sistemy,* 2001, no. 10.
3. Chernyak, L. Metadata and Integration of Content, *Otkrytye sistemy*, 2006, no. 2.
4. Khokhlov, Yu.E., Arnautov, S.A., A Review of Metadata Forms. Portal *Rossiiskiye elektronnye biblioteki*. http:www.elbib.ru/index.phtml?page=elbib/rus/methodology/md rev
5. Diann Rush-Feya, International Initiatives in the Field of Metadata—Recent Achievements, *Elektronnye biblioteki*, 2000, vol. 3, no. 4.
6. Bezdushnyi, A.A. et al., Proposals on Sets of Metadata for Scientific Information Resources ENIP RAN, *Elektronnye biblioteki,* 2004, vol. 7, no. 5.
7. Zitserman, V.Yu., Kobzev, G.A., and Fokin, L.R., Possibilities and Prospects of Information Technologies in Preparation and Dissemination of Reference Data: Properties of Substances and Materials, *NTU, Ser. 1*, 2004, no. 2, pp. 7–14.
8. Yerkimbayev, A.O., Zitserman, Construction of Information Systems by Properties of Substances Using Technologies of Feebly Structured Data, in *Proceedings of the XI. Russian Conference on thermal Physical Properties of Substances*, St. Petersburg, 2005.
9. Yerkimbayev, A.O., Zitserman, V.Yu., Kobzev, G.A., and Fokin, L.R., Logical Structure of Physical-Chemical Data. Issues of Standardization and Exchange of Digital Data, *Zhurn. Fiz.khim.*, 2008, vol. 82, no. 1, pp. 20–31.
10. Frenkel, M., Global Communications and Expert Systems in Thermodynamics. Connecting Property Measurement and Chemical Process Design, *Pure Applied Chem.*, 2005, vol. 77, no. 8, pp. 1349–1367.
11. Kogalovski, M.R., *Entsiklopedia teknologii baz dannykh* (Encyclopedia of Data Base Technologies), Moscow: Finansy i statistika, 2002.
12. Grinev, M., Systems of Management of Semistructured Data, *Otkrytye sistemy*, 1999, nos. 05–06.

13. *Termicheskiye konstanty veshchestv. Spravochnoye izdaniye v 10 vypuskakh* (Thermal Constants of Substances. Reference edition in 10 issues), Glushko, V.P., Ed., Moscow: VINITI, 1965–1982.

14. Bartunov, O., *What is PostgreSQL*? http://www.sai.msu.su/megera/postgres/talks/what is postgresql.html

15. Bartunov, O.S., Sigayev, F.G., Specialized Types of Data for Digital Libraries, *Trudy 9-i Vserossiiskoi nauchnoi konferentsii "Elektronnye biblioteki: perspektivnye metody i tekhnologii, elektronnye kollektsii"* (Proceedings of the Russian Conference: Electronic Libraries: Prospective Method and Technologies, electronic Collections), RCDL'2007, Pereyaslavl-Zalesski, 2007.

16. Stonebreaker, M. et al., *Is One Size Suitable for All*? Part 2: Results of Tests [in Russian]. IT-Portal SITFORUM.ru.

17. Shennon, C., Works on Theory of Information and Cybernetics [in Russian], Moscow: Izd. inostr. lit., 1963.

18. Kolmogorov, A.N., Three Approaches to the Definition of the Notion "Quantity of Information," *Problemy peredachi informatsii*, 1965, no. 1, pp. 1–7.

19. Gilarevskii, R.S., *Osnovy informatiki: kurs lektsii* (Basic Informatics Course of Lectures), Moscow: Ekzamen, 2003.