Ontological Approach to the Description of a Common Digital Space of Scientific Knowledge

Olga Ataeva[1] [0000-0003-0367-5575], Nikolay Kalenov[2] [0000-0001-5269-0988],
Vladimir Serebryakov[3] [0000-0003-1423-621X]

[1,2,3]Dorodnicyn Computing Center FRC CSC of RAS, Vavilov str., 40, 11933, Moscow, Russia
[2]Supercomputer Center of the RAS, 32a, Leninsky Prospect, Moscow, Russia, 119334
[1]oli@ultimeta.ru, [2]nekalenov@mail.ru, [3]serebr@ultimeta.ru

Abstract. Despite the development of technical means, the processes associated with the search for complete and accurate scientific information in a huge number of data sources are becoming more complicated. To reach a new level in the use of information processing technologies, first of all, a transition to a semantically meaningful representation is necessary for scientific knowledge extracted from information in a digital environment. In modern conditions, characterized by multidisciplinary research, the desired effect can be achieved by developing universal approaches to the storage and presentation of scientific knowledge. These approaches are reflected in the concept of the Common Digital Space of Scientific Knowledge. The paper presents an overview of the basic concepts in this area, which are used both to represent the elements of space and to provide access to them not only for humans, but also for software agents. Semantic libraries are considered as tools for constructing the knowledge space.

Keywords: knowledge space, digital knowledge space, ontologies, metadata, scientific knowledge, metadata levels, ontology design, semantic libraries.

1 Introduction

The development of digitalization of many aspects of society's life has put it before the need to accumulate and process a large amount of information. There is an intensive development of information resources of a new type, new ones are emerging that widely use the digital representation of scientific resources. A large number of information sources have appeared and provide data in different forms and formats and representations. Despite the development of technical means, the processes associated with the search for complete and accurate scientific information become more complicated, and the time required for information processing increases dramatically. With the appearance of the Semantic Web paradigm, attempts to formalize knowledge in various fields of science based on the developed ontologies are being made to solve these problems. This enables the semantic processing of information, the extraction of new knowledge.

To reach a new level in using the capabilities of today's rapidly developing information processing technologies, first of all, it is necessary to move to a semantically meaningful representation of scientific knowledge extracted from information in the digital environment. Although each field of science has its own specifics, in modern conditions, characterized by multidisciplinary research, interpenetration of scientific directions, the necessary effect can be achieved only by developing universal approaches to the storage and presentation of scientific knowledge. These approaches are reflected in the concept of creating a Common Digital Space for Scientific Knowledge (CDSSK) [1, 2]

The development of technology allows us to take a broader look at the definition of the CDSSK and summarize the accumulated experience in the implementation of various solutions in this area. The following part is an overview of the basic concepts in this area, which are used both to represent the elements of space and to provide access to them not only for humans, but also for software agents, which opens up wide possibilities for their processing and use in various areas of consumption by interested participants in scientific activity.

The consistency of scientific information [4, 5] implies reliance on the study of various dependences. The specificity of such information is a clear structure of the organization of scientific data in hierarchical structures, permeated with horizontal links. As a result, an unambiguous interpretation of scientific knowledge by various researchers is provided. The main problem of presenting scientific information is the complexity of the concepts used and the relationships between them, and, most importantly, they are subject to more frequent changes in data structures, which inevitably leads to the need to make improvements to its description.

The definition of scientific knowledge is closely related to the concept of scientific information, which is defined according to State standard GOST 7.0-99 [6] as logically organized information obtained in the process of scientific cognition and reflecting the phenomena and laws of nature, society and thinking. As can be seen from the definition, these two concepts, scientific information and scientific knowledge are often used interchangeably. Further in the text, the term scientific knowledge is used, which, in our opinion, most accurately reflects the meaning.

2 Components of CDSSK

The space of scientific knowledge is understood as a system of knowledge tested by the scientific community from various fields of science. At the same time, *the digital space of scientific knowledge* (DSSK) is a digital environment which information resources and objects that have been proven by the scientific community. The scientific knowledge from a certain field of science are integrated in that space. The DSSK subspace is part of the space bounded by the framework of a certain subject area. In fact, the CDSSK consists of a set of subspaces related to various areas of science, built according to common principles.

Despite the fact that there are some examples of formalization of knowledge in different subject areas [7, 8–15], there is no generalized approach to defining the digital

space of scientific knowledge. An analysis of examples of the formalization of the knowledge space in various fields indicates that the main components of the DSSK in general and each of its subspaces in particular are *ontology* and its *content*.

A set of digital copies of real-world objects and a description of their metadata profiles are considered as *content*, while an *ontology* includes a universal description of the CDSSK data structure. So the ontology of the CDSSK contains classes of objects reflected in each subspace, the types of relationships between these classes and their objects both within one subspace and between subspaces, as well as the rules for reflecting objects in the CDSSK.

3 Approaches to the ontology construction of the CDSSK

The construction of the ontology of the CDSSK subspace can be represented in terms of two orthogonal approaches:

1. terms are introduced and characterized the scientific subject area under consideration, connected by various links, both hierarchical and horizontal;
2. a set of definitions is introduced at a more abstract level, describes the set of objects of a scientific subject area, in fact, setting the structure of their description and relations between them.

In various studies [3, 7–10, 16, 17], in both cases, one speaks either about the construction of a domain thesaurus, or about the construction of a domain ontology. But, in fact, these are two completely different approaches to describing the subject area, which are not mutually exclusive at the same time, but should complement each other. This approach allows, on the one hand, to separately focus only on the types of information resources that are elements of the knowledge space, and to describe the basic concepts characteristic of this subject area. On the other hand, speaking about a thesaurus, one must bear in mind a set of concepts and terms that provide terminological support for the concepts of the domain ontology. Based on the foregoing, a knowledge space thesaurus is a complete systematized set of terms of any area of knowledge, largely and more related to the vocabulary used in a specific area, while an ontology describes the resources of the subject area and their interrelationships. For each subject area, the set of resources may differ both in format and in the set of resources themselves.

The ontology of the scientific space of knowledge is a complex multi-level system of concepts describing resources and objects of the subject area, concepts, terms and connections between them, characterized by an open hierarchical and dynamic structuring and serving both for storing existing knowledge and their structuring, and for extracting new ones.

4 Designing the ontology of the CDSSK

Based on the classical definition of an ontology according to Gruber [18], the content of the CDSSK, identified as a separate component, is an ontology of a certain concrete

subspace (specification of conceptualization), which is based on a more abstract system of concepts of the CDSSK ontology.
Designing the system-wide part of the CDSSK ontology involves the selection of a set of universal and system-wide classes, the definition of their attributes. Among the universal classes, there are system-wide classes, whose object instances can simultaneously refer to different subspaces. These include persons (one scientist can work in different fields of science), organizations (one organization can be engaged in polythematic research), geographic concepts, polythematic journals, collections, databases, etc.

Along with such classes, in each subspace there may be classes specific to this particular subspace. Designing the content of a thematic subspace includes the creation of its subject ontology, the definition of object classes specific for a given scientific direction and their attributes. The subject ontology of a subspace includes a set of indexes of classification systems, key terms with their thesaurus connections related to a given scientific area, a set of metadata specific to the subspace.

5 CDSSK ontology metadata levels

In fact, solving the problem of ontology design, we come to the need to use metadata of different levels:

1. metadata as universal concepts of the CDSSK;
2. metadata as part of the description of the objects of the application area or a subset of the CDSSK;
3. Application area metadata as such.

In a such ontology, at the top level, used concepts are essentially related to high-level ontologies and are not related to the specifics of any specific subject area. At the second level, concepts are used for describing the subject area, while being instances of classes defined at the first level, but at the same time used as class definitions to describe data of the third level already in a specific subject area. In other words, at the first level are given definitions of the basic concepts that are used in the formation of the CDSSK, including:

- thematic subspace;
- content of the CDSSK (a set of information objects);
- information object (a digital copy of a real world object or a specially created digital object that reflects certain properties of a real object);
- information object identifier – a data element that makes it possible to unambiguously identify an object in the CDSSK;
- attributes of a digital object (a set of metadata (object properties) that characterize the object from the point of view of the tasks of the CDSSK);
- data source (real world object containing information mapped in the attributes of a digital object);

- subject ontology of a subspace – a set of indexes of classification systems, key terms with their thesaurus connections related to a given scientific direction;
- the subject ontology of the CDSSK – a set of subject ontologies of individual subspaces;
- thesaurus links – links between two elements of subject ontology A and B, which take one of 4 values: "A is equivalent to B", "A is included in B", "A contains B", "A intersects with B";
- local class of objects – objects belonging to one thematic subspace;
- universal class of objects – objects associated with several thematic subspaces.

At the second level, we describe the concepts of a specific domain as instances of the first level classes, i.e. for example, a specific thesaurus, specific types of information resources, types of data sources, etc.

Second-level concepts are used as class definitions at the third level when filling the ontology with data that are instances of second-level classes.

At the same time, if the new introduced concepts are at the second level instances of the designated resources of the first level, then when filling the ontology of the CDSSK we use them as classes for describing data. Considering instances as classes is called metamodeling. And although even the direct semantics of the OWL2 ontology language used to describe ontologies does not allow such metamodeling, this limitation in the language is circumvented using a syntactic trick known as punning. This means that when an instance identifier is found in a class axiom, it is treated as a class, and when the same identifier occurs in a separate statement, it is treated as an instance.

So, when building the ontology of the CDSSK subspace or a specific subject area, in fact, a three-level ontology is constructed, in which the first-level instances are high-level concepts and are used for class definition on second level used in their turn for filling the ontology with data in third level.

6 Semantic library as a tool for constructing the CDSSK

The formation of a model with the listed properties meets the requirements of constructing an ontology of a semantic scientific library, which is close, in fact, to high-level ontologies [20] for the subject areas of science. In fact, the concepts are divided into three categories: the first includes definitions of the concepts of the content of the semantic library and the second category refers to the definition of the concepts necessary to support the terms in the subject area thesaurus and the third includes the definitions necessary to define the processes of integration of the content of these resources [21–24]. Based on these definitions, basic processes are described, such as, for example, integrating data from different sources, categorizing / classifying, mapping different data models of sources to a given subject area, building equivalence classes, etc. This approach is in good agreement with the above-described three-level ontology and allows us to talk about semantic libraries as a tool for constructing subspaces of the CDSSK.

The semantic library must support a data model for describing scientific resources and allows you to not be limited in development to a strictly delineated set of resources.

The application of the described model allows one to reduce the complexity (dimension) of both the data model itself and the systems developed on its basis. The resulting models are more abstract, consist of fewer concepts with simpler relationships and are not tied to specific subject areas. The use of this data model makes it possible to dynamically transform and interpret the data model in the application, and allows you to customize solutions for a specific subject area. In fact, it becomes possible to reproduce and maintain in the development process the description of various structures and processes used in the subject area under consideration. This approach makes it possible to significantly improve the quality of processing and search for incoming resources and data within a limited subject area, not only through the use of its thesaurus, but also through the flexibility of describing the presentation of available resources. It also allows you to structure and link various resources, extract from them and contextualize a variety of data, turning it into knowledge.

Here are the main types of tasks that are implemented in the semantic library designed to construct the CDSSK subspace:

- description of the information system content;
- implementation of tasks of data integration from external sources;
- support for collections;
- search and navigation through system objects;
- user support.

7 Subject area "Mathematics"

Let us consider as an example the implementation of the CDSSK space for the area "Mathematics" and its subspaces of ordinary differential equations (hereinafter ODE). Based on the proposed approach, a multilevel ontology was constructed. The ODE thesaurus [19] was used as the thesaurus. The peculiarity of this thesaurus is that it contains not only the concepts and terms themselves, but also links to publications in which these concepts are introduced / defined, their mathematical records. Also, various mathematical classifiers are used, such as MSC and the mathematical part of the UDC, articles of the mathematical encyclopedia. The structure of the concepts of the mathematical encyclopedia does not have a hierarchy as such, but thanks to the use of MSC codes related to concepts, it was possible to distinguish thematically related terms of individual sections of mathematics. Formulas were singled out separately and a set of corresponding formulas was compared to each concept, if possible.

Resources such as events, theorems, persons, publications were used here as information resources. Formulas stand out separately, since mathematics implies their presence. It is a semantic object with different relationships. Formulas can be associated with different objects, have different labels, etc. Two large sources were used as data sources: DBpedia and MathNet.

About 4000 publications, formulas, persons, articles of a mathematical encyclopedia were used as content. Formulas were extracted from descriptions of mathematical texts and on the basis of these data additional links were formulated and derived: between MSC and UDC, between formulas and MSC, formulas and UDC, etc.

Let us briefly consider how, to describe the ODE thesaurus, the basic ontology of the thesaurus is extended at the second level in order to take into account all the features of the model of this thesaurus. Consider the concepts necessary for describing at all levels of the ontology and the relationship between them:

1. At the first level, classes are used that are necessary to describe the general model, such as *information resource, thesaurus, concept, thesaurus attribute*, etc.
2. At the second level, the concepts of a specific subject area are described as instances in terms of the first level:
   a. *Mathematical notation* is an instance of the thesaurus attribute class. Used to store the formula string;
   b. *Math note* is also an instance of the thesaurus attribute class. Used to store text with formulas;
   c. *Literature* is an instance of the information resource class for describing the literature included in the ODE thesaurus.
3. At the third level, we use the concepts of the first level and instances of the second level as class definitions at the third level when filling the ontology with data.

To support formulas, the concept of *Formula* was introduced into the ontology at the second level, which allows you to store the original line of the formula from the source and is associated with relations with *information objects* and *concepts* of the thesaurus. Thus, it is possible to build a network of connections of the formula with various objects that make up the content of the subspace under consideration.

Using this approach to describing the ontology for each publication, on the basis of its title, annotation, and keywords, links with the ODE thesaurus were identified. The terms of the mathematical encyclopedia were used as semantic labels. This linking made it possible to identify, with a certain degree of probability, articles related to the ODE subject area in the existing set of publications, to identify intersubject connections and headings, and to organize them in a collection based on the thesaurus and identified semantic labels.

8 Conclusions

In this article, the basic principles of building an ontology of the CDSSK were considered. A set of basic concepts for constructing a description of an arbitrary subject area was considered. An example of the development of an CDSSK ontology for the "mathematics" subject area is demonstrated. Further work is focused on the use of the mathematical apparatus underlying the descriptive logics on which the ontologies are based, and the use of means of inference of new facts based on those available in conjunction with algorithms from the text mining field for text processing. This approach makes it possible to reveal hidden knowledge and find contradictions in the existing ones, which increases the reliability of knowledge.

This work was supported by the Russian Foundation for Basic Research, projects No. 20-07-00324, 18-00-00297, 18-00-00372.

References

1. Antopolskiy, A.B., Kalenov, N.E., Serebryakov, V.A., Sotneykov, A.N.: Common digital space of scientific knowledge. Vestn. Ros. akad. Nauk, 89 (7), 728–735 (2019).
2. Antopolskiy, A.B. i dr.: Principy postroeniya i struktura edinogo cifrovogo prostranstva nauchnyh znanij Nauchno tekhnicheskaya informaciya. Ser. 1, (4), 9–17 (2020).
3. Muromskij, A.A., Tuchkova, N.P.: Predstavlenie matematicheskih ponyatij v ontologii nauchnyh znanij. Ontologiya proektirovaniya, 9 (1) (2019).
4. Gubanov, N.I., Gubanov, N.N., Volkov, A.E.: Kriterii istinnosti i nauchnosti znaniya. Filosofiya i obshchestvo, (3(80)), 78–95 (2016). URL: https://cyberleninka.ru/article/n/kriterii-istinnosti-i-nauchnosti-znaniya.
5. Il'in, V.V., Kalinkin, A.T.: Priroda nauki: Gnoseologicheskij analiz. M., Vysshaya shkola. (1985).
6. GOST 7.0-99 Mezhgosudarstvennyj standart GOST 7.0-99 “Sistema standartov po informacii, bibliotechnomu i izdatel'skomu delu. Informacionno-bibliotechnaya deyatel'nost', bibliografiya. Terminy i opredeleniya”. Vveden v dejstvie postanovleniem Gosstandarta RF ot 7 oktyabrya. N 334-st (1999).
7. Gurevich, I.B., Trusova, Yu.O.: Tezaurus i ontologiya predmetnoj oblasti “Analiz izobrazhenij”. Vserossijskaya konf. s mezhdunar. uchastiem “Znaniya – Ontologii – Teorii” (ZONT–09). Novosibirsk: Institut matematiki im. S.L. Soboleva SO RAN (2009).
8. Hlava, M.M.K.: The Taxobook: History, Theories, and Concepts of Knowledge Organization, Part 1 of a 3-Part Series. Synthesis Lectures on Information Concepts, Retrieval, and Services. 6(3) (2014).
9. Hlava, M.M.K.: The Taxobook: Principles and practices of building taxonomies, Part 2 of a 3-Part Series. Synthesis Lectures on Information Concepts, Retrieval, and Services, 6(4) (2014).
10. Hlava, M.M.K.: The Taxobook: Applications, Implementation, and Integration in Search: Part 3 of a 3-Part Series. Synthesis Lectures on Information Concepts, Retrieval, and Services, 6(4) (2014).
11. Bezdushnyj, A.N., Zhizhchenko, A.B., Kulagin, M.V., Serebryakov, V.A.: Integrirovannaya sistema informacionnyh resursov RAN i tekhnologiya razrabotki cifrovyh bibliotek, Programmirovanie, (4), 3–14 (2000).
12. Ahlyostin, A.Yu., Lavrent'ev, N.A., Fazliev, A.Z.: Sistematizaciya nauchnyh graficheskih resursov po molekulyarnoj spektroskopii. Nauchnyj servis v seti Internet: trudy XIX Vserossijskoj nauchnoj konferencii, 34–42 (2017). doi:10.20948/abrau-2017-39.
13. Sotnikov, A.N. i dr.: Principy postroeniya i formirovaniya elektronnoj biblioteki “Nauchnoe nasledie Rossii”. Programmnye produkty i sistemy (4) (2012).
14. Elizarov, A.M., Zhizhchenko, A.B., Zhil’tsov, N.G., Kirillovich, A.V., and Lipachev, E.K.: Mathematical Knowledge Ontologies and Recommender Systems for Collections of Documents in Physics and Mathematics. Doklady Mathematics, 93 (2), pp. 231–233 (2016), https://doi.org/10.1134/S1064562416020174.
15. Mitrofanova, O.A., Konstantinova, N.S.: Ontologii kak sistemy hraneniya znanij. Vserossijskij konkursnyj otbor obzorno-analiticheskih statej po prioritetnomu napravleniyu “Informacionnotelekommunikacionnye sistemy” (2008).
16. Dextre, Clarke S.G., Zeng, M.L.: From ISO 2788 to ISO 25964: The evolution of thesaurus standards towards interoperability and data modelling. Information Standards Quarterly (ISQ), 24(1) (2012).

17. Kostin, V.V.: Obzor semanticheskih modelej, opisyvayushchih nauchnye publikacii i nauchno-issledovatel'skuyu deyatel'nost'. Elektronnye biblioteki: perspektivnye metody i tekhnologii, elektronnye kollekcii (2014).
18. Gruber, T.: Ontology of folksonomy: A mash-up of apples and oranges. International Journal on Semantic Web and Information Systems (IJSWIS), 3(1), 1–11 (2007).
19. Muromskij, A.A., Tuchkova, N.P.: O tezauruse dlya predmetnoj oblasti "Obyknovennye differencial'nye uravneniya". Vychisl. centr im. A.A. Dorodnicyna RAN (2004).
20. Mascardi, V., Cordi, V., Rosso, P.: A Comparison of Upper Ontologies (2007).
21. Katsis, Y., Papakonstantinou, Y.: View-based data integration. Encyclopedia of Database Systems, 3332–3339 (2009).
22. Xu, L., Embley, D.W.: Combining the Best of Global-as-View and Local-as-View for Data Integration. ISTA, 48, 123–136 (2004).
23. Noy, N.F.: Semantic integration: a survey of ontology-based approaches. ACM Sigmod Record, 33(4), 65–70 (2004).
24. Zhao, L., Ichise, R.: Ontology integration for linked data. Journal on Data Semantics, 3(4), 237–254 (2014).