Informational Infrastructure of the Common Digital Space of Scientific Knowledge

Olga Ataeva[1][0000-0003-0367-5575], Nikolay Kalenov[2][0000-0003-5269-0988],
Vladimir Serebriakov[3][0000-0003-1423-621X], Alexander Sotnikov[4][0000-0002-0137-1255]

[1, 3] Computing Center FRS “Computer Sciences and Control” of Russian Academy of Sciences,
Vavilova str., 40b, 119333 Moscow, Russia

[2, 4] Joint Supercomputer Center of Russian Academy of Sciences – Branch of Federal State Institution “Scientific Research Institute for System Analysis” of Russian Academy of Sciences, Leninskiy pr., 32a, 119334, Moscow, Russia

[1] oli@ultimeta.ru, [2]nkalenov@jscc.ru, [3]serebr@ultimeta.ru,
[4]asotnikov@jscc.ru

Abstract. The space of scientific knowledge is understood as a system of knowledge from various fields of science tested by the scientific community. The Common Digital Space of Scientific Knowledge (CDSSK) is a digital environment into which information resources and objects that have been tested by the scientific community and that reflect scientific knowledge from a different fields of science are integrated. The common digital space of scientific knowledge consists of a set of thematic subspaces related to various areas of science, built according to common principles. The first step towards the implementation of the CDSSK is the development of its ontology, i.e. a multi-level system of concepts describing resources and objects of subject areas, concepts, terms and connections between them, characterized by an open hierarchical and dynamic structuring, oriented both for storing existing knowledge and for extracting new ones. Since the CDSSK is a fundamentally new information environment, there are a number of significant problems on the way to its implementation. The paper deals with the problems and tasks of forming such a space and developing tools for implementation and support the CDSSK.

Keywords[1]: digital scientific information, structure of knowledge space, ontologies, knowledge space content, knowledge space design

[1] CDSSK–2020: International Conference “Common Digital Space of Scientific Knowledge”, November 10–12, 2020, Moscow, Russia
EMAIL: oli@ultimeta.ru (Olga Ataeva); nekalenov@mail.ru (Nikolay Kalenov); serebr@ultimeta.ru (Vladimir Serebriakov); asotnikov@jscc.ru (Alexander N. Sotnikov)
ORCID: 0000-0003-0367-5575 (Olga Ataeva); 0000-0003-5269-0988 (Nikolay Kalenov); 0000-0003-1423-621X (Vladimir Serebriakov); 0000-0002-0137-1255 (Alexander Sotnikov)

1 Introduction

The avalanche-like increase in the volume of scientific information appearing in the world in the form of printed and digital publications, documentary and factual databases, creates serious problems with its processing and use. A researcher dealing with a particular scientific problem, in order to familiarize himself with all the information related to his sphere of interest, and to highlight the necessary and useful information, requires a considerable time, distracting him from his scientific activity. On the other hand, lack of important data obtained in his scientific field can lead to a senseless waste of time for research, the results of which have already been obtained by other scientists.

In this regard, it becomes necessary to filter scientific information and extract knowledge from it that is novel and unique [1, 2]. The aggregate of such information in digital form, together with the means to ensure its actualization, preservation and provision to users, is defined as a single digital space of scientific knowledge (CDSSK) [3, 4]. The CDSSK is an integral part of the “electronic knowledge space”, the formation of which is an important issue raised in a number of policy documents related to the development of the information society and the digital economy.

2 Basic definitions

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 CDSSK is a digital environment into which information resources and objects that have been tested by the scientific community and that reflect scientific knowledge from a specific field of science are integrated.

The CDSSK consists of a set of thematic subspaces related to various areas of science, built according to common principles.

Thematic subspace is a part of the general space, limited by the framework of a specific subject area. Despite the fact that there are some examples of formalization of knowledge in different subject areas [5–15], there is no generalized approach to defining the digital space of scientific knowledge. An analysis of examples of the formalization of the knowledge space [16, 17] in various areas indicates that the main components of the CDSSK in general and each of its thematic subspaces in particular are ontology and content.

The ontology includes a universal description of the CDSSK data structure. This is a multi-level system of concepts that describes the 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 mapping objects into the CDSSK.

The content of the CDSSK is actually scientific information (objects, properties, ontologies), as well as a terminological description of the field of knowledge [15, 20–23] (thesauri, dictionaries, classifiers). The content, in particular, includes a set of digital copies of real world objects and a description of their metadata profiles [18, 19, 23, 24].

The local subspace of the CDSSK is a certain “isolated” part of the CDSSK (possibly polythematic), which has the same properties as the general space. Local subspace:

• is defined on a certain subset of the CDSSK data (content);

• is built on the same principles as the CDSSK;

• is intended for use by a group of users;

• can exist in isolation from the CDSSK itself, without violating the structure of the knowledge system contained in it;

• is dynamically linked to the main (base) space means by import/export.

Integration is the process of linking data from external sources with the CDSSK content.

The CDSSK design is the process of mapping the state of the subject area into a digital representation available for research and machine processing.

The Semantic Library is an the CDSSK design tool that provides:

- the formation of a thematic subspace (TS) of the CDSSK, including automated processing and loading of content into TS with the establishment of all ontological links between its elements;
- accumulation and editing of new information;
- multidimensional data search and navigation through the content of the CDSSK;
- the formation of a local subspace according to the given characteristics.

3 The ontology of the CDSSK

The first step towards the implementation of the CDSSK is the development of its ontology. By ontology in this context, we mean a multi-level system of concepts that describe resources and objects of subject areas, concepts, terms and connections between them, characterized by an open hierarchical and dynamic structuring, oriented both for storing existing knowledge and for extracting new ones.

The CDSSK ontology has three levels.

First level. Conceptualizing the CDSSK. This level describes the basic concepts that make up the CDSSK. It includes the following concepts:

• 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 allows to unambiguously identify an object in the CDSSK;

• Thematic subspace – a set of information objects with their properties and connections related to a certain scientific direction;

• Local class of objects – objects belonging to one thematic subspace;

• Universal object class – objects associated with several thematic subspaces. Universal classes include objects such as persons, publications, documents, events, organizations, etc.;

• Attributes of a digital object – structured properties (metadata) that characterize the object from the point of view of the objectives of the CDSSK;

• Metadata profile – a set of attributes describing objects of a particular class;

• Subject ontology (thesaurus) of a thematic subspace – a set of indexes of classification systems, key terms with their thesaurus connections related to a given scientific direction;

• Subject ontology of the CDSSK – a set of subject ontologies of individual subspaces;

• Thesaurus connections are links between two elements of a subject ontology with a given set of values that can be established within a given thematic subspace;

• Attribute links – links between different objects both within one class and between objects of different classes;

• Type of attribute links – takes one of two values: unambiguous link (object A is linked by link b with only one object B of a particular class) and multivalued link (object A can be linked by link b with several objects of a given subset B). There is an unambiguous relationship between the object "city" and "country" from the class "geographic concepts", while the feedback "country" – "city" is multivalued;

• Kind of attribute links – specific relationships between different objects of the same or different classes, both local and universal. The objects of the universal classes "person" and "publication" can have multivalued connections of the form "author", "editor", "compiler", etc. There can be connections between a museum object and a person of the form "model author", "collection author", "donator", etc.

At the second level of the ontology, the concepts of the first level are concretized in relation to the general space and one or another thematic subspace, in particular, in accordance with the definition given at the first level, thesauri of specific subject areas are determined. Specific local classes of objects are determined by the specifics of the field of science to which they belong. For microbiology, these can be strains, collections and databases of microorganisms; for linguistics – dictionaries, languages, text corpora, etc., for chemistry – reactions, substances, chemical elements, etc.

At the third ontological level, the metadata profiles of the CDSSK objects and the relationships between them are established. As examples of unambiguous attribute relationships within the same class, one can cite relationships between publications of different levels: an article (analytical level) is linked by the relationship "published in ..." with a certain issue of the journal (monographic level), which, in turn, is associated with the description of the journal (summary level).

4 Levels of design and functionality of the CDSSK software

The result of the design of the CDSSK should be a "three-layer" software package, the layers of which correspond to the levels of the ontology of space and have the corresponding functionality. The design of the CDSSK includes three levels:

- System-wide (based on the first level ontology), intended for the work of specialists in the field of informatics – system analysts of a wide profile. It provides tools for the formation and editing of the main vocabularies of the system, object types and types of links between objects, etc.;

- The thematic layer of the software complex, determined by the second level of ontology, has functionality that ensures the formation of thematic subspaces in a particular field of science. It is focused on the work of authorized users – specialists in a specific scientific field. At this level, they define the basic concepts of the subspace and the relationship between its elements, describe and connect the necessary data sources, using the introduced concepts of the first and second levels of the ontology, and gain access to batch loading of data. Within this layer, user interfaces of the system, specific to the thematic subspace, are formed;
- Applied layer (implementation of specific decisions made at the first and second levels), developed by specialists in the field of creating software systems. This layer of the software package is based on the third level of the ontology and provides external users with access to the CDSSK. The functionality of this layer includes the provision of the possibility of a multidimensional search for data within the CDSSK, their visualization and export, and navigation through the space. At this level, the user can organize data into collections, classify them according to various criteria, and form one or another local subspace.

5 Problems of creating the CDSSK

Since the CDSSK is a fundamentally new information environment, there are a number of significant problems on the way to its implementation.

Formalization problems

Currently, there are no generally accepted approaches to defining the content of the CDSSK, taking into account the fact that specific scientific knowledge is specific to certain fields of science. The problem of identifying the subspace of the CDSSK that belongs to a separate subject area is quite complicated, since there is no formal generally accepted approach to this problem; it would seem that to solve this problem it would be possible to rely on existing classification systems (such as SRSTI, UDC, MKI, etc.), however, as analysis shows [22], no classification provides the necessary completeness and accuracy of reflection of scientific knowledge in a particular area.

Linguistic problems

The basis of each thematic subspace is a subject ontology – an extended thesaurus with types of connections between its elements specific for each field of science, which comprehensively describes this subject area. For the formation of such thesauri and their dynamic support, it is necessary to develop a special methodology based on Semantic Web technologies.

Expertise problem

The CDSSK should contain reliable scientific information with minimal duplication of both the data itself and their sources. Accordingly, it is necessary to develop a mechanism for the selection of materials intended for inclusion in the CDSSK. Expert selection should be one of the components of such a mechanism.

Content generation and data integration issues

There is no unified approach to integrating data from various sources into a single information environment. Most of the data sources to be loaded into one or another thematic subspace of the CDSSK do not have a means of exporting in any unified format. To retrieve and load formalized data that meet the requirements of ontologies into the thematic subspace of the CDSSK, in each specific case, it is necessary to have specialized technical knowledge and use special software tools.

6 Data integration issues

Let's consider the concept of interaction of the software package with external data sources. Data loading from external sources into the knowledge space (downloading) is carried out through its ontology. The priorities for creating the CDSSK data integration tools include the following.

Development of an ontology of the integration process. It should be related to the CDSSK ontology and include concepts such as: Display, Source, Link and others.

Development of source requirements. The data source to be integrated must be provided with a semantic layer. Each object in the source must have a set of properties to uniquely identify it and a unique URI.

Constructing the mapping. The mapping is built at the level of conceptual concepts of the subject area, while:

- a set of standard operations is introduced for cleaning and bringing data;
- a set of rules is introduced according to which an object from a source can be represented within the framework of the concepts of the subject area.

7 The CDSSK Toolkit

Basic technological requirements for the CDSSK toolkit:

- universality – the ability to describe different types of resources;
- structuredness – support of links between resources;
- adaptability – the ability to port to various platforms;
- dynamism – the ability to adapt the system to the evolution of resources and customize user interfaces.

Functionality of the CDSSK toolkit

The main functionality should ensure the solution of the following tasks:

• formation of a multilevel model of the subject area;
• integration of data sources;
• provision of user and machine readable interfaces.

List of functions:

• create / view / edit of:
  - information resources and their copies;
  - thesaurus;

- collections of information objects;

- connection of data sources;
- batch loading of data that make up the content of the library;
- attribute / semantic / full-text search;
- provision of data in a machine-readable format;
- highlighting additional links between instances of information resources;
- support for semantic labels to describe thematic orientation.

8 Basic principles for the implementation of the toolkit for supporting the local subspace of the CDSSK

The CDSSK functionality is based on the following principles of a personal open semantic digital library [27, 28]:

1. The library must support the ability to use media resources or refer to them when describing objects, including text, images, 3D models, audio and video files, or any combination of them;

2. The types of resources used and the connections between them should be described by means of the system within the framework of certain concepts that make up the semantic description of the resources of the library content. The set of these concepts is thematically limited to the terms of the subject area from a certain thesaurus;

3. The library should be an integration node, providing the ability to link local subspace data with data from different sources, in particular, presented in a form that meets the requirements of Linked Open Data (LOD). It should also be able to retrieve the data from this library in a structured format;

4. Library users should be able to organize their local subspaces of the collection according to the scientific direction they are interested in, adding new terms to the subject thesaurus, thus clarifying the area of their interests. Also, users should be able to search not only among objects within the system, but also on data sources without the need to use special knowledge for search queries. A personal open semantic digital library must support:

- input and editing of heterogeneous digital objects (such as text, images, audio and video files, 3D models of objects and phenomena);
- customizable options for importing and exporting data presented in formats that meet the requirements of the Semantic Web;
- multidimensional data search and navigation through related objects;
- data visualization, taking into account the specifics of subject areas;
- authorization of users with customizable access rights to certain data;
- protection of information from unauthorized access;
- the ability to quickly recover data in case of technical failures.

9 Priority tasks for the design of the CDSSK

1. Development of an ontology of the first level of the CDSSK, containing concepts common to various scientific fields, on the basis of which it is possible to reflect knowledge from a certain field of science in the CDSSK.
2. Development of requirements for the integration of data sources and ensuring the solution of the data integration problem based on the developed ontology.
3. Development of specific requirements for the semantic library used as a tool for constructing the CDSSK subspace and providing navigation through it. Such a library should be equipped with data source integration tools; must support mechanisms for linking library data with data from other sources; have customizable user interfaces; provide import / export of data in machine-readable formats.

10 Conclusion

The development and maintenance of the CDSSK is a national challenge. It should involve leading organizations in the field of information technology (in terms of defining the first-level ontology), and in various fields of science (for the development of second-level ontologies of individual thematic subspaces), the main organizations-keepers of scientific resources (libraries, archives, museums).

Research is carried out within the framework of state assignments No. 0580-2021-0016 (Joint Supercomputer Center of Russian Academy of Sciences) and No.18-00-00297 (Federal Research Center “Computer Science and Control” of the Russian Academy of Sciences).

References

1. Gubanov, N.I., Gubanov, N.N., Volkov, A.E.: Kriterii istinnosti i nauchnosti znaniya. Filosofiya i Obshchestvo 3 (80). S. 78–95 (2016).
URL: https://www.socionauki.ru/upload/socionauki.ru/journal/fio/2016_3/078-095.pdf
2. Il'in, V.V., Kalinkin, A.T.: Priroda nauki: Gnoseologicheskij analiz. M.: Vysshaya shkola, 230 p. (1985).
3. Antopolskiy, A.B., Kalenov, N.E., Serebryakov, V.A., Sotnikov, A.N.: Common digital space of scientific knowledge. Vestn. Ros. Akad. Nauk. T. 89. № 7. p. 728–735 (2019).
https://doi.org/10.31857/S0869-5873897728-735
4. Antopolskiy, A.B., Bosov, A.V., Savin, G.I., Sotnikov, A.N., Cvetkova, V.A., Kalenov, N.E., Serebryakov, V.A., Efremenko, D.V.: The principles of construction and structure of a unified digital space of scientific knowledge (UDSSK). Nauchno-tekhnicheskaya Informaciya. Ser. 1. № 4. S. 9–17 (2020).
https://doi.org/10.36535/0548-0019-2020-04-2
5. Muromskiy, A.A., Tuchkova, N.P.: Representation of mathematical concepts in the ontology of scientific knowledge [In Russian]. Ontology of Designing 9(1). P. 50–69 (2019).
https://doi.org/10.18287/2223-9537-2019-9-1-50-69

6. Gurevich, I.B., Trusova, Yu.O.: Tezaurus i ontologiya predmetnoj oblasti “Analiz izobrazhenij”. In: Vserossijskaya konf. s mezhdunar. uchastiem “Znaniya – Ontologii – Teorii” (ZONT–09). Novosibirsk: Institut matematiki im. S.L. Soboleva SO RAN. 10 s. (2009). URL: http://math.nsc.ru/conference/zont09/reports/95Gurevich-Trusova.pdf
7. 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). 80 p. (2014). https://doi.org/10.2200/S00602ED1V01Y201410ICR035
8. 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). 164 p. (2014). https://doi.org/10.2200/S00603ED1V02Y201410ICR036
9. 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). 156 p. (2014). https://doi.org/10.2200/S00604ED1V03Y201410ICR037
10. Bezdushnyj, A.N., Zhizhchenko, A.B., Kulagin, M.V., Serebryakov, V.A.: Integrirovannaya sistema informacionnyh resursov RAN i tekhnologiya razrabotki cifrovyh bibliotek. Programmirovanie. (4) S. 3–14 (2000).
Ahlyostin, A.Yu., Lavrent'ev, N.A., Fazliev, A.Z.: Sistematizaciya nauchnyh graficheskih resursov po molekulyarnoj spektroskopii. In: Nauchnyj servis v seti Internet: trudy XIX Vserossijskoj nauchnoj konferencii (18–23 sentyabrya 2017 g., g. Novorossijsk). M.: IPM im. M.V. Keldysha. p. 34–42 (2017). URL: http://keldysh.ru/abrau/2017/39.pdf https://doi.org/10.20948/abrau-2017-39
11. Kalenov, N.E., Savin, G.I., Serebriakov, V.A., Sotnikov, A.N.: Scientific heritage of russia digital library: construction and sources aggregation philosophy. Programmnye Produkty i Sistemy 4 (100). P. 30–40 (2012).
12. Elizarov, A.M., Zhizhchenko, A.B., Zhil’tsov, N.G., Kirillovich, A.V., Lipachev, E.K.: Mathematical knowledge ontologies and recommender systems for collections of documents in physics and mathematics. Doklady Mathematics 93 (2). p. 231–233 (2016). https://doi.org/10.1134/S1064562416020174
13. Mitrofanova, O.A., Konstantinova, N.S.: Ontologii kak sistemy hraneniya znanij. In: Vserossijskij konkursnyj otbor obzorno-analiticheskih statej po prioritetnomu napravleniyu “Informacionno-telekommunikacionnye sistemy”. p. 54 (2008). URL: https://lib.nsu.ru/xmlui/handle/nsu/8979
14. Muromskij, A.A., Tuchkova, N.P.: O tezauruse dlya predmetnoj oblasti “Obyknovennye differencial'nye uravneniya”. Vychisl. centr im. A.A. Dorodnicyna RAN. S. 52–55 (2004).
15. System of standards on information, librarianship and publishing. Information and librarian activity, bibliography. Terms and definitions. 27 p. (1999). URL: https://docs.cntd.ru/document/1200004287
16. Kostin, V.V.: Obzor semanticheskih modelej, opisyvayushchih nauchnye publikacii i nauchno-issledovatel'skuyu deyatel'nost'. Elektronnye biblioteki: perspektivnye metody i tekhnologii, elektronnye kollekcii. S. 131–136 (2014).
17. Kalenov, N., Sobolevskaya, I., Sotnikov, A., Kirillov, S.: The use of 3D visualization technology web-collections for the formation of virtual exhibitions. CEUR Proceedings of the 21st Conference on Scientific Services & Internet 2543. P. 382–390 (2020).
18. Kalenov, N., Sobolevskaya, I., Sotnikov, A.: Mathematical modeling of the processes of interdisciplinary collections formation in the digital libraries environment. CEUR Proceedings of the 21st Conference on Scientific Services & Internet 2543. P. 391–398 (2020).

19. Nikolay Kalenov, Irina Sobolevskaya, Alexander Sotnikov: Hierarchical Representation of Information Objects in a Digital Library Environment. Communications in Computer and Information Science 1093. p. 93–104 (2019). https://doi.org/10.1007/978-3-030-30763-9_8
20. Dextre Clarke, Stella G., Zeng, Marcia Lei: From ISO 2788 to ISO 25964: The evolution of thesaurus standards towards interoperability and data modelling. Information Standards Quarterly (ISQ) 24 (1). 8 p. (2012). URL: http://eprints.rclis.org/16818/1/SP_clarke_zeng_isqv24no1.pdf
21. Tsvetkova, V., Kharybina, T., Mokhnacheva, Yu., Beskaravaynaya, E., Mitroshina, I.: Combining classification systems and building the array of keyworks for defining the space of microbiological knowledge. Scientific and Technical Libraries (11). p. 25–43 (2019). https://doi.org/10.33186/1027-3689-2020-12-83-98
22. Kalenov, N.E., Sobolevskaya, I.N., Sotnikov, A.N.: Ierarhicheskie urovni predstavleniya informacionnyh ob'ektov v srede elektronnyh bibliotek. Informaciya i Innovacii 13 (2). S. 25–31 (2018). https://doi.org/10.31432/1994-2443-2018-13-2-25-31
23. Gruber, T. Ontology of folksonomy: A mash-up of apples and oranges. International Journal on Semantic Web and Information Systems (IJSWIS) 3 (1). P. 1–11 (2007).
24. Antopol'skij, A.B., Ataeva, O.M., Serebryakov, V.A.: Sreda integracii dannyh nauchnyh bibliotek, arhivov i muzeev "LibMeta". Informacionnye Resursy Rossii (5). S. 8–12 (2012).
25. Serebryakov, V.A., Ataeva, O.M.: The basic concepts of a semantic libraries formal model and its integration process formalization. Programmnye Produkty i Sistemy (4). S. 180–187 (2015). https://doi.org/10.15827/0236-235X.112.180-187
26. Ataeva, O.M., Serebryakov, V.A. Personal semantic open digital library libmeta. construction of the content. integration with lod sources. Informatika i eyo Primenenie 11 (2). P. 85–100 (2017). https://doi.org/https://doi.org/10.14357/19922264170210
27. Ataeva, O.M.: An information model of LibMeta semantic library. Programmnye Produkty i Sistemy (4). P. 36–44 (2016). https://doi.org/10.15827/0236-235X.114.036-044
28. Ataeva, O.M., Serebryakov, V.A.: Ontologiya cifrovoj semanticheskoj biblioteki LibMeta. Informatika i eyo Primeneniya 12. S. 2–10 (2018). https://doi.org/10.14357/19922264180101