Ontology of the Universal Subspace of the Common Digital Space of Scientific Knowledge

N.E. Kalenov[1], I.N. Sobolevskaya[1], A.N. Sotnikov[1], S.A. Vlasova[1]

[1] Joint Supercomputer Center of the Russian Academy of Sciences — Branch of Federal State Institution "Scientific Research Institute for System Analysis of the Russian Academy of Sciences" (JSCC RAS — Branch of SRISA), 119334, Moscow, Leninsky av., 32 a, Russia

ABSTRACT

The work is a development of research conducted by the authors in the field of creating a Common Digital Space of Scientific Knowledge (CDSSK). In the framework of previous studies, a unified structure for representing the CDSSK ontology elements (subspaces, classes and attributes of objects, relations between objects or attributes) was proposed. This paper presents a variant of the universal subspace (USS) ontology of the CDSSK, proposed by the authors, built in accordance with the developed structure. There are defined 10 subject and 10 objects auxiliary classes in the described model of USS. Among the subject classes are “persons”, “publications”, “museum objects”, “events”, etc. Among the auxiliary ones are “formats”, “units of measurement”, “languages”, etc. The paper contains reference books of each class, is built in accordance with the ontology structure model as well as an object attributes list, their directories and static dictionaries examples.

Keywords: digital space of scientific knowledge, ontologies, structuring, linked data.

1. INTRODUCTION

For a researcher working on a specific scientific problem, it takes a significant amount of time to familiarize themselves with all the information relevant to their field of interest and extract the necessary and useful knowledge. This time-consuming process can distract them from their actual scientific work. On the other hand, the absence of important data obtained in their scientific field can lead to wasted time on research that has already been conducted by other scientists.

In this regard, there is a need for the organization of scientific information and the extraction of novel and unique knowledge from it. The collection of such information in digital form, along with the tools that ensure its relevance, integrity, and provision to users, is referred to as a Common Digital Space of Scientific Knowledge.

It is necessary to create a digital environment that contains not only documentary information but also facts reflecting the results of scientific activity, with references to the priority sources where they are published, and the full texts of these sources. CDSSK serves as such an environment, which is a multifunctional and versatile structure that caters to various user segments. To accomplish the task of forming the CDSSK, it is necessary to develop a general ontology for the CDSSK , determine its functional and organizational structure, analyze existing potential sources of information, and formulate criteria for selecting objects to be included in the content space. Additionally, a technology for content population and updating needs to be developed, service components of the CDSSK and legal aspects of its use and provision to users should be addressed, and requirements for linguistic, technical, and software tools of individual components of the CDSSK and the space as a whole should be formulated and implemented. The results obtained from these studies are reflected in a series of reports presented at the All-Russian

Conference "Common Digital Space of Scientific Knowledge: Issues and Solutions"[1] as well as in [1, 2, 5, 7, 8, 12, 13, 14, 18].

The work presented in this paper is a continuation of research conducted by the Joint SuperComputer Center of the Russian Academy of Sciences (JSCC RAS) related to the development of a Common Digital Space of Scientific Knowledge (CDSSK) as a structured integrated information environment reflecting advancements in various fields of science [1-4]. At previous stages of the work, the architecture of building CDSSK was defined [5], issues of network support for CDSSK were considered [6], and the reflection of 3D models of multimedia objects in CDSSK was studied [7-11].

The model of the ontology structure of CDSSK proposed by us was published in [12]. A modified version of the model is presented in [13]. According to this model, CDSSK is represented as a 5-level hierarchical structure (CDSSK - subspaces - object classes - object class attributes - attribute values), supplemented with three types of relations - (universal, quasi-universal and specific). Information about all elements of CDSSK is stored in directories and dictionaries. Elements of each hierarchical level are described in directories with a fixed structure for each level. Directories also describe relations of each of the three types. Attribute values and relations are stored in dictionaries, the information about which is contained in the corresponding directories. Dictionaries are divided into two groups - static and dynamic. The first group contains values of "standardized" attributes (such as a list of scientific degrees, titles and positions, subject categories etc.). These dictionaries are filled out during the CDSSK initial installation or its fragments and are corrected by administrators with appropriate privileges. Dictionaries of the second type are filled as CDSSK content is formed with specific data (person surnames, publication titles, links to network resources, etc.).

The CDSSK universal and thematic subspace includes a universal subspace and thematic subspaces.

The universal subspace contains objects of a multidisciplinary nature (persons, events, units of measurement, etc.) and their connections with other objects in the CDSSK.

The thematic subspace (such as the "informatics", "space research", "chemistry", etc. subspaces) contains elements directly related to a specific scientific field, as well as connections with elements from the universal and other thematic subspaces.

Below, we consider the classes of objects in the universal subspace, examples of attribute reference books, and attribute value dictionaries.

2. UNIVERSAL SUBSPACE OBJECT CLASSES AND THEIR ATTRIBUTES

The proposed structure implementation was modeled by creating elements of the **Universal subspace of** CDSSK, which contains multidisciplinary objects and their connections. In the initial approach, 10 subject classes and 10 auxiliary object classes were identified in the universal subspace.

The subject classes include:

- Persons
- Publications

[1] https://dirsmsc.ru/konf/

- Qualification works
- Documents
- Museum exhibits
- Images and multimedia objects
- Events / activities
- Organizations
- Universal databases and resource catalogs
- Awards

The auxiliary classes include:

- Data formats
- Universal classification systems
- Person groups
- Location (geographical characteristics)
- Time characteristics
- World constants
- Measurement units
- Numeric values
- Languages
- Collections

For every subject and auxiliary class, we created lists of attributes and included excerpts from static dictionaries.

When determining the attributes of each object, we considered the extensive experience gained from operating the digital library "Scientific Heritage of Russia" [14], which has now evolved into a model component of CDSSK [15-18]. This modern version incorporates a diverse range of interconnected data. These papers also describe a technique for creating thematic subspaces.

To illustrate, we present a selection of attribute lists for certain subject and auxiliary classes. Each attribute is accompanied by an indication of whether it is mandatory (m) or optional (o) in parentheses.

3. SOME EXAMPLES OF SUBJECT CLASSES ATTRIBUTES

Attributes of the "Persons" class objects:

- Surname (m),
- Name (m);
- Patronymic (o);
- Pseudonym (o);
- Date of birth (m);
- Place of birth (m);
- Date of death (o);
- Place of death (o);

- Academic degree (o);
- Academic title (o);
- Biography (m);
- Bibliography of the person as an author (m),
- Bibliography about the person (o).

Additional information (scientific interests, work activity, scientific discoveries, identifiers in databases, etc.) is presented as named relations with corresponding objects.

Attributes in the "Qualification works" class objects:

- Title (m);
- Work type (m);
- Date of issue (thesis defense date) (o);
- URL of the full text (o).

Additional information is presented as relations with persons (author, scientific supervisor, opponent, etc.), organizations (the execution place, the defense place, the leading organization, etc.), classification systems, etc.

Attributes in the "Museum exhibits" class objects:

- Title (m);
- Name of the museum where the original is stored (m);
- Type of source of receipt (m);
- Date of receipt at the museum (m);
- Date of discovery (creation) of the object (m). In many cases the date can only be determined approximately, but, in any case, the relevant information should be indicated;
- Description of the object (m).

Museum objects can be connected with specific relations to a person (one of the possible connection values is "collector" for natural science collections), a geographic object (place of object detection), publications, etc.

4. SOME EXAMPLES OF AUXILIARY CLASSES ATTRIBUTES

The "Data Representation Formats" class is of utmost importance in the ontology of CDSSK as it serves as a foundational element. Objects within this class define a well-organized collection of rules that describe attributes of other classes. References to data representation formats are found throughout all directories of objects and relations. These formats provide essential information for ensuring the formal and logical integrity of input data during streaming uploads and manual data input processes. Additionally, this class can serve as a valuable information resource for CDSSK users, providing insights into a wide range of data formats encountered.

The attributes associated with objects in the "**Data Representation Formats**" class include:

- Data representation type (mandatory): attribute can take values such as "text", "integer", "date in the format yyyy[.mm[.dd]]", "relations", etc.

- Format type (optional): attribute specifies the format, such as "pdf", "jpeg", "URL", and a set of values in its dictionary describe the structure of various CDSSK relations.
- Mandatory or optional (o) attribute value (m) (mandatory);
- Unique (u) or multiple (t) attribute value (m) (mandatory);
- Structural restrictions (optional) - This attribute specifies a specific structure that a certain attribute must adhere to. Examples include "all-Union State Standard 7-1.2003: Bibliographic Description" (referencing the official description of the all-Union State Standard), "ISBN control algorithm" (with a description of the control algorithm), or "required requirements for the structure of email addresses" (with the formulation of the requirements), etc.
- Reference to a detailed description of the format (optional).

The attributes associated with objects in the "**Units of Measurement**" class include: the following attributes:

- Name of the unit of measurement (m).
- Measured object (m).
- Abbreviations (m).
- Additional information (o).

To automate the generation of attribute directories and dictionaries for their values, a dialog program was developed. This program was used to create dictionaries for the aforementioned classes, attributes, and static directories.

Here is a list of some attribute directories, fragments of dictionaries, and static dictionaries.

Class.1: Persons; UN; UNPS; A|UNPS; ; information about people, in some way associated with scientific research

A|UNPS.1: Surname; UNFT.10; N|A|UNPS.1; D; the surname is selected from the dictionary, and if it is not present, it is entered and checked for equivalence with other spellings

A|UNPS.8: Qualification (academic degree); UNFT.10; N|A|UNPS.8; S; selected from the dictionary;

Fragment of the static values dictionary for the attribute "qualification (academic degree)":

N|A|UNPS.8.1: doctor of physical and mathematical sciences

N|A|UNPS.8.2: doctor of technical sciences

N|A|UNPS.8.3: doctor of chemical sciences

Class.3: Qualification papers; UN; UNDS; A|UNDS; dissertations, abstracts, etc.

A|UNDS.1: Title; UNFT.1; N|A|UNDS.1; D;

A|UNDS.2: Date of release (defense) of the paper; UNFT.4; N|A|UNTC.2; D;

A|UNDS.3: Type of work; UNFT.10; N|A|UNDS.3; S;

A|UNDS.4: URL of the full text; UNFT.17; N|A|UNDS.4; D; filled in if the text is not available in CDSSK, when the full text is available in CDSSK, a specific "qualification work-document" or "qualification work-image" relationship is created

A|UNDS.5: Additional information; UNFT.12; N|A|UNDS.5; D;

Fragment of the static dictionary of values for the attribute "Type of work":

N|A|UNDS.3.1: doctoral dissertation

N|A|UNDS.3.2: Ph.D. dissertation

Class.16: Data representation formats; UN; UNFT; A|UNFT; formats for representing attributes of objects (numbers, time, date, text, etc.).

A|UNFT.1: Data representation type; ; N|A|UNFT.1; S;

A|UNFT.2: Format type; ; N|A|UNFT.2; S;

A|UNFT.3: Mandatory (m) or optional (o) attribute value; ; N|A|UNFT.3; S;

A|UNFT.4: Unique (u) or multiple (t) attribute value; ; N|A|UNFT.4; S;

A|UNFT.5: Structural restrictions; ; N|A|UNFT.5; D;

A|UNFT.6: Link to the format description; ; N|A|UNFT; D;

Fragments of the dictionaries of attribute values for the "Formats" class:

N|A|UNFT.1.1: text

N|A|UNFT.1.2: image

N|A|UNFT.1.3: video

N|A|UNFT.1.4: connections

N|A|UNFT.2.1: PDF

N|A|UNFT.2.2: JPG

N|A|UNFT.2.3: MP4

N|A|UNFT.2.6: simple link of the first type between objects, attributes, or values O1 and O2 in the form <URNc>:<URNO1><URNO2>, where URNc is the URN of a specific relation. Example: the name of a language is equivalent to its code; the surname "Petrov" is equivalent to "Petrov"; an article is included in an encyclopedia, etc.

N|A|UNFT.2.7: simple relation of the second type, indicating the subject, object, URN of the link, and URN of the link value. The format of the relation representation is: <URNc>:<URN of the subject><URN of the object>=<URN of the dictionary values element corresponding to the link attribute value>. Example: person P1 is an employee of organization O1 in the position of an engineer (attribute value).

N|A|UNFT.3.1: r

N|A|UNFT.3.2: f

N|A|UNFT.4.1: u

N|A|UNFT.4.2: t

N|A|UNFT.5.1: Arabic digit

N|A|UNFT.5.2: bibliographic description according to Russian Standard https://docs.cntd.ru/document/1200034383

N|A|UNFT.6.1: https://habr.com/ru/post/454944 [description of the JPEG format]

N|A|UNFT.6.2: https://open-file.ru/types/mp4 [description of the mp4 format]

5. SUBJECT ONTOLOGY OF THE CDSSK

The subject ontology of the Universal subspace CDSSK is represented by the class "universal classification systems", which includes the Code of State Categories Scientific and Technical Information, Universal decimal classification, the nomenclature of specialties of the Higher Attestation Commission, patent classification, and others, included in the class as needed. In addition to the universal subspace, subject ontologies are formed in each thematic subspace of CDSSK.

Subject ontology of each thematic CDSSK subspace consists of organized encyclopedic concepts that are interconnected as well as related to objects of universal classes. The structure of the subject ontology can be developed based on the sections found within existing lists of scientific information headings. Examples of such lists include UDC (for general scientific information) [16], IN IS (for nuclear physics) [17], and more. Below is an illustration of how the subject ontology of the thematic subspace "Astronomy" is structured using the State Rubricator of Scientific and Technical Information [18] as a foundation.

ASTRONOMY

- General problems of astronomy
- Theoretical astronomy. Celestial mechanics
- Astrometry
- Astrophysics
- Solar system
- Sun
- Stars
- Nebulae. Interstellar medium
- Star systems
- Cosmology
- Observatory. Instruments, devices and methods of astronomical.

Each of the 11 sections highlighted at the second level of the hierarchy is subdivided into subsections of the third level. In particular, the following subsections are highlighted in the "Solar system" section:

- Solar system
- General problems of solar system research
- Structure and origin of the solar system
- Planets and their satellites
- Moon. Lunar eclipses
- Comets
- Meteors. Zodiacal light. Interplanetary environment.
- Meteorites

Within each subsection of the third level, there are further subsections or individual objects. Each section or subsection in the subject ontology is considered an object within the CDSSK. Every object establishes both universal and specific connections with other objects within the same subspace, across different subspaces, and with objects in universal classes.

For astronomical objects, these connections can involve persons in the form of "discovered," "described," or "calculated." In the case of publications, relations can include "first published," "textbook for school," or "the most complete monograph." Connections with objects from the "Mathematics" subspace can be of the type "described by equations," and so on.

Objects belonging to the subclass "Astronomical observatories," found in the final section of the second level in the subject ontology of the software program "Astronomy," are linked to objects in the "Location" class through the mandatory link "located in," and so forth.

The universal class "Location" represents geographic objects and provides general information without detailed geography, but with varying accuracy to determine the location (ranging from continent to house number and coordinates with seconds accuracy). This universal class is designed to process generalized queries such as "archaeological excavations in Cyprus", or "herbaria collected in Tenshu," or "astronomical observations conducted in Mexico." Despite specific references to "Paphos" for archaeological findings, or the "mountains of southern Kazakhstan" for herbarium descriptions, or the "Sierra de San Pedro Mártir" for astronomical observations, the universal class allows for processing such queries.

Objects in the universal class "Location" are associated with elements of the thematic subspace "Geography," which contains comprehensive descriptions of geographic objects. The "Location" class includes subclasses such as Land and Water, which further contain additional subclasses.

Land

- continent
- part of the world;
- natural area;
- part of the land that has a geographical name
- country
- subject of the country
- locality (city, town, village)
- the named part of the settlement (district, street, square, etc.)
- address
- coordinates

Water space

- oceans
- seas
- lakes
- rivers
- other bodies of water that have a name (waterfalls, swamps ...)

Along with universal connections, specific connections of the type “washed” (connection between a continent or country and the sea or ocean), “is an inflow” (connection between rivers), “stands on” (connection between a city and river), etc.

CONCLUSIONS

The use of the proposed CDSSK ontology model allows for standardizing the algorithms for content creation, developing a typical interface for adding new elements regardless of the subspace and the specific type of data, simplifying and speeding up the algorithms for searching and navigating through linked data [19-23]. Currently, research is being carried out at the JSCC RAS on the development and specification of the proposed model in algorithmization terms of the of nested links formation, as well as modeling the formation of the CDSSK fragments based on real data, including the content of the digital library "Scientific Heritage of Russia."

The work is being carried out at the Joint SuperComputer Center of the Russian Academy of Sciences – Branch of Federal State Institution “Scientific Research Institute for System Analysis of the Russian Academy of Sciences” within the framework of state assignment FNEF-2023-0014.

REFERENCES

1. Antopol'skiy A.B., Kalenov N.Ye., Serebryakov V.A., Sotnikov A.N., // O yedinom tsifrovom prostranstve nauchnykh znaniy // Vestnik Rossiyskoy akademii nauk. - 2019. - T. 89, - № 7. - pp. 728-735
2. Kalenov N. et al. Assessment of Efforts for Content Creation for the Common Digital Space of Scientific Knowledge //CHIRA. – 2021. – pp. 131-138.
3. Matos F. et al. Knowledge, People, and Digital Transformation. – Springer. - 2020. - p. 303
4. van Meeteren M. et al. Rethinking the digital transformation in knowledge-intensive services: A technology space analysis //Technological Forecasting and Social Change. – 2022. – T. 179. – pp. 121631.
5. Kalenov N.Ye., Sotnikov A.N. Arkhitektura yedinogo tsifrovogo prostranstva nauchnykh znaniy // Informatsionnyye resursy Rossii. - 2020. - № 5. - pp. 5-8. DOI: 10.51218/0204-3653-2020-5-5-8.
6. Abramov A.G., Gonchar A.A., Yevseyev A.V. Natsional'naya issledovatel'skaya komp'yuternaya set' novogo pokoleniya kak infrastrukturno-servisnaya platforma Yedinogo tsifrovogo prostranstva nauchnykh znaniy // Informatsionnyye resursy Rossii. - 2020. - № 5. - pp. 43-46
7. Sobolevskaya I., Sotnikov A. Multimedia Objects Representation in the Digital Knowledge Space //CEUR Workshop Proceedings. – 2021. – pp. 227-237.
8. Irina Sobolevskaya. Some Aspects of 3D-objects Presentation in a Common Digital Space of Scientific Knowledge // CEUR Workshop Proceedings (CEUR-WS.org). - 2021. - Vol. 2990. - pp. 117-124. - DOI: 10.51218/1613-0073-2990-117-124
9. Gothandaraman R., Muthuswamy S. Virtual models in 3D digital reconstruction: detection and analysis of symmetry //Journal of Real-Time Image Processing. – 2021. – T. 18. – №. 6. – pp. 2301-2318.
10. Bacciaglia A., Ceruti A., Liverani A. Surface smoothing for topological optimized 3D models //Structural and Multidisciplinary Optimization. – 2021. – T. 64. – №. 6. – pp. 3453-3472.
11. Peyre J. et al. Detecting unseen visual relations using analogies //Proceedings of the IEEE/CVF International Conference on Computer Vision. – 2019. – pp. 1981-1990.

12. Kalenov N.Ye., Sotnikov A.N. O strukture ontologii Yedinogo tsifrovogo prostranstva nauchnykh znaniy // Nauchnyy servis v seti Internet: trudy XXIV Vserossiyskoy nauchnoy konferentsii. - 2022. - pp. 203-221. DOI: 10.20948/abrau-2022-23
13. Kalenov N.Ye., Sotnikov A.N. Unifitsirovannoye predstavleniye ontologii yedinogo tsifrovogo prostranstva nauchnykh znaniy // Elektronnyye biblioteki, - 2023. - T. 26, - № 1. - pp. 80-103. DOI: 10.26907/1562-5419-2023-26-1-80-103.
14. Pogorelko K.P. Dinamika ispol'zovaniya elektronnoy biblioteki "Nauchnoye naslediye Rossii" // Informatsionnoye obespecheniye nauki: novyye tekhnologii: Sb. nauch. tr., M. - 2017. - pp. 192-200.
15. Konstantin Pogorelko A New Version of the Software for the Information System “Scientific Heritage of Russia” // CEUR Workshop Proceedings (CEUR-WS.org). – 2021. - Vol. 2990. - pp. 110-116. DOI: 10.51218/1613-0073-2990-110-116
16. Kalenov N.Ye., Pogorelko K.P., Sotnikov A.N. O razvitii elektronnoy biblioteki “Nauchnoye naslediye Rossii” kak sostavlyayushchey Yedinogo tsifrovogo prostranstva nauchnykh znaniy // Informatsionnyye protsessy. - 2022. - T. 22, - № 3. - pp. 155-166. DOI: 10.53921/18195822|2022|22|3|155
17. Bader S. et al. The international data spaces information model–an ontology for sovereign exchange of digital content //The Semantic Web–ISWC 2020: 19th International Semantic Web Conference, Athens, Greece, November 2–6, 2020, Proceedings, Part II. – Springer International Publishing, 2020. – pp. 176-192.
18. Serebryakov V. A., Ataeva O. M. Ontology Based Approach to Modeling of the Subject Domain ‘‘Mathematics’’in the Digital Library //Lobachevskii Journal of Mathematics. – 2021. – T. 42. – pp. 1920-1934
19. Gawrysiak P. et al. Retrieval and Management of Scientific Information from Heterogeneous Sources //Intelligent Tools for Building a Scientific Information Platform. – 2012. – pp. 37-48.
20. Coneglian C. S. et al. Ontological semantic agent in the context of big data: A tool applied to information retrieval in scientific research //New Advances in Information Systems and Technologies. – Springer International Publishing, 2016. – pp. 307-316.
21. Studer R. et al. New dimensions in semantic knowledge management //Towards the Internet of Services: The THESEUS Research Program. – 2014. – pp. 37-50.
22. Biswas C. et al. Privacy-aware supervised classification: An informative subspace based multi-objective approach //Pattern Recognition. – 2022. – T. 122. – pp. 108301.
23. Yang M. et al. Learning unseen concepts via hierarchical decomposition and composition //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. – 2020. – pp. 10248-10256.

N.E. Kalenov. DSc. Joint Supercomputer Center of the Russian Academy of Sciences — Branch of Federal State Institution “Scientific Research Institute for System Analysis of the Russian Academy of Sciences”, 119334, Moscow, Leninsky av., 32 a, Russia. e-mail: nkalenov@jscc.ru

I.N. Sobolevskaya. PHd. Joint Supercomputer Center of the Russian Academy of Sciences — Branch of Federal State Institution “Scientific Research Institute for System Analysis of the Russian Academy of Sciences”, 119334, Moscow, Leninsky av., 32 a, Russia. e-mail: ins@jscc.ru

A.N.Sotnikov. DSc. Joint Supercomputer Center of the Russian Academy of Sciences — Branch of Federal State Institution “Scientific Research Institute for System Analysis of the Russian Academy of Sciences”, 119334, Moscow, Leninsky av., 32 a, Russia. e-mail: asotnikov@jscc.ru

S.A. Vlasova. PHd. Joint Supercomputer Center of the Russian Academy of Sciences — Branch of Federal State Institution “Scientific Research Institute for System Analysis of the Russian Academy of Sciences”, 119334, Moscow, Leninsky av., 32 a, Russia. e-mail: svlasova@jscc.ru