Fundamentals of Common Digital Space of Scientific Knowledge Building

Nikolay Kalenov[1[0000-0003-5269-0988], Gennadiy Savin[2[0000-0003-4189-125562],
Alexander Sotnikov[3[0000-0002-0137-1255]

[1–3] 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,
[1] nkalenov@jscc.ru, [2]savin@jscc.ru, [3]asotnikov@jscc.ru

Abstract. The article discusses the general issues of creating a Common Digital Space of Scientific Knowledge (CDSSK) as a modern integrated structure focused on supporting the tasks of information support for science and education, popularizing and storing knowledge reflected in various digital objects. The tasks of the CDSSK are formulated, user groups are determined, the architecture of the space is considered. The CDSSK includes a set of subspaces related to various scientific fields. The unity of space is ensured by unified principles for constructing subspaces and ontological connections between their objects. Each subspace includes digital objects, metadata containing facts associated with objects, and subject ontologies that provide advanced searches and navigation through space. All information is reflected in the CDSSK according to the rules of the "Semantic Web". The content of each subspace includes a core (time-tested reliable scientific results) and a superstructure - new scientific results that have passed preliminary examination article describes

Keywords[1]: scientific information, scientific knowledge space structure, scientific knowledge space architecture, ontologies, information provision.

1 Introduction

The modern information space contains enormous corpus of scientific information displayed in publications and databases, presented in digital form. This corpus is growing

[1] CDSSK–2020: International Conference "Common Digital Space of Scientific Knowledge", November 10–12, 2020, Moscow, Russia

EMAIL: nekalenov@mail.ru (Nikolay Kalenov); savin@jscc.ru（Gennadiy Savin); asotnikov@jscc.ru (Alexander N. Sotnikov)
ORCID: 0000-0003-5269-0988 (Nikolay Kalenov); 0000-0003-4189-125562（Gennadiy Savin); 0000-0002-0137-1255 (Alexander Sotnikov)

at an increasing rate in all fields of science. If we talk about absolute numbers, the following data on the annual number of published articles in journals and conference proceedings reflected in the Russian Science Citation Index (RSCI) [1] in five fundamentally different fields of science -informatics, microbiology, linguistics, social sciences in general, and ecology – can be given as an example. These data for three years of the last decade (2010, 2014 and 2019) are shown in Table 1.

Table 1. Growth dynamics of scientific publications.

Field of science	Type of publications	2010	2014	2019
Computer science	Articles in journals	24968	40293	38744
	Conference proceedings	7501	17591	42113
	Total	**32469**	**57884**	**80857**
Microbiology	Articles in journals	10471	11607	14793
	Conference proceedings	999	2178	4990
	Total	**11470**	**13785**	**19783**
Linguistics	Articles in journals	25900	47182	64929
	Conference proceedings	8047	20670	57973
	Total	**33947**	**67852**	**122902**
Social sciences in general	Articles in journals	12685	22332	17359
	Conference proceedings	3860	12423	43915
	Total	**16545**	**34755**	**51274**
Ecology	Articles in journals	15659	20807	23442
	Conference proceedings	3689	7251	19017
	Total	**19348**	**28058**	**42459**

We can see from the table that over the past decade, the number of annual publications in journals and conference proceedings in various fields of science has increased from 80 per cent (microbiology) to 260 per cent (linguistics). To keep pace with scientific developments, a researcher needs the latest and most accurate information in his field of study. Providing scientists with the information they need at any given time was traditionally the main job of scientific libraries and information centers in Russia. The development of informatics, its technical and algorithmic base, and the replacement of printed publications with digital ones have given rise to the thesis “there is everything on the Internet”, the consequence of which has been the shifting the task of information support for scientific research to the scientists themselves. This approach

may lead the researcher to spend the lion's share of his time and intellectual efforts on becoming familiar with information rather than on research itself. On the other hand, if he does not do so, the lack of information can lead to duplication of research and a pointless waste of time in getting results previously obtained by other researchers.

In this context, it is necessary to create an information environment and develop tools, including those that use advances in artificial intelligence, to provide scientists with the latest information in their fields of study, with minimum time consumption on their part. One of the ways to solve this problem is research in the field of creating the Common Digital Space of Scientific Knowledge (CDSSK) [2, 3], containing reliable and unduplicated scientific factual and documentary information on various branches of science.

2 The objectives of the CDSSK creation

Building a modern system of information support for scientists is only one of the goals of the CDSSK. The CDSSK is focused on solving a set of scientific, informational, educational, general cultural and management tasks, providing:

- information support for scientific research;
- support of educational processes from secondary school (closely related to the popularization of science) to graduate school (directly related to information support);
- the popularization of science (to promote the motivation to do science and receive the appropriate education);
- preservation of scientific knowledge;
- processes of monitoring and management of science.

The content of the CDSSK includes several "layers" of information, each of which is oriented to a specific category of users, among whom:

- researchers who should be provided with multifaceted information, retrospective, and current information on the scientific direction corresponding to their interests;
- schoolchildren and students who should receive reliable, time-tested basic information at various levels; this information should include facts referring to teaching materials, texts of classical textbooks themselves (their expert selection should be based on independent criteria), digital models of phenomena and discoveries;
- specialists – analysts and representatives of management structures, analyzing the state and trends of development of different fields of science;
- the general public, who should be made aware of the most interesting results obtained in a particular field of science, as well as the history of scientific discoveries and their authors;
- specialists and "amateurs" interested in the history of science and its creators.

3 CDSSK structure and content

The tasks to be undertaken in the context of the establishment of the CDSSK determine its architecture. The space should include a set of subspaces (SSs) for different fields of science, built on common principles based on the use of ontological standards used in the Semantic Web [4, 5]. The content of each CDSSK SS includes three components: digital objects that represent the real world; metadata that reflect the different properties of each digital object and the different types of links between them; subject ontologies (thesaurus, supplemented by indices of different classification systems) describing the scientific field.

The CDSSK digital objects are divided into two types of classes - universal and local. To describe the objects of each class, different metadata profiles are used – sets of formalized properties and connections specific to objects of this class. The specific content of metadata profiles is determined by the requirements for searching and visualizing objects of this class in the CDSSK.

Universal classes include objects that are not ontologically related to any particular domain of science. Their metadata profiles are independent of the thematic subspace to which they relate. Such classes of objects as "event", "person", "publication", "organization", "archive document", etc. are universal.

Local classes include objects that are specific to a particular scientific field. Their metadata profiles are therefore also specific. Such classes as "theorem", "equation", "set" are characteristic for mathematics; "chemical element", "reaction" – for chemistry; "body of text", "language" – for linguistics. Individual local classes may belong to several scientific fields. For example, the class "law of nature" is included in the subspace of many natural scientific directions and a number of humanities, while the "legal law" is specific to social sciences.

Identification of local classes of objects belonging to the CDSSK subspace, defining the metadata profiles of the objects of each class and the types of connections between the objects of the same and different classes are major tasks in the design of any thematic subspace of the CDSSK.

Based on the tasks to be solved within the framework of the CDSSK, each subspace must include "basis" – fundamental, time-tested, information related to a given scientific direction, and "superstructure" – new scientific results having passed preliminary examination.

The "basis" is a set of scientific laws, postulates, the main results obtained in each scientific field, with references to the sources in which they are published, and the full texts of these sources. The content of the basis includes three interrelated levels – educational, popular science, and fundamental.

The educational level, aimed at school students, is formed on the basis of information obtained from fundamental textbooks and includes, along with factual information and full texts of textbooks, verified multimedia resources and digitized museum objects. This level of content is mostly static – information is rarely updated. Popular science level, designed for users who are interested in science but are not specialists in the field. This level is quasi-static. It changes when new knowledge relevant to the field becomes available. Formation of this level is based on encyclopedic and reference information,

popular science publications, digital models of scientific phenomena and museum objects.

Fundamental level content is intended for specialists in each field of science. This level includes more in-depth, in comparison with the previous two levels, information on the given scientific field, the core of this content is static and contains multifaceted factual information with references to its sources. If the sources are published materials, their full texts (if possible, the first publications of historical value) should also be included in the content of the basis of this SS.

The content of the subspace superstructure is intended for researchers working in the given field of science and contains new scientific results published in authoritative journals or reflected in the certificates of authorship. After a certain period, according to the results of evaluation by the expert community, individual components of the superstructure can be transferred to the basis or removed from the CDSSK.

The content of the CDSSK should be formed, first of all, on the basis of existing digital scientific resources. Sources for content formation should be authoritative encyclopedias – multidisciplinary (for example, the Big Russian Encyclopedia [6]) and on individual areas of science (for example, [7–9]), factographic information systems on scientific areas (such as INIS on nuclear physics [10], "Reaxys" on chemical reactions [11], etc.); digital collections and catalogs of objects of biological, anthropological, geological, and other studies; authoritative electronic libraries (such as the National Electronic Library [12] and the electronic library "Scientific Heritage of Russia" [13-15]); collections of scientific museums and archives.

The selection of materials for the formation of the content of the CDSSK should be based on a combination of modern methods for information analysis (including the use of artificial intelligence elements) and the results of the examination by the scientific community. Technologically, the CDSSK should include the following inter-related components: a central core, a set of local cores of thematic subspaces and a distributed container.

The central core contains metadata of the objects of universal classes and generalized subject ontology; the local cores contain metadata of the objects of local classes related to each other and to the objects of universal classes. A distributed container contains digital copies of real-world objects (publications, archival documents, museum artifacts, popular science films), they can be stored on servers of their owners - scientific organizations, libraries, museums, archives, etc.

4 Conclusion

The creation of the CDSSK is a task of national scale. The stages of its design should include:

- development of the CDSSK general ontology; its functional and organizational structure;
- analysis of existing possible sources of information and development of selection criteria for objects to be included in the content space;

- development of principles and technology for content development and content updating;
- development of requirements for linguistic, technical and software tools for individual components of the CDSSK and the space as a whole;
- study legal issues on the use and provision of CDSSK elements to users.

National researchers have a serious groundwork for the implementation of the above tasks. As a basis for further development of the work towards the CDSSK creation we can consider the research related to the creation and support of the digital library "Scientific Heritage of Russia" conducted at Joint Supercomputer Center of Russian Academy of Sciences (JSCC RAS) – Branch of Federal State Institution "Scientific Research Institute for System Analysis" of RAS; research in the field of semantic digital libraries conducted at FIC IS [16, 17]; research in the field of development of linguistic tools [18–21]. All-Russian mathematical portal MathNet [22], information system "Socionet" [23], information system "History of geology and mining" [24, 25] can be prototypes of practical implementation of some components of the CDSSK.

The research is carried out in the JSCC RAS within the framework of the state order 0580-2021-0016 and with the support of RFBR (project 20-07-00773).

References

1. Russian Science Citation Index. https://www.elibrary.ru/, last accessed 2021/08/24.
2. Antopol'skiy, A.B., Kalenov, N.E., Serebriakov, V.A., Sotnikov, A.N.: About Common Digital Scientific Space of Knowledge. Vestnik Rossiiskoy Academii Nauk 89 (7). P. 728–735 (2019).
3. Antopol'skiy, A.B., Bosov, A.V., Savin, G.I., Sotnikov, A.N., Tsvetkova, V.A., Kalenov N.E., Serebryakov, V.A., Efremenko, D.V.: Printsipy postroyeniya i struktura yedinogo tsifrovogo prostranstva nauchnykh znaniy (YETSPNZ). Nauchno-Tekhnicheskaya Informatsiya. Ser. 1 (4). S. 9–17 (2020).
4. https://www.w3.org/standards/, last accessed 2021/08/24.
5. Bennett, M., Baclawski, K.:The Role of Ontologies in Linked Data, Big Data and Semantic WEB Application. Applied Ontology 12 (3-4). P. 189–194 (2017).
6. Scientific and Technical Network. https://www.stn-international.com/, last accessed 2021/08/24
7. Portal "Bol'shaya rossiyskaya entsiklopediya". https://bigenc.ru/, last accessed 2021/08/24.
8. Fizicheskaya entsiklopediya. https://dic.academic.ru/dic.nsf/enc_physics/, last accessed 2021/08/24.
9. Khimicheskaya entsiklopediya. https://dic.academic.ru/contents.nsf/enc_chemistry/, last accessed 2021/08/24.
10. Matematicheskaya entsiklopediya. https://dic.academic.ru/contents.nsf/enc_mathematics/, last accessed 2021/08/24.
11. International Nuclear Information System (INIS). https://www.iaea.org/resources/databases/inis, last accessed 2021/08/24.
12. Reaxys. https://www.reaxys.com/, last accessed 2021/08/24.
13. Natsional'naya elektronnaya biblioteka. https://rusneb.ru/, last accessed 2021/08/24
14. Elektronnaya biblioteka "Nauchnoye naslediye Rossii". http://heritage1.jscc.ru/, last accessed 2021/08/24.

15. Kalenov, N.E., Savin, G.I., Serebryakov, V.A., Sotnikov, A.N.: Printsipy postroyeniya i formirovaniya elektronnoy biblioteki “Nauchnoye naslediye Rossii”. Programmnyye Produkty, Sistemy i Algoritmy 4 (100). S. 30-40 (2012).
16. Pogorelko, K.P. Dinamika ispol'zovaniya elektronnoy biblioteki “Nauchnoye naslediye Rossii”. In: Informatsionnoye obespecheniye nauki: novyye tekhnologii: Sbornik nauchnykh trudov. M.: BEN RAN. S. 192–200 (2017).
17. Ataeva, O.M., Serebryakov, V.A.: Ontologiya tsifrovoy semanticheskoy biblioteki LibMeta. Informatika i ee Primeneniya 12 (1). S. 2–10 (2018).
18. Ataeva, O.M., Serebryakov, V.A., Tuchkova, N.P.: Rasshireniye predmetnoy oblasti informatsionnogo zaprosa na osnove ontologii znaniy tsifrovoy biblioteki LibMeta. Nauchniy servis v seti Internet (21). S. 63–75 (2019).
19. Antopol'skiy, A.B., Beloozerov, V.N., Kalenov, N.E., Markarova, T.S.: O razvitii terminologicheskoy bazy dannykh v vide kompleksa otraslevykh informatsionno-poiskovykh tezaurusov. Informatsionnyye Resursy Rossii (5). S. 22–30 (2018).
20. Beloozerov, V.N., Shapkin, A.V., Shchuko, Yu.N.: Set' klassifikatsionnykh sistem VINITI RAN. Programmnyye Produkty, Sistemy i Algoritmy (4). S. 20–23 (2018).
21. Antopol'skiy, A.B.: O sozdanii tsentra lingvisticheskikh resursov RAN. Izvestiya Rossiyskoy Akademii Nauk. Seriya Literatury i Yazyka 78 (4). S. 5–12 (2019).
22. Kalenov, N.E., Senko, A.M.: Interactive system of terminological dictionaries as one of the elements in the ontology of scientific knowledge. Software Journal: Theory and Applications (electronic Journal) (4) (2019). https://doi.org/10.15827/2311-6749.33.423
23. Zhizhchenko, A.B., Izaak, A.D.: The Information System Math-Net.Ru. Current State and Prospects. The Impact Factors of Russian Mathematics Journals, Russian Math. Surveys 64 (4). P. 775–784 (2009). http://dx.doi.org/10.1070/RM2009v064n04ABEH004638
24. Parinov, S., Lyapunov, V., Puzyrev, R., Kogalovsky, M.: Semantically Enrichable Research Information System Socionet. Communications in Computer and Information Science 518. P. 147–157 (2015).
25. Kalenov, N.E., Malakhova, I.G.: Integrirovannyy obshchedostupnyy informatsionnyy resurs “Istoriya geologii i gornogo dela”. Informatsionnyye Resursy Rossii (1). S. 19–23 (2017).