On the Formation of the Space of Scientific Knowledge Subject Ontology

Nikolay Kalenov[1][0000-0001-5269-0988]

[1] Joint Supercomputer Center of RAS – branch of Federal State Institution «Scientific Research Institute for System Analysis of RAS, Leninskiy pr., 32a, 119334, Moscow, Russia
nekalenov@mail.ru

Abstract. Subject ontologies are the basis of every information system that describes scientific knowledge. A particular simplified case of a subject ontology is the indexes of classification systems, such as international UDC, ICI, Russian State Rubricator for Scientific and Technical Information (SRSTI) and Library Bibliografic Classification (LBC). A more developed example of a subject ontology is thematic thesauri containing terms related to a certain field of science or technology, with their hierarchical and horizontal connections. When considering the architecture of the Common Digital Space of Scientific Knowledge (CDSSK), a subject ontology is understood as a generalized structure that includes both thesaurus elements and associated indexes of various classification systems that describe this scientific area. This article presents the results of research related to the construction of subject ontologies based on the previously developed system for supporting terminological dictionaries and proposes a methodology for identifying new key terms for its replenishment. The methodology is based on the use of existing citation databases (CDB), such as WEB of Science and Scopus for English-language publications and the Russian Citation Index for Russian-language publications. The methodology presupposes the division of the scientific area into a number of sections, the selection from the CDB of the core of articles related to each section, and from the articles - the author's key terms, which, in conjunction with the corresponding sections of the classification systems, should form the basis of the subject ontology of this scientific area.

Keywords: Scientific knowledge space, Subject ontology, Citation databases, Keywords, Scientific terms, Thesaurus, Classification systems, Terminological dictionaries.

1 Introduction

One of the important areas of modern informatics, associated with the preservation and dissemination of scientific achievements, is the creation of a Common Digital space of scientific knowledge (CDSSK). This space should reflect the reliable knowledge obtained in various fields of science. The purpose of creating a space is to provide users of various categories with multifaceted information both within individual scientific areas and at the intersection of sciences. In accordance with the concept reflected in [1–3], the CDSSK is a collection of heterogeneous information resources, grouped into thematic subspaces, united by a single ontology. A unified ontology is understood as the principles of their construction common for all subspaces – unified approaches to storing and providing information, forming object classes and relationships between them, metadata and attribute profiles, user interfaces, etc. The ontology and software shell of the CDSSK should provide a developed multifaceted search for heterogeneous information, its convenient visualization and navigation through related resources. The basis for the thematic search for information in each subspace should be its subject ontology – the most complete set of terms reflecting all aspects of the scientific direction, with the links established between them. By definition, the CDSSK should contain a variety of resources, including those retrieved from existing databases and library catalogs. Subject search for information in these databases is based on the classification system (CS) adopted in them. If we talk about Russian polythematic bibliographic resources, then they are based on one of such KS as Rubricator for Scientific and Technical Information (SRSTI), UDC, Library Bibliografic Classification (LBC), International Classification of Inventions (ICI). To ensure accurate and complete import of data from external bibliographic systems, the CDSSK subject ontology should include the indices of these CSs. The subject ontology of the CDSSK subspace is a thematic thesaurus in this scientific area, supplemented by indices of various classification systems that act as descriptors. It is obvious that the problem of the formation of subject ontologies is closely related to the traditional problems of the formation of thematic thesauri. A lot of studies, both foreign and domestic, are devoted to these problems [4–6]. Standard forms of presentation of thesauri in machine form [7], software tools for their formation and embedding into digital libraries [8] have been developed.

Theoretical developments related to the problems of constructing and presenting thesauri in digital form create a certain basis for the formation of subject ontologies of the CDSSK. However, there is no uniform methodology for their practical implementation for various scientific fields. One of the possible typical approaches to solving this problem is presented in this article.

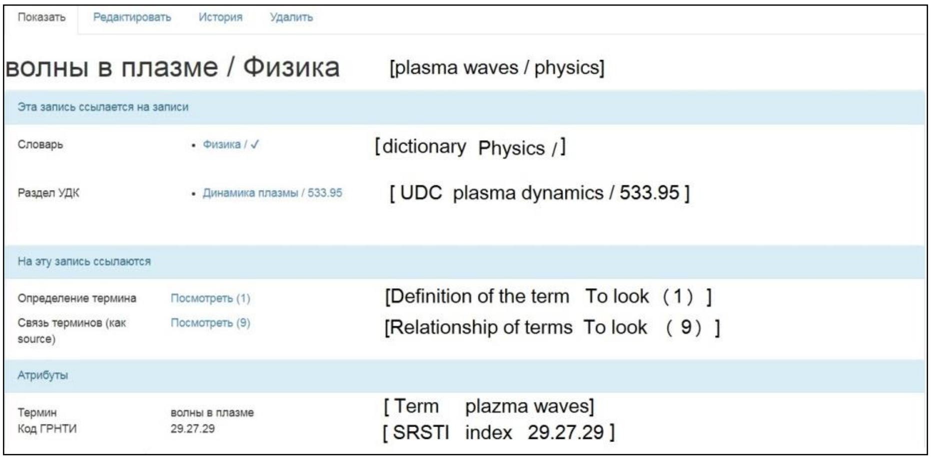
2 Basic System "Terms"

In 2017–2019, with the support of the RFBR, a team of specialists with the participation of the author conducted research in the field of creating a prototype of a subject ontology based on the use of existing information resources [9–12].

The result of these studies was the creation of a system of terminological dictionaries "Term" [13], which includes terms and their definitions corresponding to the concepts reflected in the SRSTI [14]. The informational basis for building the system was the terminological dictionaries developed at Allrussian Institute for Scientific & Technical Information (VINITI).

The system in its original form included more than 12,000 terms related to 69 thematic scientific areas, and definitions of terms with active links to their sources on the

Internet. The system provided the ability to enter and edit data, search, view and navigate through its elements. In Fig. 1 provides information about the term "plasma waves" related to the physics dictionary. Here you can see the name of the dictionary to which the term belongs, UDC indices (in this case 533.95), SRSTI code (29.27.29), links to the definition of the term.



волны в плазме / Физика [plasma waves / physics]

Эта запись ссылается на записи

Словарь • Физика / ✓ [dictionary Physics /]

Раздел УДК • Динамика плазмы / 533.95 [UDC plasma dynamics / 533.95]

На эту запись ссылаются

Определение термина Посмотреть (1) [Definition of the term To look (1)]

Связь терминов (как source) Посмотреть (9) [Relationship of terms To look (9)]

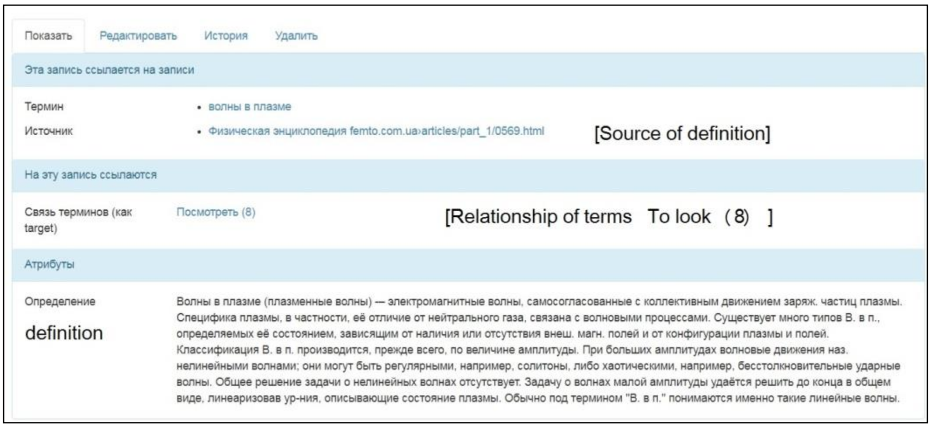
Атрибуты

Термин волны в плазме [Term plazma waves]

Код ГРНТИ 29.27.29 [SRSTI index 29.27.29]

Fig. 1. Term metadata example

Clicking on the link "View" in the "Definition of term" line opens a window (Fig. 2), which contains the definition of the term and an active link to its source.



Эта запись ссылается на записи

Термин • волны в плазме

Источник • Физическая энциклопедия femto.com.ua›articles/part_1/0569.html [Source of definition]

На эту запись ссылаются

Связь терминов (как target) Посмотреть (8) [Relationship of terms To look (8)]

Атрибуты

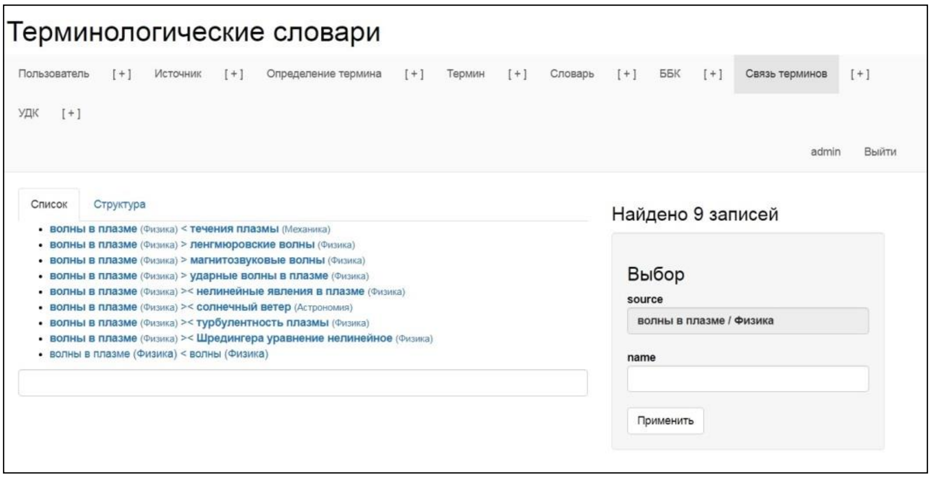
Определение definition

Волны в плазме (плазменные волны) — электромагнитные волны, самосогласованные с коллективным движением заряж. частиц плазмы. Специфика плазмы, в частности, её отличие от нейтрального газа, связана с волновыми процессами. Существует много типов В. в п., определяемых её состоянием, зависящим от наличия или отсутствия внеш. магн. полей и от конфигурации плазмы и полей. Классификация В. в п. производится, прежде всего, по величине амплитуды. При больших амплитудах волновые движения наз. нелинейными волнами; они могут быть регулярными, например, солитоны, либо хаотическими, например, бесстолкновительные ударные волны. Общее решение задачи о нелинейных волнах отсутствует. Задачу о волнах малой амплитуды удаётся решить до конца в общем виде, линеаризовав ур-ния, описывающие состояние плазмы. Обычно под термином "В. в п." понимаются именно такие линейные волны.

Fig. 2. Term definition window

When the development and filling of the first version of the system were completed, the idea arose of forming the relationships of terms by identifying the presence of each term in the definition of other terms, both within "one's" dictionary and in external dictionaries. A corresponding software algorithm was developed and implemented, as a result of which more than 300,000 pairs of terms related to each other in the indicated sense were generated in the system, and the ability to view and edit these relationships was provided. The resulting relationships related to several subject dictionaries were edited by experts in these subject areas – each relationship was assigned one of five types of meanings: identical, close, contains another, contained at another, intersect. In the system, along with verbal designations of the type of relationships, for clarity, symbolic ones are used, respectively – "=", "~", ">", "<", "> <".

The metadata pages for terms contain links to the relationships of that term to others. On the page (Fig. 1) there is a link to relationships with those terms in the definition of which this concept is included. The page containing the definition of a term (Fig. 2) contains a link to relationships with terms that are included in the definition of this concept. In the example shown in Fig. 1 there are 9 such links, in Fig. 2 – 8. Clicking on the link "To look 9" opens a window with the specification of links (Fig. 3). The system offers not only 7 relationships of terms inside the dictionary "Physics", but also connections of the concept “waves in plasma” from the “Physics” dictionary with the term “plasma flow” from the “Mechanics” dictionary (first line) and “Solar wind” from the “Astronomy" (sixth line).


Терминологические словари
Пользователь [+] Источник [+] Определение термина [+] Термин [+] Словарь [+] ББК [+] Связь терминов [+]
УДК [+]
admin Выйти
Список Структура

- волны в плазме (Физика) < течения плазмы (Механика)
- волны в плазме (Физика) > ленгмюровские волны (Физика)
- волны в плазме (Физика) > магнитозвуковые волны (Физика)
- волны в плазме (Физика) > ударные волны в плазме (Физика)
- волны в плазме (Физика) >< нелинейные явления в плазме (Физика)
- волны в плазме (Физика) >< солнечный ветер (Астрономия)
- волны в плазме (Физика) >< турбулентность плазмы (Физика)
- волны в плазме (Физика) >< Шредингера уравнение нелинейное (Физика)
- волны в плазме (Физика) < волны (Физика)

Найдено 9 записей

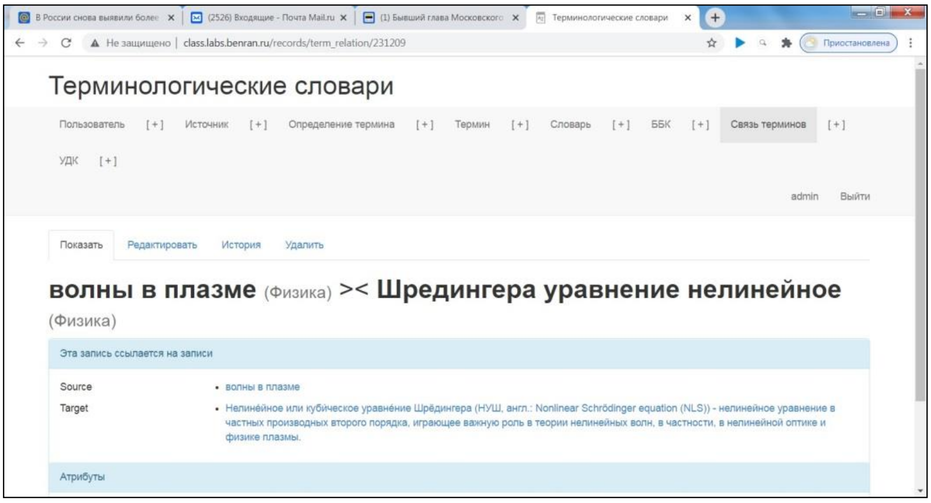
Выбор

source: волны в плазме / Физика

name:

Fig. 3. An example of term relationships (words of this term are included in the definition of others).

Each link is an active link, when you click on it, a window opens with the definition of the term associated with the considered one (Fig. 4). An authorized user with the appropriate rights can edit the link type or delete it.


В России снова выявили более
(2526) Входящие - Почта Mail.ru
(1) Бывший глава Московского
Терминологические словари
Не защищено | class.labs.benran.ru/records/term_relation/231209
Приостановлена
Пользователь [+] Источник [+] Определение термина [+] Термин [+] Словарь [+] ББК [+] Связь терминов [+]
УДК [+]
admin Выйти
Показать Редактировать История Удалить

Терминологические словари

волны в плазме (Физика) >< **Шредингера уравнение нелинейное** (Физика)

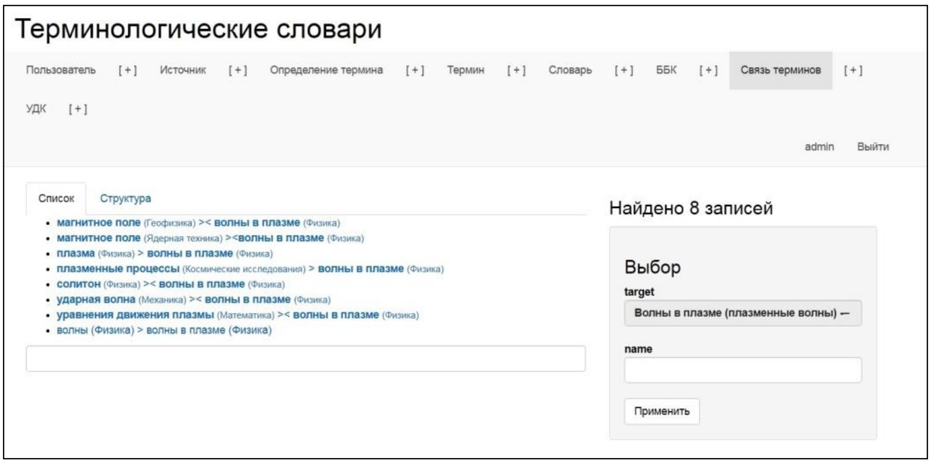
Эта запись ссылается на записи

Source	• волны в плазме
Target	• Нелинейное или кубическое уравнение Шрёдингера (НУШ, англ.: Nonlinear Schrödinger equation (NLS)) - нелинейное уравнение в частных производных второго порядка, играющее важную роль в теории нелинейных волн, в частности, в нелинейной оптике и физике плазмы.

Атрибуты

Fig. 4. Relationship details window

A similar window opens when you click on the "To look 8" link in Fig. 2. Here we see the connections of the term "waves in plasma" with terms from the dictionaries "Geophysics", "Nuclear Engineering", "Space Research" and "Mathematics" (Fig. 5).


Пользователь [+] Источник [+] Определение термина [+] Термин [+] Словарь [+] ББК [+] Связь терминов [+]
УДК [+]
admin Выйти
Список Структура
Применить

Терминологические словари

- **магнитное поле** (Геофизика) >< **волны в плазме** (Физика)
- **магнитное поле** (Ядерная техника) ><**волны в плазме** (Физика)
- **плазма** (Физика) > **волны в плазме** (Физика)
- **плазменные процессы** (Космические исследования) > **волны в плазме** (Физика)
- **солитон** (Физика) >< **волны в плазме** (Физика)
- **ударная волна** (Механика) >< **волны в плазме** (Физика)
- **уравнения движения плазмы** (Математика) >< **волны в плазме** (Физика)
- волны (Физика) > волны в плазме (Физика)

Найдено 8 записей

Выбор

target: Волны в плазме (плазменные волны) —

name:

Fig. 5. Example of term relationships (terms included in the definition of a given one).

3 Methodology for filling subject ontology

For the development of the "Term" system to a full-fledged version of the subject ontology, it is necessary to supplement it with key terms (CT) that describe in sufficient detail the selected scientific directions. It is proposed to solve this problem using the keywords indicated by the authors in their scientific articles. To automate this process and select high-quality publications, citation databases such as WEB of Science (WoS) and Scopus (for English-language publications) and the Russian Citation Index (RSCI) for Russian-speaking ones can be used.

The proposed technique includes the following processes.

1. Division of the field of science into separate sections. The degree of detail of such a division is determined by scientists together with information workers on the basis of an analysis of existing classification systems (CS), such as SRSTI, UDC, ICI, etc., well developed for this scientific area.

2. Establishing relationships between the selected sections and the indexes of the selected CS related to this field of science. In relation to each section in each CS, indices are selected that are more or less associated with this section. The pairs "section – index" are formed and one of the 5 type of relationships between terms in each pair are established (see above).

3. For each of the selected sections, queries are formulated to the citation database (WoS, Scopus, RSCI), in accordance with which publications are selected published during a certain interval of years, depending on the scientific direction. The metadata of each article retrieved upon request has attributes containing a list of key terms and links to publications citing the article.

The data obtained can serve as the basis for filling the subject ontology. For this, key terms for each scientific section should be selected from the received articles, and duplicate terms should be excluded from the resulting list. Then, experts should exclude “noise terms” that are not relevant to the given field of science. Technically, the extracting of key terms from WoS and Scopus is not difficult – both of these systems can process queries automatically and provide the ability to obtain information in various structured formats that make it easy to highlight the author's key terms from DB records. The situation with the RSCI is somewhat more complicated, in which neither automatic processing of requests nor the issuance of information in any structured format is provided. The system provides data in the form of text records. To extract the necessary information, it will be needed to develop a program that processes HTML pages containing found publications.

As a result of processing the information received for each selected section of a given scientific area, an array of Russian-language (RSCI) and English-language terms (WoS, Scopus) is formed, indicating the frequency of their occurrence during any given time interval.

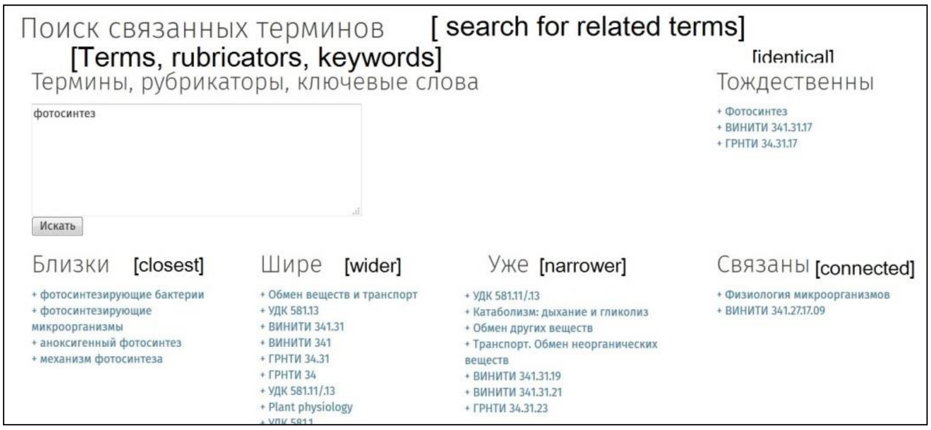
As a first approximation, it can be assumed that all the selected key terms are included in the corresponding sections of this scientific direction, which, in turn, are associated with the previously established type of connection with the indices of various COPs. Obviously, the resulting sets of key terms require editing by specialists in this scientific field, but this work is much easier than searching for key terms.

4 Model Implementation

The proposed technique was tested in 2019 on the example of modeling a subject ontology in microbiology [15]. In this scientific direction, 42 sections were allocated. For each section, a relationship was established of one of the 5 above types with SRSTI and UDC. For each of them, based on the processing of articles records received by queries to the WoS database, key terms were programmatically selected.

A total of 5865 articles were processed, of which 22715 different English key terms (KT) were identified. After semantic processing (screening out of KTs not related to microbiology), the total number of unique KTs was 7346. These terms were translated into Russian and, together with the microbiology data downloaded from the system "Terms", were loaded into a separate database, which is a simplified model of the subject ontology. Data exchange between systems was carried out on the basis of the thesaurus description format proposed within the framework of the Semantic WEB concept (SKOS recommendations) [16].

In Fig. 6 shows the form of issuing terms associated with the concept of "photosynthesis". The elements of the subject ontology located under the heading "are identical" show that "photosynthesis" is the basic term of one of the terminological dictionaries and the heading of the sections of the VINITI and SRSTI headings.



Поиск связанных терминов [search for related terms]

[Terms, rubricators, keywords]

Термины, рубрикаторы, ключевые слова

фотосинтез

Искать

[identical]

Тождественны

+ Фотосинтез
+ ВИНИТИ 341.31.17
+ ГРНТИ 34.31.17

Близки [closest]

+ фотосинтезирующие бактерии
+ фотосинтезирующие микроорганизмы
+ аноксигенный фотосинтез
+ механизм фотосинтеза

Шире [wider]

+ Обмен веществ и транспорт
+ УДК 581.13
+ ВИНИТИ 341.31
+ ВИНИТИ 341
+ ГРНТИ 34.31
+ ГРНТИ 34
+ УДК 581.11/.13
+ Plant physiology
+ УДК 581.1

Уже [narrower]

+ УДК 581.11/.13
+ Катаболизм: дыхание и гликолиз
+ Обмен других веществ
+ Транспорт. Обмен неорганических веществ
+ ВИНИТИ 341.31.19
+ ВИНИТИ 341.31.21
+ ГРНТИ 34.31.23

Связаны [connected]

+ Физиология микроорганизмов
+ ВИНИТИ 341.27.17.09

Fig. 6. An example of visualization of terms related to "photosynthesis"

The lists “Wider” and “Narrower” contain corresponding indices of three headings (GRNTI, UDC and VINITI) and the names of sections allocated by microbiologists. The “Closest” list includes key terms selected from the WoS, whose names include the word “photosynthesis”; in the list "Connected" – sections of microbiology from the "Terms" system, the definitions of which include the term "photosynthesis" (in this example, one of the VINITI rubricator indexes.

5 Conclusion

The proposed methodology has shown its efficiency. Researches on its development will be continued at the MSC RAS within the framework of state assignment 0580-2021-0016 and at VINITI with the support of the Russian Foundation for Basic Research (project No. 20-07-00103). The nearest prospect is to expand the model of subject ontology using the example of microbiology – to work out the technology for replenishing the set of key terms from Russian-language databases (primarily from the RSCI), to enter into the system of English-language terms with the establishment of synonymy and other relations with Russian-language terms extracted from the RSCI.

References

1. Antopolskij, A.B., Kalenov, N.E., Serebryakov, V.A., Sotnikov, A.N.: O edinom cifrovom prostranstve nauchnyh znanij. Vestnik Rossijskoj akademii nauk, 89(7). 728–735 (2019).
2. Antopolskij, A.B. and others: Principy postroeniya i struktura edinogo cifrovogo prostranstva nauchnyh znanij (ECPNZ). Nauchno-tehnicheskaya informaciya. Ser. 1, (4). 9–17 (2020).
3. Kalenov, N., Sobolevskaya, I., Sotnikov, A.: Mathematical modeling of the processes of interdisciplinary collections formation in the digital libraries environment. CEUR Workshop Proceedings. 2543, 391–398 (2020).
4. Mercedes, M. Martínez-González, María Luisa Alvite, Díez: The support of constructs in thesaurus tools from a Semantic Web perspective: Framework to assess standard conformance. Comput. Stand. Interfaces (65). 79–91 (2019).
5. Roche, C., Costa, R., Carvalho, S., Almeida, B.: Knowledge-based terminological e-dictionaries: The EndoTerm and al-Andalus Pottery projects Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication, (2), 259–290 (2019).
6. Beloozerov, V.N. Gurevich, I.B., Trusova, Yu.O.: Tezaurus po analizu izobrazhenij v seti terminologicheskih slovarej. Perspektivnye napravleniya issledovanij i kriticheskie tehnologii v klassifikacionnyh sistemah: materialy konf. Moskva, 35–36 (2017).
7. UNESCO Thesaurus. https://skos.um.es/unescothes/, last accessed 28.04.2020.
8. Ataeva, O.M., Serebryakov, V.A.: Personalnaya otkrytaya semanticheskaya cifrovaya biblioteka LibMeta. Konstruirovanie kontenta. Integraciya s istochnikami LOD. Informatika i eyo primeneniya, 11(2), 85–100 (2017).
9. Antopolskij, A.B., Beloozerov, V.N., Markarova, T.S.: O razrabotke ontologii na osnove klassifikatorov nauchnoj informacii i terminologicheskih slovarej. Informacionnye resursy Rossii, (5 (159)), 2–7 (2017).
10. Antopolskiy, A.B. and others: The Development of a Semantic Network of Keywords Based on Definitive Relationships. Scientific and Technical Information Processing, 44(4), 261–265 (2017).
11. Antopolskij, A.B., Beloozerov, V.N., Kalenov, N.E., Markarova, T.S.: O razvitii terminologicheskoj bazy dannyh v vide kompleksa otraslevyh informacionno-poiskovyh tezaurusov. Informacionnye resursy Rossii, (5 (165)), 22–30 (2018).
12. Beloozerov, V.N., Shaburova, N.N.: O razrabotke klassifikacionno-tezaurusnoj ontologii dlya predmetnoj oblasti fiziki i radioelektroniki. Informacionnoe obespechenie nauki: novye tehnologii: sb. nauch. tr. Ekaterinburg, 75–86 (2018).
13. 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). http://swsys-web.ru/en/interactive-system-of-terminological-dictionaries.html. last accessed 28.04.2020.
14. Gosudarstvennyj rubrikator nauchno-tehnicheskoj informacii. http://grnti.ru last accessed 28.04.2020.

15. Tsvetkova, V.A., Harybina, T.N., Mokhnacheva, Yu.V., Beskaravajnaya, E.V., Mitroshina, I.Yu.: Osobennosti sovmesheniya klassifikacionnyh sistem i formirovaniya massiva klyuchevyh slov dlya opredeleniya prostranstva znanij po mikrobiologii. Nauchnye i tehnicheskie biblioteki, (11), 25–43 (2019).
16. Zeng, M.L., Mayr, P.: Knowledge Organization Systems (KOS) in the Semantic Web: a multi-dimensional review. Int. J. Digit. Libr., 20, 209–230 (2019). https://doi.org/10.1007/s00799-018-0241-2