Информационные технологии

Classes of objects and relations in the Common Digital Space of Scientific Knowledge*

N.E. Kalenov, I.N. Sobolevskaya, A.N. Sotnikov

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), Moscow, Russia

Abstract. All over the world there are many both global and local information systems focused on solving various problems. As an integrator that allows you to solve complex information problems at the intersection of sciences and application areas of existing information systems. to the maximum extent using the information resources accumulated in them, the Common Digital Space of Scientific Knowledge (CDSSK) can be considered. The article provides the structure of the CDSSK, the requirements for its functionality and the structure of the software shell, corresponding to the principles of the Semantic WEB. All objects reflected in the CDSSK are divided into two classes − universal and local. Relationships between objects are also divided into two groups − universal and specific. The paper proposes a list of universal classes of objects, defines universal types of relations between them, gives examples of specific relations and approaches to identifying local classes and subclasses of objects in a particular field of science.

Keywords: common Digital Space of Scientific Knowledge, Semantic WEB, codification, objects classes, object reference, object relations, entity reference, relationships between objects, ontology.

DOI: 10.14357/20790279230101

Introduction

The Common Digital Space for Scientific Knowledge (CDSSK) is being formed with the aim of supporting and developing services in the field of science and education in the modern digital environment. [1 -4] The CDSSK includes heterogeneous information objects tested by the world scientific community. All over the world and in Russia, in particular, there are many both global and local information systems focused on solving various problems. In this regard, CDSSK should be considered as an integrator that allows you to solve complex information problems at the intersection of sciences and areas of application of existing information systems. maximum use of information resources accumulated in them. In particular, if we talk about Russia, then there are and are developing a number of state information systems in the digital environment, for example, the Russian Encyclopedia [5], the National Electronic Library [6], the Common State Register of Legal Entities [7], State catalog of geographical names [8], Russian Science Citation Index [9] (Fig. 1).
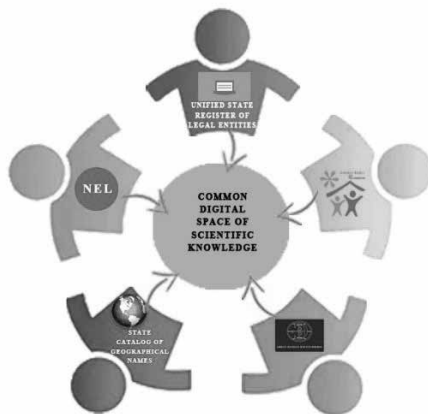
Fig. 1. CDSSK is an integrator for scientific purposes of the state information systems

1. The common digital space of scientific knowledge object classes

Formation of CDSSK as a semantic WEB-space includes the following tasks:

– selection and structuring of scientific objects presented in existing information systems and containing reliable and comprehensive information about scientific achievements in various fields of knowledge;
– metadata profiles formation of objects presented in information systems;
– link start-up and registration of various kinds connections between dissimilar objects;
– formation of RDF triplets within the chosen field of knowledge.

Using OWL and RDF objects representations, their properties and relationships, and SPARQL-based data manipulation tools, you can build an information system containing multifaceted scientific information, backed by citations from reliable, time-tested, information-based sources. systems that are constantly being updated [10-12].

2. The common digital space of scientific knowledge general ontology construction

When constructing CDSSK common ontology, it is necessary to implement the following steps:
1. Allocation of universal classes of objects.

Currently, these include:
– Persons;
– Groups of persons (people united by a certain criterion, for example: "high school students", "students studying in a given specialty", "geologists", "residents of a given country or city", etc.);
– Publications. This object contains subclasses such as: monographs, collections, serials, etc .;

– Qualification works (dissertations and abstracts, copyright certificates);
– The documents. This class contains physical units such as a specific book, handwritten materials, archival documents;
– Museum items. In particular, rare editions should be treated as museum pieces with appropriate relations;
– Events;
– Location (geographic characteristics);
– Time characteristics;
– Organizations;
– Scientific directions;
– Thesauri (subject ontologies);
– General laws of nature from all scientific fields (the law of universal gravitation, the three laws of Newtonian mechanics, the laws of Lomonosov, the laws of thermodynamics, Zipf's law, etc.).
2. Development of metadata profiles of objects of each class (subclass), including:
a. Formation of a list of metadata elements;
b. Define the characteristics and acceptable values of metadata, including:
   – type of data – text, number (range of numbers), date (dates range), link to another object, e-mail address, URL of an external object;
   – mandatory or optional;
   – unique or repetitive;
   – presentation format;
   – selection from permissible values linear table (single or multiple);
   – choice from a hierarchical structure (single or multiple);
   – free value (in accordance with the established view) with possible formal control within the view (text – according to dictionaries, number – for the validity of characters, date – for the established format, link – for checking the existing id, URL – for structure and accessibility).
3. Relations type establishing between objects. I.e., normalized tables formation of relations values between objects inside and outside each class;
4. Development of the data warehouse structure and the relations organization between them;
5. Development of a customizable administrator interface that implements the attributes list formation, tuning tables, data types selection, control type;
6. Development of an operator interface for entering metadata objects;
7. Development of the system internal organization (the formation of the name space of objects and relations (URN), the space of identifiers (URI), the formation of RDF triplets).

When defining metadata profiles for universal classes objects, it makes sense to focus on semantic re-

lations with international and domestic organizations (ResearchGate, ORCID, RSCI – for persons, RAR and USRLE for organizations, RSCI and WEB of Science for publications, SCGN for geographic objects) [13, 14].

3. Universal relations types

1. Equivalence. For persons it means different spellings of surnames and first names;
 – for publications – translated versions of one publication, reprints of books (stereotyped);
 – for organizations – different names of the same organization. For example, Moscow State University – Lomonosov Moscow State University – Moscow University, etc.;
 – for temporal characteristics;
 – for geographical names. For example, RF Russian Federation – Russia.

 Each object has required attributes:
 – a unique identifier;
 – name and relations of a given type with other objects.

 Synonyms in subject ontologies. It is possible to select one of the equivalent objects as the base (descriptor), while the rest of the equivalent objects have enough three attributes – id, the name and the relationship of the type "equivalent" with the base.

2. "To be part of" (subordination of terms in subject ontologies). For organizations it means subordination of subdivisions;
 – for publications: article – journal, collection;
 – for geographical objects: country – continent, city – country, street – city, sea – ocean;
 – for museum items: object – collection;
 – for archives: document – inventory.

3. Contains (the prototype of "To be part of").

4. Intersects (subject ontologies are intersection of classification indices, country and natural zones, desert on the territory of several countries; rivers and countries; international organizations and countries, etc.).

4. Specific relations

Specific relations exist in both generic and local classes [15].

Examples of specific connections.

"Publication" – "Person":
 – author;
 – editor;
 – compiler;
 – interpreter;
 – painter;
 – sponsor;

 – contains information about a person (in the library terminology "about him");
 – reviewer;
 – other roles (technical editor, proofreader, etc.)

"Publication" – "Organization":
 – author (collective author – in library terms);
 – publishing house;
 – contains information about the organization;
 – sponsor.

"Qualification work" – "Person":
 – author;
 – scientific adviser;
 – opponent / reviewer.

"Qualification work" – "Organization":
 – place of work performance;
 – leading organization

"Document" – "Person";
 – author;
 – owner;
 – contains information about the person;
 – the document contains notes of this person.

"Document" – "Organization":
 – author;
 – owner + location specification (for archival documents – the number of the document, case and inventory, for the library the storage code, for the museum – the inventory number).
 – mentioned in the document
 – sponsor.

"Museum Item" – "Person":
 – author – manufacturer;
 – collection author (for natural science collections)
 – source of income (donor / seller)
 – restorer
 – is associated with this person (photography, film-video, audio recording, etc.).

"Museum item" – "organization":
 – author – manufacturer;
 – owner (+ clarification of location – inventory number)
 – source of income (donor / seller)
 – is associated with this organization (photography, film-video, audio recording, etc.).

5. The common digital space of scientific knowledge subspace structure

The factographic basis of each thematic CDSSK subspace (its subject ontology) is structured encyclopedic concepts related to each other and to objects of universal classes. The structure of subject ontology can be based on the sections of existing heading lists of scientific information. For example: UDC (for general scientific) [16], INIS (for nuclear physics) [17], etc.

Below is an example of structuring the subject ontology of the thematic subspace "Astronomy" on the State Rubricator of Scientific and Technical Information [18] basis.

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, subsections of the next level or individual objects are allocated. Each section (subsection) of a subject ontology is an object of the CDSSK. For each object, universal and specific connections are established with other objects of this subspace, other subspaces and with objects of universal classes. For astronomical objects, these can be connections with persons of the form "discovered", "described", "calculated"; with publications -relations of the form "first published", "textbook for school", "the most complete monograph", etc .; with objects from the subspace "Mathematics" – connections of the type "described by equations", etc. Objects of the subclass "Astronomical observatories" included in the last section of the second level of the subject ontology of the software program "Astronomy" are connected with objects of the "Location" class by the obligatory link "located in", etc.

Location (geographic objects) as a universal class contains general information about an object, not detailed from the point of view of geography, but allowing to determine the location with varying accuracy (from the mainland to the house number and coordinates with an accuracy of seconds). The purpose of distinguishing this universal class is to process generalized queries such as "archaeological excavations in Peru," or "her-baria collected in Altai," or "astronomical observations carried out in Chile," and so on. despite the fact that the description of the object of archaeological finds may indicate "Machu Picchu" or "Easter Island", when describing the herbarium, the surroundings of Biysk were indicated, and in astronomical observations, the Atacama Desert was indicated.

Objects of the universal class "location" are associated with elements of the thematic subspace "Geography", which contains comprehensive descriptions of geographic objects. The location class includes the subclasses Land and Water, which in turn include the following 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

When software development for a particular scientific direction, it is necessary to move along the path of identifying classes and subclasses of objects, forming objects metadata profiles of each subclass, establishing relations between objects of this class, within this software and with objects of universal classes. The result of the design should be a set of RDF triplets, which will allow implementing mechanisms for finding answers to complex queries based on the SPARQL language.

References

1. *Antopolskij A.B., Kalenov N.E., Serebryakov V.A., Sotnikov A.N.* O edinom cifrovom prostranstve

nauchnyh znanij // Vestnik Rossijskoj akademii nauk, 2019. – T. 89, – № 7. – S. 728-735. DOI 10.31857/S0869-5873897728-735

2. *Savin G.I.* Edinoe cifrovoe prostranstvo nauchnyh znanij: celi i zadachi // Informacionnye resursy Rossii, 2020. – № 5. – S. 3-5. DOI: 10.51218/0204-3653-2020-5-3-5

3. *Nikolay Kalenov, Gennadiy Savin, Alexander Sotnikov.* Fundamentals of Common Digital Space of Scientific Knowledge Building // CEUR Workshop Proceedings (CEUR-WS.org) , 2021. – Vol. 2990. – P. 93-99. DOI: 10.51218/1613-0073-2990-93-99

4. *Olga Ataeva, Nikolay Kalenov, Vladimir Serebryakov, Alexander Sotnikov.* Informational Infrastructure of the Common Digital Space of Scientific Knowledge // CEUR Workshop Proceedings (CEUR-WS.org) , 2021. – Vol. 2990. – P. 1-10. DOI: 10.51218/1613-0073-2990-1-10

5. https://bigenc.ru/ (the last access 12.2021)

6. https://rusneb.ru/ (the last access 12.2021)

7. https://fedresurs.ru/?attempt=1 (the last access 12.2021)

8. https://cgkipd.ru/science/names/reestry-gkgn.php (the last access 12.2021)

9. https://elibrary.ru/project_risc.asp? (the last access 12.2021)

10. *Millar D., Braines D., D'Arcy L., Barclay I., Summers-Stay D., Cripps P.* Embedding Dynamic Knowledge Graphs based on Observational Ontologies in Semantic Vector Spaces // Artificial intelligence and machine learning for multi-domain operations applications III. Vol. 11746., № 117461O. (2021).

11. *Wang Q., Ji YD., Hao YS., Cao J.* GRL: Knowledge graph completion with GAN-based reinforcement learning // Knowledge-based systems. Vol. 209., № статьи 106421. (2020).

12. *Hansen C., Hotz I., Ynnerman A.* Visualization in Public Spaces // Ieee computer graphics and applications. 40 (2). pp. 16-17. (2020).

13. *Piplai A., Ranade P., Kotal A., Mittal S., Narayanan SN., Joshi A.* Using Knowledge Graphs and Reinforcement Learning for Malware Analysis // 2020 IEEE international conference on big data (big data). pp. 2626-2633. (2020).

14. *Dessi D., Osborne F., Recupero DR., Buscaldi D., Motta E.* Generating knowledge graphs by employing Natural Language Processing and Machine Learning techniques within the scholarly domain // Future generation computer systems-the international journal of escience. Vol. 116. pp. 253-264. (2021).

15. *Nikolay Kalenov, Irina Sobolevskaya, Alexander Sotnikov.* Hierarchical Representation of Information Objects in a Digital Library Environment // Communications in Computer and Information Science. Vol. 1093. pp. 93-104 (2019).

16. UDC: https://udcc.org/index.php/site/page?view=factsheet (the last access 12.2021)

17. INIS: https://www.iaea.org/sites/default/files/19/09/en-2019-09.pdf (the last access 12.2021)

18. SRSTI: https://grnti.ru (the last access 12.2021)

Kelenov N.E. 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

Sobolevskaya I.N. 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 (correspondent author)

Sotnikov A.N. 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