An approach to development of ontological knowledge base in the field of scientific and research activity in Russia

M Sh Murtazina and T V Avdeenko

Novosibirsk State Technical University, 20, Karla Marksa Ave., Novosibirsk, 630073, Russia

E-mail: murtazina@corp.nstu.ru

Abstract. The state of art and the progress in application of semantic technologies in the field of scientific and research activity have been analyzed. Even elementary empirical comparison has shown that the semantic search engines are superior in all respects to conventional search technologies. However, semantic information technologies are insufficiently used in the field of scientific and research activity in Russia. In present paper an approach to construction of ontological model of knowledge base is proposed. The ontological model is based on the upper-level ontology and the RDF mechanism for linking several domain ontologies. The ontological model is implemented in the Protégé environment.

1. Introduction

Over the past 20 years, the Russian sector of scientific knowledge creation is assessed as insufficiently effective in the documents offering the strategy and the direction of the development of the national scientific and technological complex. One of the reasons for this could be the inability of the national research teams to implement the full cycle of fundamental and applied research terminating in the release of high-tech products, which in turn may be due to the insufficient level of structuring and completeness of the data provided by means of information resources in the Internet, containing scientific, technical and technological knowledge, as well as the limited search mechanisms in the field of scientific solutions, technologies and industries.

Indeed, in the modern era of big data, the data themselves are no longer of great value. The thing that is really valuable is knowledge that could be extracted from the data. In this regard, the ever-increasing volume of scientific data in the information resources of the Internet requires the proper organization to find information that is actually relevant to the query of the user. Better organization of data and information could be achieved through information technologies of knowledge management based on the models of data representation in the special semantic form.

The purpose of semantic models of data representations is to give possibility of recording information in the form that can be processed by a computer taking into consideration the meaning of the information. An integral component of semantic technologies is the ontology representing a formalized description of basic concepts of the application domain and relationships between them.

The demand for semantic technologies is growing every year. It is semantic technologies that have become the most promising IT trend in 2013 by Gartner analysts [4]. In the past few years, a number of major search engines started to implement semantic search technologies. In particular, in 2012 the Knowledge Graph semantic tool has been added to the Google search engine [7]. In 2015 the project Semantic Scholar was launched which is a search engine for scientific publications based on artificial intelligence methods [9]. The algorithms for sorting the results of search engines take into account the
scientific significance of the work. The significance of the paper is determined by its citation index, but not only by simple counting the number of citations, but determining the context in which the paper is referenced (literature review, description of the scientific research methodology, etc.). The locations used while mentioning the work are also taken into account. To date the Semantic Scholar system performs the analysis for English language publications.

The development of modern methods of artificial intelligence and data analysis has led to the actively applying worldwide intelligent (semantic) search methods and creating knowledge bases and ontologies, which could significantly improve the quality of management in this area due to the greater awareness of the decision-maker. However, these new intelligent methods are used still not enough for the organization of knowledge in the scientific and technical solutions and technologies. This is especially evident concerning non-English sources. In this regard, it is insufficient level of structuring and completeness of knowledge, as well as the limited mechanisms of semantic search in the field of scientific and technical solutions, can be one of the reasons for the lack of effectiveness of the Russian sector of the scientific knowledge generation. Thus, the task of creating model of information space in the field of scientific data that will make it possible to efficient use of the data collected from open Russian-language online sources is urgent. This requires the creation of ontological knowledge base containing information from multiple sources in the field of scientific and technological solutions presented in various forms that can be used in decision-making both at the strategic and the tactical and operational levels of management.

In the present paper, a model of the ontological knowledge base for intellectual support of scientific and research activities has been proposed. The paper is organized as follows. Section 2 considers the problem of semantic search for the scientific data and assesses the current state of its solution concerning the Russian-speaking sources. Section 3 provides an overview of existing ontologies to systematize scientific publications and the scientific publishing process. Section 4 describes an approach to organization of the ontological knowledge base in the field of scientific research. Section 5 describes the implementation of ontology in Protégé environment. In Section 6 the conclusions about the prospects of using semantic technologies for informatization of the scientific and research activities are given.

2. Semantic search of scientific data

To solve the problem of knowledge retrieval in the scientific domain, it is advisable to distinguish between such concepts as "data for scientific research" and "scientific data". The data for scientific research are the source information used for the research and expressed in numerical or any other form. At the same time scientific data is that representing the results of scientific research.

Since 1989 when Tim Berners-Lee proposed the concept of the World Wide Web, the Internet accumulated great volume of data. As a result a special organization of the data search relevant for the user’s query and elicitation of knowledge from the data became very urgent tasks. The main problem in solving these tasks is the lack of semantic content in the data. Therefore already in 1998 Tim Berners-Lee proposed the concept of Semantic Web, and in 2007 - further development of this concept - Giant Global Graph [3].

The technical part of the Semantic Web is a set of markup languages for creating ontologies and schemas. Key technologies of Semantic Web are:

- XML (Extensible Markup Language) is a content markup language which provides marking that defines the semantic structure of the document;
- RDF (Resource Description Framework) is a language for describing triplets based on the model of "subject-predicate-object";
- RDFS (Resource Description Framework Schema) is a language for describing RDF schemas;
- OWL (Web Ontology Language) is a language for describing ontologies;
- SPARQL (Protocol and RDF Query Language) is a technology for creating warehouses of RDF-data, as well as a query language for retrieving data from RDF:
SWRL (Semantic Web Rule Language) is a language for implementation of logical inference in OWL;

LOD (Linked Open Data) is an approach to structuring public data in which each data item has its own URI, which provides access to a semantic description, and accordingly, the possibility of building semantic query to published data. The standardized model RDF is most often used for describing data objects in the Semantic Web.

Analysis of search possibilities of Russian resources has shown that the access to Russian-language semantic information in the field of science and technology solutions is at his early stage in comparison with a number of foreign countries. Existing information resources providing search services for Russian-language scientific publications can be divided into four groups:

- scientific search catalogs containing data on scientific publications located in the Internet;
- information systems with a web interface allowing the users to search objects over a particular set of parameters in data bases of information systems;
- social networks of scientists;
- websites containing scientific publications with the results of research, but not including built-in search tools.

The representatives of the first group include a number of international and foreign projects that support the indexing of Russian sources, such as Google Scholar (http://scholar.google.ru/) and the Science Direct (http://www.sciencedirect.com). Among the Russian developments is a search engine for scientific publications Scholar.ru (http://www.scholar.ru/). The project Scholar.ru is a catalog of references to scientific papers located in various Internet sites. The main goal of the Scholar.ru project is to gather information about free downloadable scientific publications. The search engine of the system allows searching for publications by branch of science, author, institution, journal name, year of publication (specified period), URL of website the publication is located.

The second group includes information resources such as the information system of the Russian Foundation for Basic Research (http://www.rfbr.ru/rffi/ru), the Database of Science and Technology Forum (http://www.sciteclibrary.ru/rus/catalog/tecs/), the Internet portal RSCL.RU (http://www.rsci.ru/), the website of the Russian Humanitarian Science Foundation (http://www.rfh.ru/), the website of Russian Science Foundation (http://www.rscf.ru/), etc. The analysis of search forms on these resources has shown that search engines in them are very limited: the morphology and the semantics of words is not taken into account, it is not possible to search for information within the context. In most cases the following attributes are used during the search:

- the name and/or the author(s) of the work;
- the field of knowledge (branch of science);
- the year of publication of the work (or the application for scientific research).

Among this group of resources, one should emphasize the scientific electronic library (e.g., eLIBRARY.RU) and scientometric data systems such as intelligent information systems "ISTINA" (https://istina.msu.ru/). In the "ISTINA" system developed at Moscow State University, it is used the ontology, methods of linguistic and statistical analysis of texts, contextual analysis of the Internet search results [1]. Technologies of organizing and searching information with use of ontologies have been also applied in the Information system "Natural Resources of Karelia" [8].

Examples of the resources from the third group are SciPeople (http://scipeople.ru/), Social Network "Scientists of Russia" (http://www.russian-scientists.ru) and Social Research Network Scientific Social Community (https://www.science-community.org/ru).

The fourth group includes the resources that host the archives of journals, archives of conference proceedings, competitions of scientific projects announcements, reports about the results of scientific research. However, there is no any search form on these materials, or there is only search of the entered text fragment over the site pages as, for example, on the website of the Institute of Economics of the Russian Academy of Sciences (http://inecon.org/publikacii/ katalog-izdanij-ie-ran.html) and the website of Advanced Research Fund (http://fpi.gov.ru/). Searching for information is a very
laborious task from the target user point of view, because the data is not structured in the fourth resource group. Texts of scientific papers and reports with the results of scientific research are simply attached to a web page in the form of text (pdf, doc(x) files) or even graphic files.

Thus, the review of Russian information resources in the Internet, containing data on scientific and technological solutions, showed that searching information in them is organized in most cases according to the principle of exact matching with the values of the database fields. Some sites just connected services of the known search engines. In the description of the existing resources RDF model of data representation is almost never used, and therefore, there hardly ever is SPARQL access to Russian-language data.

The trend of recent years is the use of ontologies in the websites of scientific knowledge. However, while these sites are very few and they specialize in specific fields of knowledge. In this connection, the task of creating a model of the ontological knowledge base in the field of scientific information is topical. The main carrier of scientific data is scientific publications which can be grouped into a collection of texts on a given subject. In the next section, the existing ontologies describing the scientific publications are considered.

3. Review of existing ontologies for description of scientific publications

To solve the problem of knowledge retrieval in the scientific domain it is advisable to distinguish between such concepts as “data for scientific research” and “scientific data”. The data for scientific research is the source information used for the research and expressed in numerical or any other form. At the same time scientific data is that representing the results of scientific research.

To date it is developed several ontologies to describe the process of scientific publications and scientific publishing: BIBO, complex of ontologies SPAR, CERIF, SWRC, EXPO, FRBR, SKOS, DublinCore, ISIS and etc. [5, 6, 10]. Further, the application of these ontologies is considered.

The Ontology BIBO includes the basic concepts and properties to describe bibliography references in the Semantic Web in RDF (i.e., quotes, books, papers, etc.).

Complex of ontologies SPAR allows us to describe the publishing process by using RDF. The basic ontologies of the set:

- FaBiO (FRBR-aligned Bibliographic Ontology) is the ontology for recording and publishing bibliographic records objects containing bibliographic references (journal articles, conference proceedings, books, etc.);
- CiTO (Citation Typing Ontology) is the ontology for describing the nature of citations in scientific publications (fact or statement);
- BiRO (Bibliographic Reference Ontology) is the ontology for describing bibliographic records and references, and their compilation into bibliographical collections and bibliographies;
- C4O (Citation Counting and Context Characterization Ontology) is the ontology that permits the number of in-text citations of a cited source to be recorded;
- DoCO (Document Components Ontology) is the ontology that contains structured dictionary of document components, includes structural units (e.g., paragraph, section, chapter) and function blocks (e.g., introduction, conclusion, acknowledgement, references, figures, appendices);
- PSO (Publishing Status Ontology) is the ontology that is intended to describe the state of publication at each stage of publishing process;
- PRO (Publishing Roles Ontology) is the ontology, describing the role of agents - individuals, legal entities and computational tools in the process of publication. Agents can be an author, editor, reviewer, editor or librarian;
- PWO (Publishing Workflow Ontology) is the ontology to describe the steps in the business process associated with the publication of the document.

CERIF (Common European Research Information Format) is the ontology to describe the process of research activities. On the upper level there are the essences "Person", "Project", "Organizational
block", which are associated with essences of other levels such as, for example, "Publication", "Product", "Patent".

SWRC (Semantic Web for Research Communities) is the ontology for modelling objects of research communities, such as persons, companies, publications (bibliographic metadata) and their relationships.

EXPO (EXPeriment) is the ontology to describe scientific experiments including about 200 concepts.

FRBR (Functional Requirements for Bibliographic Records) is the ontology for recording and publishing bibliography. It is divided into three groups: the first group allows one to describe the results of intellectual work (expression, manifestation, instance), the second group describes the person responsible for the results of intellectual work (person, group of persons, legal person), the third group includes essences associated with the first and the second group (concept, object, event, place).

SKOS (Simple Knowledge Organization System) is the ontology to describe the thesaurus for RDF model.

DublinCore is the ontology which includes two sets of metadata: simple and extended. The first set consists of 15 elements, the second one consists of 22 classes and 55 elements. Dublin Core Specification has the status of an official international standard ISO 15836:2009.

ISIS (Integrated Scientific Information Space) is a project aimed at integrating scientific data of Russian Academy of Science institutions. The ontology used in the project is based on the Dublin Core ontology. Four main groups of essences are distinguished in the ISIS ontology (participants of the scientific activity, scientific activity, scientific results, documents and publications).

A review of existing ontologies shows that the most well developed are ontologies to describe the "bibliographic essence" domain. However, these ontologies are insufficient for the analysis of scientific data in many cases. For example, if one needs to analyze the necessity of solving a certain scientific problem, it is necessary not only to analyze already existing publications with description of the scientific results, but also to obtain the slice of knowledge about current needs. Such knowledge could be extracted from the data about grants and projects of various funds. In addition, most of the existing ontologies were developed abroad, and therefore they are not well suited to describing bibliographic entities in accordance with Russian standards for librarianship and publishing of a series GOST 7.

### 4. Description of the ontological knowledge base

The model of the ontological knowledge base in a general form can be presented as follows (see formula 1):

\[
Z = \langle O_m, O_p, M \rangle
\]

where \(O_m\) is an upper-level ontology (metaontology) which includes the most abstract concepts and the relations between them; \(O_p\) is a set of domain ontologies; \(M\) is a mechanism of knowledge inference.

It seems reasonable to build the ontological knowledge base for intelligent support of scientific and research activity on the basis of the metaontology (ontology of the upper-level) shown in Figure 1.
Each node in the upper-level ontology is the top of the class hierarchy for the ontologies of low-level which could be presented as follows [2]:

\[ O_p = \langle C, R, S, G, T, D_s, D_g, E \rangle \]  \hspace{1cm} (2)

where \( C = \{c_i|i = 1, \ldots, n\} \) is a finite non-empty set of classes describing the notions of the domain; 
\( R = \{r_i|i = 1, \ldots, m\} \) is a finite set of binary relations between the classes, \( R \subseteq C \times C, R = \{R_{ISA}\} \cup R_{ASS} \), where \( R_{ISA} \) is an antisymmetric, transitive, non-reflexive hierarchy relation "class-subclass" defining a partial order on the set of classes; \( R_{ASS} \) is a finite set of associative relations; 
\( S = \{s_i|i = 1, \ldots, k\} \) is a finite set of slots (class attributes); 
\( G = \{g_{s_i}|i = 1, \ldots, l\} \) is a finite set of facets (slot attributes); 
\( T \) is a finite non-empty set, which determines the controlled vocabulary of the domain terms, built on a set of basic terms \( B = \{b_i|i = 1, \ldots, n\} \) being a set of names of the ontology classes:

\[ T = \bigcup_{i=1}^{n} T_i \]  \hspace{1cm} (3)

\[ T_i = \{b_i\} \cup E q(b_i), \cap_{i=1}^{n} T_i = \emptyset \]  \hspace{1cm} (4)

\( Eq(b_i) \) is a set of synonymous terms, each term is associated with base term \( b_i \in B \); 
\( D_s \) is a finite set of slots types; 
\( D_g \) is a finite set of facet types; 
\( E = \{e|i = 1, \ldots, u\} \) is a finite set of class instances.

The structure of the class is defined as follows:

\[ c = \langle Name_c, (is\ a \ c_{parent}), \langle s_1, s_2, \ldots, s_{n(c)} \rangle \rangle \]  \hspace{1cm} (5)
where \( c, \ c_{\text{parent}} \in C \) are the ontology classes connected by the hierarchy relations \( R_{\text{ISA}} \). \( s_j \in S \) are the slots, \( Name_c \in B \) is the class name being the base term of the controlled vocabulary \( T \).

Hierarchy of classes is formed by means of the relation "is-a".

The set of classes is divided into two non-intersecting sets \( C = C_{\text{abstract}} \cup C_{\text{concrete}} \). For the classes from the subset \( C_{\text{concrete}}, \) it is possible to determine instances \( e \in E \). The structure of the class-instance is analogous to the structure of class \( c \), which forms the instance:

\[
e(c) = \langle Name_e, (s^e_1, s^e_2, ..., s^e_{n(c)}) \rangle
\]

where \( s^e_1, s^e_2, ..., s^e_{n(c)} \) are samples of slots filled with specified values of the attributes.

Definition of associative relations constituting the set \( R_{\text{ASS}} \) is carried out by means of explicit indicating as the slot value the name of associated class, as well as the type of associative relation existing between these classes. To implement the associative relations, the types \( D_{\text{class}} \) (type "Class") and \( D_{\text{instance}} \) (type "Instance") are used along with standard types \( D_{\text{SS}} \) (symbol, string, float, ...):

\[
D_S = D_{SS} \cup \{D_{\text{class}}\} \cup \{D_{\text{instance}}\}.
\]

Defining the types \( D_{\text{class}} \) and \( D_{\text{instance}} \) involves specifying the additional argument – the associated class. If one of the slots of the class \( c_1 \) has the type \( D_{\text{instance}} \) with associated class \( c_2 \), then as of the slot values when creating instances of the class \( c_1 \) can be used the instances of the set \( Tr(c_2) \) being transitive closure of \( c_2 \) related to \( R_{\text{ISA}} \), including class \( c_2 \) and all its subclasses below the hierarchy:

\[
Tr(c_2) = \{c_2\} \cup \{c_1 \in C | \exists R_{\text{ISA}} (c_1, c_2) \}
\]

In this case, classes \( c_1 \) and \( c_2 \) are related with associative relation, i.e. \( \exists R_{\text{ASS}} (c_1, c_2) \). If one of slots of the class \( c_1 \) has the type \( D_{\text{class}} \) with associated class \( c_2 \), then as the slot value when creating instances of the class \( c_1 \) classes of the set \( Tr(c_2) \) can be used. In this case, the classes \( c_1 \) and \( c_2 \) are also related with associative relation, i.e. \( \exists R_{\text{ASS}} (c_1, c_2) \). Thus, as the slot value can not be only a sample of associative class but also a base term. The latter can be used for describing complex objects like the requirements types with terms of the controlled vocabulary.

Semantic relationships between the individual domain areas could be described with use of a standardized model RDF for describing data objects for Semantic Web. This will make it possible to connect the data presented in a machine-readable format.

5. Implementation of the ontology in Protégé environment

Footnotes should be avoided whenever possible. If required they should be used only for brief notes that do not fit conveniently into the text.

The ontology has been developed in Protégé editor. Protégé is a free, open source platform for building domain models and an application that supports the creation of ontologies. Ontology built in the editor, can be saved in a variety of formats including RDF, OWL and JSON-LD. The latter format is used for the semantic representation of the linked data. The editor supports the ability to work with SWRL and allows one to ask questions to the ontology with use of SPARQL language.

The basis of the ontology consists of the following classes:

- class "PublicationProduct" which corresponds to the concept "publication" and consists of subclasses describing the types of publications;
- class "ResearchResults" which corresponds to the concept "scientific result" and consists of subclasses reflecting such concepts as "the original attempt", "assessment of state", "theory", "paradigm", "tool", "model";
- class "ScientificEvent" which corresponds to the concept "scientific event" and consists of subclasses reflecting such concepts as "announcement of competition", "conference", "forum";
- class "ScientificProblem" which corresponds to the concept "scientific problem" and consists of subclasses reflecting such concepts as "substrate problem", "structural problem", etc.;
• class "Space" which corresponds to the concept "space" and consists of subclasses reflecting such concepts as "geographical location", "political space", "cultural space", "natural space", etc.;
• class "SubjectOfResearch" which corresponds to the concept "research subject" and consists of subclasses reflecting such concepts as "person", "research team", "organization";
• class "Time" which consists of subclasses reflecting such concepts as "period", "year", "quarter", "month".

The class hierarchy for the concept "scientific publication" is shown in figure 2.

![Figure 2. Hierarchy of classes](image)

Development of the first version of the hierarchy of classes for this concept is made on the basis of GOST 7.60-2003 and GOST R 7.83-2013. According to GOST 7.60-2003, edition is "a document intended for spreading the information contained in it, past editorial and publishing process, self-designed, has an output".

By physical characteristics all publications are divided into the print editions and the electronic ones. The classes "Print" and "ElectronicPublishing" have been created to display this division. Electronic editions are in turn divided into independent electronic editions, derivative electronic publications and electronic copies of the publications.

All publications can be divided into two groups – already published and unpublished yet. The first group includes such publications as monographs, collections of scientific papers, conference, congress, symposium proceedings, thesis abstract, preprint, an inventor’s certificate or a patent. The second group includes the scientific and technical reports, theses, deposited manuscripts. To display this division classes "PublishedDocument" and "UnPublishedDocument" have been created.

6. Conclusion
Currently, semantic technologies are widely used in various data management systems to provide for extraction of knowledge from linked data. Even elementary empirical comparison of the results of
search engines implementing the idea of "knowledge graph" and "a giant global graph" with the results of classical search engine shows that the first engines are indisputable advantages. Present paper discusses the possibility of semantic technologies for data description. The extent of application of semantic technologies to the organization of information resources, providing search services for Russian-language scientific publications, was investigated. The ontology in the field of scientific research in Russia was proposed based on the study of a number of approaches given in [1] and [8]. The authors see their further work in creating a detailed mechanism of linking the domain ontologies through the developed upper-level ontology in the integrated ontological knowledge base.

Acknowledgments
The reported study was funded by Russian Ministry of Education and Science, according to the research project No. 2.2327.2017/4.6.

References
[1] Afonin S A et al. 2014 Intellectual system of case research of scientific and technical informationed V A Sadovnichiy(Moscow: Moscow State University)
[2] Avdeenko T V, Bakaev M A 2016 Scientific Bulletin of NSTU 3 84
[3] Berners-Lee T 2007 Giant global graph. Retrieved from: http://dig.csail.mit.edu/breadcrumbs/node/215
[4] Gorshkov S. 2016 Vvedenie v ontologicheskoe modelirovanie: uchebnoe posobie (Ekaterinburg)
[5] Peroni S, Shotton D 2012 Web Semantics: Science, Services and Agents on the World Wide Web 17 33-43
[6] Ruiz-Iniesta A, Corcho O 2014 Proceedings of 4th Workshop on Semantic Publishing (SePublica)
[7] Singhal A 2012 Introducing the knowledge graph: things, not strings. Retrieved from: https://googleblog.blogspot.ru/2012/05/introducing-knowledge-graph-things-not.html
[8] Vdovicyn V T, Lebedev V A Information resources of Russia 1 7
[9] Xiong C, Power R, Callan J 2017 Explicit Semantic Ranking for Academic Search via Knowledge Graph Embedding. Proceedings of the 26th international conference on world wide web pp 1271-1279
[10] Zaharov A A, Filipov V I 2009 Proceedings of the XI All-Russian Research Conference RCDL’2009 (Petrozavodsk)