Solutions for system analysis and information support of the various activities in the Arctic

Andrey Oleynik*, Pavel Lomov, Alexey Shemyakin, Alexey Avdeev

Institute for Informatics and Mathematical Modelling of Technological Processes of the Kola Science Center Russian Academy of Sciences, Fersman st. 24a, Apatity, Murmansk region, Russia

Abstract

Comprehensive use of data and knowledge obtained within different disciplines is necessary for the scientific substantiation of activities in the Arctic zone and for a system analysis of the possible consequences of this activity. Information resources created so far allow the access to a variety of data on the Arctic. The authors propose the solution for task of data consistency ensuring in the field of combined presentation and use of data and knowledge of interdisciplinary research. The proposed solution is based on the joint use of relational database and ontology. The developed structure and mechanisms of the database maintenance provide a uniform representation of the information about results of the researches executed in the framework of various disciplines. The ontology is a high-level global schema of the information system and it provides a dictionary that is used to formulate a database query in terms of a subject domain. In this work, ontology is implemented as a system of small fragments - ontology design patterns. The patterns use makes it possible to perform efficient preliminary database indexing, which ensures faster execution of user queries.

Key words: activity in Arctic, interdisciplinary researches, information and analytical support, database, ontology

DOI: 10.5817/CPR2017-2-27

Introduction

In the last decades, qualitative changes in approach to the Arctic research have happened. The Arctic and subarctic territories are becoming a zone of expanding economic activity of the human community. The experience accumulated by mankind shows that this activity can substantially change the existing ecosystems of the ter-

Received May 25, 2017, *accepted* November 8, 2017.

*Corresponding author: A. Oleynik <oleynik@iimm.ru>

Acknowledgements: The authors thank the Russian Foundation for Basic Research for the support in the frame of the projects 15-29-06973 "Development of methodology, model toolkit and information technologies for system risk assessment of new exploration of the Arctic region" and 16-07-00562 "Development of the model, methods of integration and use of interdisciplinary knowledge and data for management support of the Russian Arctic complex development". We would also like to thank the researchers of the Pitirim Sorokin Syktyvkar State University (Syktyvkar, Komi Republic, Russia) for the constructive discussion of the database structure.

ritories. For the Arctic zone, the danger of irreversibility of the consequences of human activity is particularly high. In this regard, it is extremely important to predict and analyze the possible consequences of any activity in the Arctic zone. To solve the tasks of analysis, adequate information support is needed, which is to provide the user with accessible data and knowledge on the problem of interest to him.

To date, there are many information systems that provide efficient storage and access to data related to various aspects of the Arctic macrosystem. The macrosystem is the aggregate of different systems (biological, social, technological and others) in interaction. Some of them were presented at the Session "Arctic Data and Information Science meets System Science" of the Arctic Science Summit Week 2017.

Among the presented solutions are: Metadatabase "Global Index of Vegetation-Plot Databases" (GIVD) [1] in which you can find metadata about major vegetation-plot databases of the world, including data on Arctic territories; "PolarData Catalogue" [2] which is a database of metadata and the data that describe indexes and provides access to diverse data sets generated by Arctic and Antarctic researchers. Very promising is the projects "Arctic Data Ecosystem Map" [3] aimed to establishing a map of the arctic data management "universe". The map should provide information on projects, several services and relationships, as well as geographical localization of information entities. The Fennoscandian Ore Deposit interactive map and Database (FODD) are presented on the Geological Survey of Finland web-site [4]. The results of the geological surveys of Finland, Norway, Russia and Sweden are accumulated in the FODD. This information resource is essential from the point of view of the mining industry development in the western part of the Arctic zone of the Russian Federation (AZRF).

Recently, several Russian developments focused on the integration of information on scientific research exist. They are *e.g.* web-server of the Siberian Branch of the Russian Academy of Sciences [5], the information system "Archives of the Russian Academy of Sciences" (IS ARAN) [6], the information-analytical system "Natural Resources of Karelia" (Vdovitsin et Lebedev 2012). Variant of the information base structure for support of the academic and university science interaction were offered earlier by the authors of this paper (Oleynik et Shtivelman 1998). The work of the team of authors, which includes Serebryakov V.A., Bezdushny A.N., *etc*. is devoted to the project "Integrated Information Resources System (ISIR) RAS" using Semantic Web technologies (Bezdushny et al. 2006, Bezdushny et Serebryakov 2010, Serebryakov 2014).

However, existing information resources do not fully ensure the effective sharing of data and knowledge obtained in the various subject areas (ecology and mining geophysics and oceanography, demography, sociology and others). The creation of tools for informational support of interdisciplinary research in the Arctic remains an urgent task. In the context of this work, knowledge is a result of cognitive activity which was systematized and recorded in a form suitable for further use. Printed fact sheets, papers, report materials, models, algorithms and *etc*. are the possible forms of knowledge representation.

In this paper, we consider the combined use of the relational database management system (DBMS) and ontology based information system for the creation of information support tools to interdisciplinary research and various activities in the Arctic.

Material and Methods

Research results of the various institutions of the Kola Science Center of the Russian Academy of Sciences [7] was used by the authors as a "test information source" of interdisciplinary data and knowledge. The tools for information support of interdisciplinary research of the Arctic are the two types of systems that complement each other. The first is a relational database of researches results from various scientific fields.

The classic top-down approach was applied for development of the database structure. At the first stage we define a set of basic information entities which should be represented in the database. The authors proceeded from the assumption that the "point of intersection" of investigations of various disciplines is an object under research (research object). It can be either a real physical object, or an abstract virtual object. Each research (research project) is carried out by the researcher (research team) in order to investigate some specific properties of the research object. The properties under investigation determine the subject of research. The researcher gets results using certain methods and research tools. The initial conceptual model of the interdisciplinary research database being created includes five basic entities cover to research: "object", "subject", "researcher, "method" and "result". The relation between these entities is realized by the introduction of the entity "project" (Fig. 1.).

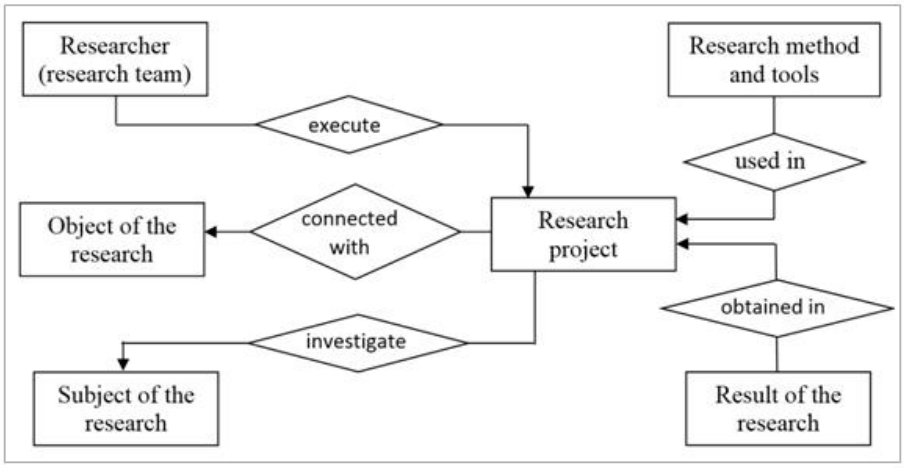

Fig. 1. ER-diagram of basis information entities for interdisciplinary database.

The choice of the relational model is determined by the orientation to the database distributed realization with horizontal and vertical fragmentation. This, on the one hand, will increase the efficiency and fault tolerance in data management, and on the other hand, in some cases, will speed up the response of the system to user queries. Data on the research should be entered into the database directly by their authors, or by "authorized" representatives of research teams.

The Django 1.8 web framework, the Python 3.5.2 programming language were selected as the tools for development of the database. The framework Django signifi-

cantly facilitate the development of complex web-applications on the language Python.

An essential disadvantage of databases is the use of explicit values of attributes both for providing links between tables and for targeted access to stored data.

In recent decades, ontology has been used to structure, formalize and unify the representation of knowledge for the purpose of its repeated and flexible use in information systems. As noted, in particular, in (Lapshin 2010), the term "ontology" first appeared in the work of T. Gruber (Gruber 1991), in which various aspects of the interaction of intellectual systems between themselves and with human were considered. Nowadays, ontology is considered as a declarative knowledge description, made in a formal language and provided with a certain classification of specified knowledge, which allows a human to conveniently perceive it (Lapshin 2010). Resource Description Framework (RDF) [8] and Web Ontology Language (OWL) (Grau et al. 2008) are used as the means of describing ontologies. RDF allows to describe an ontology as a set of triplets of the form: “subject-property-object”. OWL provides great expressive possibilities and at the same time ensures formal decidability of ontologies, described by it. This provides extensive inference capabilities on the ontologies described with OWL.

The main advantage of using ontology as the basis for knowledge storage is the description of the subject area in the form of a logical theory that includes the concepts of the subject domain and relations. The disadvantage of ontologies in information-analytical systems is the lower performance of query execution than when using relational DBMS. This is cost of the expressive possibilities offered by ontologies and the absence of a "rigid" data scheme.

The use of ontologies in information systems can be carried out in two ways. The first one is using an RDF triple store with an ontology loaded into it. In this case, the ontology is treated as an RDF document consisting of a set of triplets. The RDF repository allows to access its content by executing SPARQL queries (Kollia et al. 2011). Another way is to combine an ontology with a relational database to provide ontology-based database access (ODBA). In this scenario, the ontology is a high-level global schema and provides a dictionary that is used to formulate the query in terms of the subject domain. Based on the specified mapping scheme, this request is rewritten into a traditional SQL query and passed to the DBMS for execution. Thus, the ontology is used to provide access to a data source represented by a given database, whereas it does not contain data itself.

To develop an information support system for interdisciplinary research, it was decided to use the ODBA approach taking into account a predictably large amount of data, which would require a good query performance. Ontology in this case will play the role of semantic representation of the information source.

Results

Relation database of interdisciplinary research

The set of the basic information entities is presented in Fig. 1. All of them have hierarchical, in fact – recursive, structure. For example, an instance of the top level entity “researcher (research teams)”, as a rule, is some research organization. The organization consists of various research units, which can also be composite entities. At the lower level of the hierarchy of “researcher” is a researcher-person. Special relations

(tables) are introduced into relational structure for the correct representation of recursive and multiple relationships of conceptual level. Therefore, a researcher-person should only be associated with the research unit of the lower level (*e.g.* laboratory). His belonging to larger research units (*e.g.* department or branch which include laboratory) and, ultimately, to the research organization will be derived from hierarchical relations between corresponding units. The proposed scheme allows to connect a person with several organizations and correctly to describe temporary research teams. A similar principle is also used when representing hierarchical relationships between entities of other classes. The developed relational structure provides storage of structured information about the objects of the Arctic zone, researchers (research teams) who investigate these objects from various points of view, research projects, methods and results of research.

Four levels of administration for the maintenance of the database. The first level is a system (database) administrator who response for the "technical" side of the system.

The second level is administrator of the general data. He manages shared data, including user management, form and modify database structure (if it is needs), input and systematize basic objects of research, registrants a new research teams and *etc*.

The next level is administrator of an organization (research team) data. They responses for structure and maintenance of data about their organization (team). This group can grant the rights to the projects administrators (which are controlled by the organization) or maintenance this projects data by themselves.

Administrators of the projects are responsible for maintaining the data about "their" projects. They use a common data (objects, organizations, funding sources) entered into the database by administrators of the higher levels and load information about concrete projects (subject, methods and results of the research) into database.

Database users are divided into two categories: authorized users and guests. Researcher can contribute to the database information about himself after registration in the system approved by the administrator a certain level. In this case he is authorized user. Guest can use database without registration. The guest cannot enter the information into the database, but can search and view data of interest.

The test version of the database was implemented for Russian-speaking users. The main goal of the first implementation is a practical test of the efficiency and effectiveness of the solutions proposed by the authors.

Implementation of ontology-based database access to the database

Consider the general structure of the developed ontology of the interdisciplinary research and its application as component of ODBA system. The ontology consists of several modules, defined in accordance with the level of abstraction of their concepts and the functional purpose of module contents. Each module is a file containing a separate ontology, described by OWL (Fig. 2). The Information Artifact Ontology (IAO) [9] is an extension of one of the most widespread upper level ontologies - the Basic Formal Ontology (BFO) (Grenon 2003). Elements from BFO are used to describe abstract objects, processes and events that are invariant to subject domains. Elements of IAO represented processes of obtaining information, their participants, as well as information sources and information carriers.

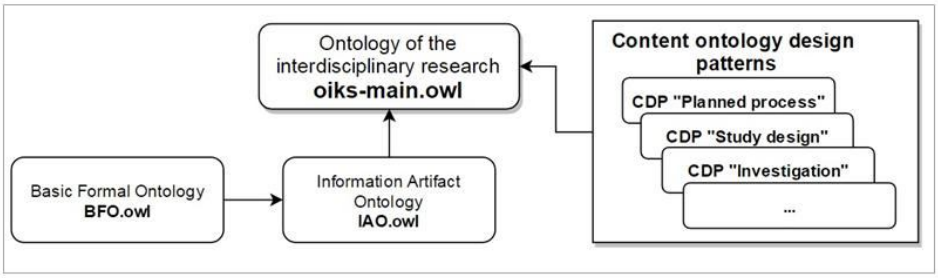

Fig. 2. Module structure of ontology of the interdisciplinary research.

At the same time, they not only form a conceptual system, but also define the domain knowledge representation methods which are correct in accordance to the conception of IAO.

The main part of the developed ontology is a set of Ontology Content Design Patterns (CDP) (Gangemi 2005). Each CDP is represented as a mini-ontology and solves one ontological modeling problem. The patterns contain concepts and relationships, defined in the IAO. The use of CDP is the basis of eXtreme Design methodology (XD-methodology) (Blomqvist et al. 2010, Masolo et al. 2003) for ontology development. Segmentation of the ontology into separate patterns allows to save its user from manipulating the entire conceptual system. In addition, as the authors of the XD-methodology emphasize, this allows to ensure the quality of the ontology being developed, since each pattern is a proven solution that has proved its effectiveness.

To apply developed patterns for domain concepts representation user chooses a suitable CDP and makes its specialization, which consists in determining the heirs of its elements. Selection is based on a set of qualification questions associated with each pattern. They are formulated in natural language and indicate what information can be obtained by using an ontology containing corresponding pattern.

As an example, consider the "Investigation specification" pattern. The UML diagram of its classes and relations is shown in Fig. 3.

This pattern allows a user to define parts of the research specification: description of the objective of the study (Objective specification), actions (Action specification), object (Investigation object specification) and subject (Investigation subject specification) of the study and the applied method (Study design).

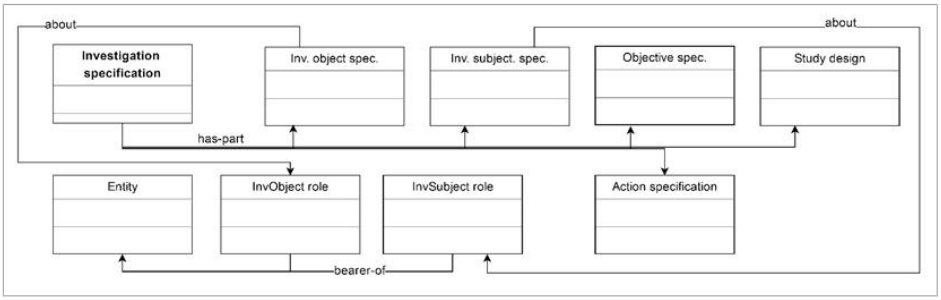

Fig. 3. UML schema of CDP "Investigation Specification".

For this pattern, the following qualification questions were defined:

- What method is planned to be used to study this object?
- What are the objectives/plan/method of research?
- In what studies was planned to apply this method/pursue this goal?

Next, consider the implementation of ODBA using the developed ontology. In this case ontology contains only expressions described classes (concepts) and relations between them (Tbox). Expressions about concrete instances of classes (Abox) are generated from results of execution SQL queries. Such generation as well as rewriting the initial query is based on mapping that determines a correspondence between DB schema and Tbox of ontology. Thus, the main work in ODBA implementation is to define a mapping that is performed semiautomatically or manually by a knowledge engineer.

For each of the developed patterns, DB-scheme fragment was defined. The mapping of ontology elements and data schema was performed in accordance with W3C Recommendation "A Direct Mapping of Relational Data to RDF" [10]. Fragment of the scheme corresponding to the CDP "Investigation specification" is shown on Fig. 4. As an implementation of OBDA system, Ontop was used (Kontchakov et al. 2014). Key features of Ontop are its solid theoretical foundations, a virtual approach to OBDA, extensive optimizations, and its support for all major relational databases.

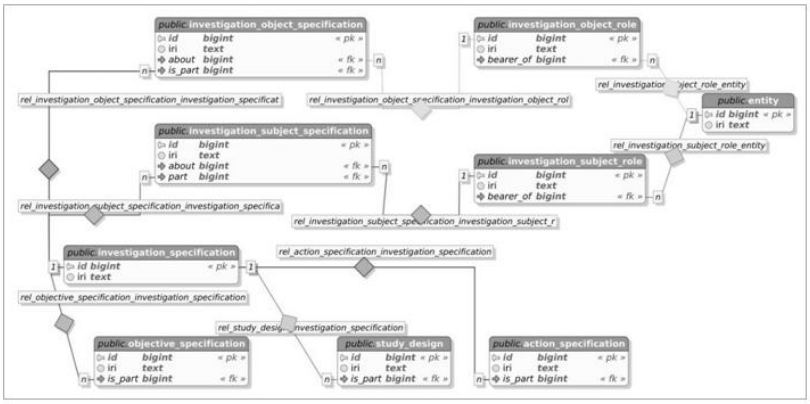

Fig. 4. DB schema for CDP "Investigation specification".

For each pattern, a set of mapping rules in Ontop format was developed. Each rule corresponds to one database table and includes a target - a set of triplets and a source - SQL query, the results of which are used to generate the triplets. For example, consider the rule for the "Study design" table, linked by the foreign key with the "Investigation specification" table:

```
mappingId   [study_design]
target      :study_design-{id} a                    obo:study_design;
            :study_design-{id} obo:is-part-of  :investigation-specification-{is_part}.
source      select "id", "is_part" from "public"."study_design"
```

This rule cretate two triplets ("something" – "is an instance of" – "study_design", "something" – "part of" – "study_design") from a SQL query. Note that mapping is also used by the Ontop system for rewriting SPARQL query.

An effective way to increase the performance of executing SQL queries is to build indexes for DB tables. In case of patterns, it makes sense to create indexes on the basis of their qualification questions. It is assumed that answers to them will be of primary interest to a user. Building indexes in this case assumes the formation of a SPARQL query corresponding to the qualification question, its rewriting via an ODBA system into an SQL query, and an analysis the execution plan of the last one.

To assess the effectiveness of using indexes built on the base of qualification queries, an experiment was conducted. It included estimation of the execution time of a series of SPARQL queries using the Ontop system over a test set of data before and after its indexation. The test suite volume was 6300000 rows in each of the tables that correspond to the "Investigation specification" pattern. As the RDBMS, PostgreSQL 9.6.3 was used.

The series consisted of 20 SPARQL queries, corresponding to the previously discussed qualification question "What method is planned to be used to study this object?". Prior to indexing, the average execution time of one query was 1700 ms, after indexing - 210 ms.

Discussion

A combination of the database and the ontology allows us to present the contents of a data source at a conceptual level. Such combination will provide a familiar dictionary for generating queries for a user.

When rewriting the ODBA query, the system will perform its logical analysis and extension of the final SQL query based on the results of the logical deduction. A high speed of SQL query execution is provided by a relational database. ODBA allows us to use several heterogeneous data sources for querying when defining mapping schemes. This simplifies the horizontal scaling of the received information system.

A drawback of the proposed approach is the presence of some restrictions on expressiveness of the ontology description language. However, this is a necessary condition for ODBA systems which ensure an acceptable complexity of rewriting the original query into so-called first order queries (FO-queries) (Kontchakov et al. 2014). This reduces logical inference capabilities, but it doesn't really matter in our case.

The used approach of developing the ontology of an interdisciplinary research by defining separate patterns makes it possible to simplify its application by the user. When using a particular pattern, the user operates a limited conceptual system, which addresses to one specific task of ontological modeling and its results are clearly defined as the set of qualifying questions. The use of patterns makes it possible to provide predictability of queries and makes it possible to efficiently index the database tables. The conducted experiments showed that indexing significantly improves the speed of such requests.

References

BEZDUSHNY, A. N., SEREBRYAKOV, V. A. (2010): *Yedinoye nauchnoye informatsionnoye prostranstvo (YENIP) RAN* [The Unified Scientific Information Space (ENIP) of the Russian Academy of Sciences], available at: http://www.benran.ru/magazin/inaros/seminar/2010/1.doc (Accessed at 24 May 2017) (In Russian).

BEZDUSHNY, A. A., BEZDUSHNY, A. N., SEREBRYAKOV, V. A. and FILIPPOV, V. I. (2006): Integratsiya metadannykh Yedinogo Nauchnogo Informatsionnogo Prostranstva RAN [Integration of the metadata of the Single Scientific Information Space of the Russian Academy of Sciences] available at: http://www.icsti.su/portal/newproblem/index.php?m=25 (Accessed at 24 May 2017) (In Russian).

BLOMQVIST, E., PRESUTTI, V., DAGA, E. and GANGEMI, A. (2010): Experimenting with eXtreme Design. *Proceedings of EKAW 2010*, LNCS 6317. Springer. Berlin/Heidelberg/New York: 120-134.

GANGEMI, A. (2005): Ontology Design Patterns for Semantic Web Content. *Proceedings of the Fourth International Semantic Web Conference*, Springer. Galway, Ireland: 262-276.

GRAU, B., HORROCKS, I., MOTIK, B., PARSIA, B., PATEL-SCHNEIDER, P. and SATTLER U. (2008): OWL2: The next step for OWL. *Web Semantics: Science, Services and Agents on the World Wide Web*, 6(4): 309-322.

GRENON, P. (2003): Spatio-temporality in Basic Formal Ontology: SNAP and SPAN, Upper-Level Ontology, and Framework for Formalization: PART I. *IFOMIS Report 05/2003*, Institute for Formal Ontology and Medical In-formation Science (IFOMIS), University of Leipzig, Leipzig, Germany. 166 p.

GRUBER, T. R. (1991): The role of common ontology in achieving sharable, reusable knowledge bases. Principles of Knowledge Representation and Reasoning. *Proceedings of the Second International Conference.* J.A. Allen, R. Fikes, E. Sandewell - eds. Morgan Kaufmann: 601-602.

KOLLIA, I., GLIMM, B. and HORROCKS, I. (2011): SPARQL Query Answering over OWL Ontologies, *Proceedings of The Semantic Web: Research and Applications: 8th Extended Semantic Web Conference, ESWC 2011*, Heraklion, Crete, Greece, May 29-June 2, 2011, Part I, G. Antoniou, M. Grobelnik, E. Simperl, B. Parsia, D. Plexousakis, P. De Leenheer, and J. Pan eds. Berlin, Heidelberg: Springer Berlin Heidelberg: 382-396.

KONTCHAKOV R., REZK, M., RODRÍGUEZ-MURO, M., XIAO, G. and ZAKHARYASCHEV, M. (2014): Answering SPARQL queries over databases under OWL 2 QL entailment regime. *In*: The Semantic Web – ISWC 2014. Lecture Notes in Computer Science, vol. 8796. Springer, Heidelberg, pp. 552-567.

LAPSHIN, V. A. (2010): *Ontologii v komp'yuternykh sistemakh. Rol' ontologiy v sovremennoy komp'yuternoy nauke* [Ontologies in computers systems. The role of ontologies in modern computer science], available at: http://rsdn.org/article/philosophy/what-is-onto.xml (Accessed at 24 May 2017) (In Russian).

MASOLO, C., BORGO, S., GANGEMI, A., GUARINO, N., OLTRAMARI, A. and SHNEIDER, L. (2003): WonderWeb Deliverable D18, available at: http://wonderweb.man.ac.uk/deliverables/documents/D18.pdf (Accessed at 24 May 2017).

OLEYNIK, A. G., SHTIVELMAN, YA. E. (1998): *Razrabotka struktury yedinoy informatsionnoy bazy dannykh v ramkakh integratsii nauchnykh issledovaniy i vysshego obrazovaniya v regione* [Development of a unified information database in the framework of integrating research and higher education in the region] *Sistemy informatsionnoy podderzhki regional'nogo razvitiya*. Apatity, KSC RAS: 19-24 (In Russian).

SEREBRYAKOV, V. A. (2014): Research and development in the computing centre of RAS in the field of distributed information systems. *Novosibirsk State University Journal of Information Technologies*, 12 (3): 100-123, available at: http://www.nsu.ru/xmlui/bitstream/handle/nsu/6870/09.pdf (Accessed 24 May 2017) (In Russian).

VDOVITSIN, V., LEBEDEV, V. (2012): *Tekhnologii informatsionnogo obespecheniya nauchnykh issledovaniy v IAS «Prirodnyye resursy Karelii»* [Technologies of information support of scientific research in IAS "Natural Resources of Karelia"] *Informatsionnie resursi Rossii* ,1: 7-12 (In Russian).

Web sources / Other sources

[1] Global Index of Vegetation-Plot Databases, available at: http://www.givd.info/givd/faces/databases.xhtml. (Accessed 24 May 2017).

[2] PolarData Catalogue, available at: http://www.polardata.ca (Accessed 24 May 2017).

[3] Arctic Data Ecosystem Map, available at: http://arcticdc.org/products/data-ecosystem-map (Accessed 24 May 2017).

[4] Fennoscandian Ore Deposit interactive map and Database (FODD) available at: http://en.gtk.fi/informationservices/databases/fodd/ (Accessed 24 May 2017).

[5] Information Resources of Siberian Branch of the Russian Academy of Sciences, available at: http://www-sbras.nsc.ru/win/nsc-net/info99.html (Accessed at 24 May 2017).

[6] Information System “Archives of Russian Academy of Sciences”, available at http://isaran.ru/?q=welcome (Accessed at 24 May 2017) (In Russian).

[7] Kola Science Centre of the Russian Academy of Sciences, available at http://www.kolasc.net.ru/english/ (Accessed at 24 May 2017).

[8] W3C Recommendation “RDF Schema 1.1”, available at: https://www.w3.org/TR/2014/REC-rdf-schema-20140225 (Accessed at 24 May 2017).

[9] The Information Artifact Ontology (IAO), available at: https://github.com/information-artifact-ontology/IAO/ (Accessed at 25 May 2017).

[10] W3C Recommendation “A Direct Mapping of Relational Data to RDF”, available at: - https://www.w3.org/TR/rdb-direct-mapping (Accessed at 24 May 2017).