Ontological Approach to Matching of the Domain Competence of Specialists for Research Projects

Rogushina J.V.
Institute of Software Systems of National Academy of Sciences of Ukraine
Kyiv, Ukraine
ladamandraka2010@gmail.com

Gladun A.Y.
The International Teaching scientific center of information technology and systems NAS of Ukraine, Kyiv, Ukraine
glanat@yahoo.com

Abstract—objective methods of competence evaluating of research project developers based on the semantic comparison of the project description and documents that characterize the competence of developers in the chosen subject domain are proposed. We propose to acquire ontological knowledge from the Web open environment - Wikis, scientometric databases, personal blogs, official websites of organizations and metadata, domain ontologies etc. Specialized ontology of scientific activities oriented on unified describing of qualification terminology is developed.

Keywords—research project; ontology; competence; scientometric indicators.

I. INTRODUCTION

Today it is difficult to imagine practically all spheres of human activities without the use of information technologies and, consequently, without the development of research projects that allows new innovative technologies and applied intelligent information systems (IIS). Preparing of the request for research project is a complex interdisciplinary problem which solution needs in use of the modern technologies of the intelligent information processing. In particular, these technologies allow to evaluate the originality and relevance of the project, to find the most relevant experts and actors, to predict the likelihood of its success.

These technologies are oriented on information processing on the semantic level by use of apply knowledge from subject domain of the project, as well as knowledge on research activities in general. The analysis of international experience shows that the use of ontological models for knowledge formalizing is one of the most promising approaches to such problems.

II. PROBLEM STATEMENT

We analyze a particular case of complex information retrieval task that deals with estimation of matching of researcher qualification with scientific research project. As experience shows, this parameter is crucial for predicting of the project success and therefore it should be taken into account first of all in decision-making on project funding and grants, especially for new and interdisciplinary research fields where traditional formal methods are not efficient.

We propose to use semantic processing of information from open sources (for example, from information resources (IRs) from the Web) to match the scope of researchers competencies to the subject domain of proposed project: specialist can be highly qualified in some domain but be very poorly prepared to participate in project from another domain from the same sphere of knowledge. This matching problem can be divided by the following subtasks:

- Generation of the set of natural language IRs that describe the project;
- Selection of documents that describe qualification and experience of particular researchers (with indicating of the importance level and trust of each source);
- Acquisition of the formalized project model from its description;
- Building of the formalized profile of researcher on base of his documents (directly proposed by applicant or retrieves from the Web and other open sources);
- Matching of researchers profiles with project domain model and a quantitative estimation of their proximity.

The main source of information about the project is its natural-language description (application, request or requirements specification), as well as additional IRs such as external ontologies, Wiki pages proposed by project authors that contain structured and semantically meaningful information regarding the considered domain.

III. COMPETENCE AND EXPERTISE OF RESEARCHERS

The most difficult from these subtasks is an analysis of information about the project participants. Some part of this information about them is clearly formalized and can be clearly evaluated without taking into account the semantics of domain and additional knowledge about the project (level of education, experience in the relevant field, the presence of previously developed projects). But often this information is not sufficient to determine competence on the development of a research project in the new and rapidly changing domains.

It is necessary to distinguish concepts "competence" and "expertise". The person knowledge and experience that have to provide successful execution of various tasks in accordance with some rules, laws, etc. characterize competence, and expertise is the relation between the person and the competence which means that certain person has this competence.

In our case, competencies are the characteristics and tasks that are necessary for the development of a research project, and

expertises are specialist properties, their characteristics, experience and capabilities. Expertise can be defined based on the analysis of professional activity of specialist, his awareness of the science and technology achievements, his understanding of the investigated problems and ways of their solving. In the field of research projects development there are both formal and informal requirements for specialists that can be considered as competencies.

One of the most objective criteria of evaluation of competence sphere for scientists is an analysis of their publications presented by the Web – various scientific articles, papers, reports and presentations that usually are represented by natural language texts or structured metadata.

Document pertinence to discovered project depends on such parameters as the number of references to the main terms in the document and the number of main project terms used in the document. There is a lot of works regarding the automatic determination of competence on the basis of documents.

However, different information sources of the Web have different assessment of the information quality. it is important to take into account the evaluation of researcher activities by the scientific community – by the presence of references by other authors in their work. In addition, information about researchers can be imported from a knowledge bases of intelligent applications that provide personalized information services. For an objective assessment of competence of developers and experts it is advisable to use external quantitative parameters that reflect the overall efficiency and the intensity of their scientific activities.

Thus, we propose to use as a source of information about the researcher expertise the following IRs [1]:

- official documents acknowledge education and experience (for example, university diploma, academic degree, certificates and awards);
- IR that describe the semantics of these official documents (passports of specialties and disciplines, the requirements for obtaining of scientific titles and degrees, job descriptions, taxonomy of national academic degrees, etc.);
- texts of published articles, abstracts, monographs, textbooks, technical reports, patents and other intelligent property presented in the form of natural language documents and published by the Web that are rate by scientometric databases;
- Wiki pages of persons and organizations that provide structured presentation of information;
- personal blogs and pages from social nets;
- official Web pages of organizations and institutions deal with applicants (for example, membership in the international scientific and technical societies, editorial boards of scientific journals, cooperation with the National Academies of Sciences, educational institutions).

IV. SCIENTOMETRIC CHARACTERISTICS OF RESEARCH ACTIVITIES

The effectiveness of scientific activities of individuals, groups and organizations can be evaluated using both qualitative and quantitative indicators. Qualitative ratings are based on opinions of domain experts. However, the subjectivity of such assessments significantly reduces the reliability of results, and the lack of a quantitative expression complicates their use.

The term "Scientometrics" was introduced in 1969 by V.V.Nalimov [2]. The increased interest in scientometric indices is caused primarily by the ability to automate the evaluation of the results of scientific activity [3].

Scientometric indicators are suitable for estimating the results of fundamental research which demand is assessed by the references of the scientific community. Scientometrics is a science that involves statistical studies of the structure and dynamics of scientific information flows. It studies the evolution of science through a numeric measurement of scientific information, such as the number of scientific articles for a certain period of time, citation, etc.

Now generation of the researcher rankings use various parameters such as number of publications (in total or separately for types – monographs, articles, theses, publications in journals indexed in the Web of Science, Scopus or Google Scholar, etc.) and references to them. Sometimes the volume and impact factor of publications are took into account. Integral criteria based on these ones are formed.

The effectiveness of scientific activity can be evaluated using both qualitative and quantitative indicators. The most effective and the most common characteristics of scientific work productivity, in particular, are the Hirsch index and impact factor [4, 5].

In 2005, Hirsch proposed a new index - *h-index* [6] defined by the maximum integer h which means that the author has published h papers and each of them was referred in other articles at least h times. Hirsch index is popular because of easy calculation and insensitivity to the typical methods of factitious improvement of considered above scientometric indicators.

Hirsch index can be calculated using a free public scientometric database on the Web (for example, Google Scholar, Elibrary.ru, ADS NASA), and the database with a paid subscription (for example, Scopus or ISI Web of Science). However, many paid databases give the h-index of scientists in the public domain. It should be noted that the Hirsch index has different rating meanings of the same researcher in dependence of the indexed IR set.

Hirsch index gives more objective results in the case of withdrawal of the author references to their own articles. For example, in the ranking of scientists of Ukraine according to the Hirsch index calculation is made on the database Scopus with the withdrawal of the authors of references to their own articles.

Impact factor indicates the average number of links on each article that was published in the journal for the next x years after its release [7, 8]. This quantitative measure of the importance of a scientific journal is calculated annually by the Institute for Scientific Information (ISI) and is published in the Journal Citation Report.

Impact factor allows to compare different journals and research groups by formal parameters [9]. Generally, the calculation of the impact factor is based on a three-year period. The impact factor of the journal A for year x is calculated by the formula:

$$\text{Im } p(A,x) = \frac{Cit(A,x-2,x) + Cit(A,x-1,x)}{Pub(A,x-2) + Pub(A,x-1)},$$

where $Cit(A,y,z)$ is the number of references during the year on the z articles published in the magazine A during the year y in magazines publications monitored by Institute for Scientific Information, and $Pub(A,y)$ is the number of publications in the journal A during the year y.

Citation Index is an accepted by the scientific world measure of the significance of scientific work of some scientist or research team which is the total number of links in the indexed articles on reviewed publication. Citation depends not only on the level of scientific results but also on other factors, for example, the publication timeliness.

Scientometrics databases (SMDB) that used to obtain these estimates are bibliographic and abstract databases with tools for citation tracking of articles published in scientific journals.

Scopus of Publishing Corporation Elsevier is one of the most well known SMDB . Scopus do not apply the concept of impact factor but it widely used Hirsch index. This database is available by subscription through the Web-based interface (http://www.scopus.com). Furthermore, authors of articles can not register to view their rating page http://www.scopus.com/search/form/ authorfreelookup.uri.

Web of Science (WoS) of Thomson Reuters is one more very popular SMDB . It contains links to the full text in the original sources and lists of bibliographic references that appear in each publication.

Index Copernicus (Poland) (http: //www.indexcopernicus.com) is an international SMDB which covers indexing, ranking and abstracting of journals and is a platform for scientific collaboration and joint research projects.

Google Scholar (http://scholar.google.com/) indexes the broadest spectrum of research papers represented on the Web. This SMDB processes the full text of scientific publications in all formats and disciplines. The main scientometric indicator that generates by this SMDB is Hirsch index (both general and for the last five years).

Now many national SMDB oriented on indexing and evaluation of publications in the languages other than English are developed. For example, Web-site "Ukrainian Science Citation Index" (http: //uincit.uran.ua) provides to rate the publication activity of individual scientists and scientific institutions of Ukraine by the scientometric indicators.

V. The Role of Ontologies for Competence Estimating of Spesialists

In addition to general professional level, it is necessary to assess the expertise of researchers in domain of particular project. Formal models of such domain can be represented by its ontology.

Ontological analysis is now the most common approach of domain knowledge representation that provides the analyses and comparison of the competencies of experts and developers in new research areas [10]. In addition, the availability of the domain ontology that is known and used by researchers usually indicates their deep knowledge in this domain (especially when it concerns to information technology).

At the same time with domain ontologies it is advisable to use a common ontology of research activities that enables unambiguously establish the terminology associated with the rating publications, scientific degrees and academic titles, types of organizations, etc. [11].

We have developed the following ontology oriented on the determining of the competencies of the research project authors that can be integrated with organizational ontologies of scientific institutions, Academy of Sciences, UDC classifier and other relevant knowledge bases (Figure 1).

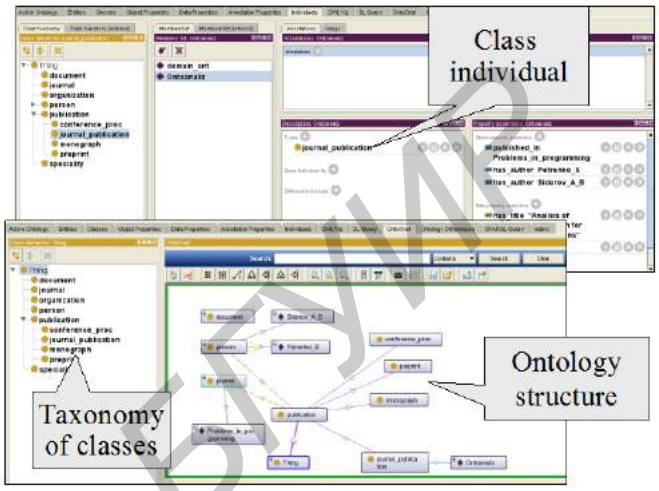

Fig. 1 – Ontology of research activities

This ontology contains such terms specific to research activities as «publication», «monograph», «research project», «diploma», «researcher», «specialty», «academic degree», «education» etc. and expresses such relationships between classes as "to be co-author of", "to work in the organization", "to be the author of the publication," "have a degree in the specialty" and properties such instances of classes like "to be publishing", "to have the Hirsch index ".

Developers of research projects can use this ontology as a model for the description and classification of the submitted documents which have to certify their competence level both in scientific work in general and in some particular research domain of information technology.

VI. Domain Thesauri and Ontologies as the Means of Matching of Natural Language Documents

We propose to generate the thesaurus of project and thesauri of IR that describes project participants: matching of these thesauri provides the evaluation of proximity of researchers qualification with project domain.

Thesaurus of the natural language IR can be considered as a projection of domain ontology [12]. Thesaurus of the project participants is generated as the join of the IR thesauri with account of weight of the individual IRs which should be considered as the significance of the document to describe the competence and level of trust to this IR. For example, the weight of thesis abstract is more then the weight of diploma, and impact factor of journal can define the weight of publication.

It is assumed that each of project developers generates a set of documents that are most pertinence to proposed research project.

For example, if the author has n scientific publications then he chooses m of them that are relevant with the project. However, the author should seek to ensure that all the concepts of the domain ontology that have linguistic equivalents in the project text of these would be present in his selected works (the weight of each comparison depends on the weight of IR due to the function of the status and rating of the document). Status of document characterizes the level of documentary evidence of this type of IR and rating of document characterizes its estimation in scientometric databases.

Project thesaurus Th_{proj} depends on the project description and of the selected domain ontology. It is a set of pairs (t_i, q_i) where $t_i \in T$, T is a set of terms of domain ontology $O_{domain} = \langle T, R, A \rangle$, and q_i is a number of matches that determines the weight of the term (if some term appears in the description of the project 10 times than it is assumed to be more important than that those one that appears only 2 times).

For each term of the domain ontology text fragments are retrieved.

The overall estimation of the team competence is determined by the array

$\left(t_i, \sum_{j=1}^{m} p_{IR_{j_i}} * v_{IR_j} \right)$, where

- t_i is a number of domain ontology terms from T;
- $p_{IR_{j_i}}$ is a number of matches with this term in the j-th IR;
- v_{IR_j} is a weight of the j-th IR.

It is important that this array does not contain all terms of domain ontology (domain in general can be much broader than it's part relevant to project), but only those ones that have matches with project.

We determine the weight of the j-th IR in such way: default weight of each document is w=1 and then it can grow under these conditions (evaluation can be summed):

- for scientific publications:

- If the article is published by journals with impact factor greater than 0.5 then w=w+5,

- If the article is published in the materials of the conference then w=w+1,

- If the article is published by the foreign edition then w=w+3,

- If the article is indexed in Google Scholar, then w=w+2,

- If the article indexed in Scopus then w=w+10;

- Passport of specialty or diploma w=w+5;
- Description of the organization profile w=w+3;
- Description of the earlier successfully fulfilled research project w=w+5;
- Description of the earlier proposed but nor realized project w=w+1;
- Abstract of a thesis w=w+4.

It is obvious that different articles have different weights for estimation of specialist competencies. Therefore we can take into account impact factor of journal and the year of publications (new publications are more important then the old ones). In this case, an overall estimation of the researcher is as follows:

$$C = \sum_{i=1}^{n} q_i * \left(\sum_{j=1}^{m} p_{IR_{j_i}} * v_{IR_j} * \mathrm{Im}\, p(IR_{j_i}) \right) \quad (1),$$

where $\mathrm{Im}\, p(IR_{j_i})$ is an impact factor of journal that publishes this IR.

In the future, it is advisable to enter various other normalized ratios that can reduce the impact of a large volume documents that are poorly saturated with the domain terms.

However, system development of these ratios requires much more detailed study of the hierarchy and content of documents submitted for examination by the authors and in a great measure depends on the project specifics. For example, different conditions are applied for young scientists and monograph reviewers.

Estimation (1) does not use domain semantics and relations among the domain ontology terms. Therefore we proposed to use the following more complex estimation:

$$C = \sum_{i=1}^{n} q_i * \left(\sum_{j=1}^{m} p_{IR_{j_i}} * v_{IR_j} \right) * s_i \quad (2),$$

where parameter s_i determines the value of the i-th term of the ontology by the number of its relations with those terms of ontology that are also included to the project thesaurus and take into account the semantic distance between them.

General qualifications of each of the developers of the research project can be took into account (in addition to the domain specialization) by their rating derived from SMDB. In particular, we propose to use information from Google Scholar and Scopus because this information is accessible for all the Web users.

In addition, it gives an opportunity to differentiate qualified individual researchers and just do not summarize their results. The following estimation of x-th researcher uses knowledge from external SMDBs:

$$C_x = \sum_{i=1}^{n} q_i * \left(\sum_{j=1}^{m} p_{IR-x\,j_i} * v_{IR-x\,j} \right) * s_i * h_x \quad (3),$$

where h_x is a sum of Hirsch index of researcher from Google Scholar and Scopus. If other SMDBs are available then their Hirsch ratings can be sum up too.

The general estimation of the qualifications of the project developers can be measured as the sum of estimation of the participants or their normalized sum. The first approach is advantageous because knowledge and experience of each can be used independently of the number of participants. Therefore, the

normalized estimation can be used only as an complementary or for teams with greatly different number of members.

The results of this objective competence estimation of scientists for new research domains characterized by high dynamics of innovation and technology were used in the preparation of a request for funding for a new project on the EU program Horizon 2020. Project Title: «Novel scalable E-call platform based on an intelligent ontological system - NEMO». The main idea of this project consists in creation of intelligent system for emergency medical care and for people with the risk to their lives.

One of the problems that exist in such systems (for example, warning system 112) is break or termination of telephone communication with affected person due to various reasons and the impossibility of obtaining comprehensive data about this person to send him the special care services.

The ontology-based approach that uses intelligent software agents allows to identify the missing information about affected people from the distributed network and helps to redirect this information to relevant support services in the shortest possible time and in appropriated form.

The definition of this project is semantically marked up by means of the Semantic MediaWiki (fig.2).

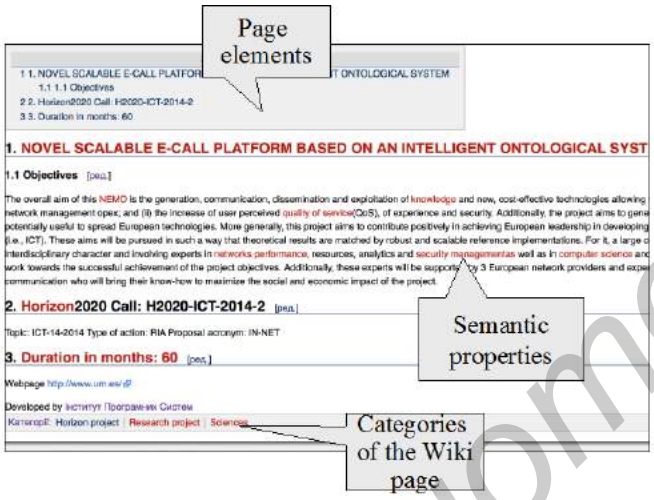

Fig. 2 – Structured description of the project

Project ontology can be built automatically by this structured description by means of the Semantic MediaWiki and be represented by OWL language [13, 14]. This ontology can be processed by Protégé [15] for knowledge visualization, extraction of some data and for integration with other external domain ontologies (fig.3).

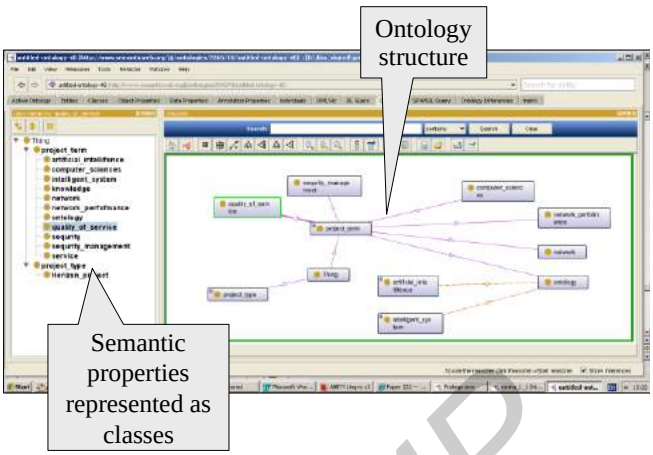

Fig. 3 – Ontology of the project domain

Then the terms of this ontology are used for semantic markup of the documents such as articles, certificates, diplomas that describe the competencies of potential project members. For example, semantic markup of the Semantic MediaWiki uses constructions that provide selection of different types of semantic properties for every Wiki page: [[semantic_property_name::semantic_property_value]], e.g. [[project term ::quality of service]] (fig.4). These elements are added to the document content at the places that are relevant to meanings of these terms.

The markup process can be performed manually or by specialized software instruments that help user in retrieval of relevant text fragments [16]. Unfortunately these tools depend hardly from the natural language of document and have to be developed for every language that is used [17].

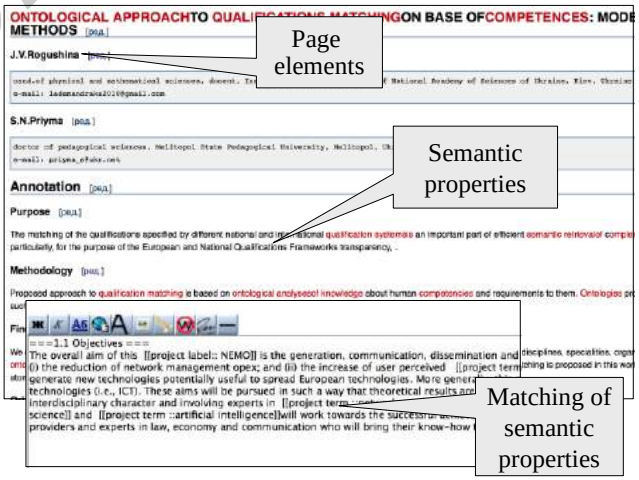

Fig. 4 – Structured description of the project

On the basis of the described above method objective competence assessment of experts and developers from different European countries was performed in order to form a consortium of developers of the different countries and institutions. Project Consortium members: Ukraine (Institute for Software Systems and the International Center of NASU), Germany (Technical University of Darmstadt), Spain (University of Murcia) and France (Paris Polytechnic School). NEMO project ontology was created

on the basis of preliminary project description, formalization of requirements and tasks to it (WPs) need to be addressed when creating the system. This ontology has been used for creation of subteam of experts and developers for the project. Currently this joint project aimed at developing a new intelligent service support for people in a situations that threaten their lives is submitted to the EU Commission Horizon 2020 and passed the first restrictive selection committee.

VII. Conclusions and prospects for further research

A new approach to solving the problem of objective competence evaluation in the context of new information and communication technologies has a lot of important specific features. These specifics is caused by which by high dynamics and heterogeneity of the Web information resources that demand semantic processing .

The approach proposed in this article is based on use of ontological knowledge: it provides matching of project thesaurus and thesauri of project participants which are based on the same domain ontology to evaluate the semantic distance between competencies of researchers and competence needs of project. Thesauri are generated by processing of the natural-language project description and information about participants (their publications, diplomas, descriptions of previous research projects, information about their organizations, etc.).

The goal of this research work is a development of the objective methods of qualification evaluation of potential project participants from the viewpoint of project domain competencies. These methods are oriented on taking into account a significant number of knowledge available by the Web. We think that these methods will we helpful for planning of research teams for various scientific research tasks and provide more efficient and high-quality research results.

References

[1] J. Rogushina, A. Gladun “Ontology-based competency analysis in new research domains”, in Journal of Computing and Information Technology. V.23, N. 4, 2012. – pp.123-134.

[2] Y.V.Granovsky “Is it possible to measure science? V.V. Nalimov's research in scientometrics", in Scientometrics 52.2, 2001, pp.127-150.

[3] P.Vinkler "Relations of relative scientometric impact indicators. The relative publication strategy index", in Scientometrics 40.1, 1997, pp.163-169.

[4] A.Van Raan "Comparison of the Hirsch-index with standard bibliometric indicators and with peer judgment for 147 chemistry research groups", in Scientometrics, 67.3, 2006, pp. 491-502.

[5] L.Bornmann, H.-D.Daniel "The state of h index research", in EMBO reports 10.1, 2009, pp.2-6.

[6] J. E. Hirsch “An index to quantify an individual's scientific research output”, in Proc. of the National academy of Sciences of the United States of America, 2005, 102(46), pp.16569-16572.

[7] L.Egghe "The Hirsch index and related impact measures", in Annual review of information science and technology 44.1, 2010, pp.65-114.

[8] M.Bordons, M.Fernández, I.Gómez. "Advantages and limitations in the use of impact factor measures for the assessment of research performance", in Scientometrics 53.2, 2002, pp.195-206.

[9] E.Garfield "The history and meaning of the journal impact factor", in Jama 295.1, 2006, pp.90-93.

[10] T.R.Gruber “Towards Principles for the Design of Ontologies Used for Knowledge Sharing”, In Inter. Journal of Human-Computer Studies, № 43 (5/6), 1994, pp.907-928.

[11] J.Rogushina, A.Gladun “Ontology-based competency analyses in new research domains”, in Journal of Computing and Information Technology. V.20, N. 4, 2012, pp..277-293.

[12] A.Gladun, J.Rogushina “Formalization of Search Context on Base of Ontologies and Multilinguistic Thesauruses”, in Computing, V.6 (3), 2007, pp.16-22.

[13] “OWL Web Ontology Language Semantics and Abstract Syntax. Section 2. Abstract Syntax”. – http://www.w3. org/ TR/owl-semantics/syntax.html.

[14] “A Practical Guide To Building OWL Ontologies Using Protégé 4 and CO-ODE Tools”. Edition 1.2., 2009. -- http://phd.jabenitez.com/wp-content/uploads/2014/03/A-Practical-Guide-To-Building-OWL-Ontologies-Using-Protege-4.pdf.

[15] “Protégé. W3C Semantic Web”. – http://www.w3.org/2001/sw/wiki/Protégé.

[16] A.Gladun, J.Rogushina “Use of Semantic Web Technologies and Multilinguistic Thesauri for Knowledge-Based Access to Biomedical Resources”, in International Journal of Intelligent Systems and Applications, 2012, №1, pp.11-20. – http://www.mecs-press.org/ijisa/ijisa-v4-n1/IJISA-V4-N1-2.pdf.

[17] J. Rogushina “Methods and tools of knowledge management at the Semantic Web environment”, in International Journal «Information Theories and Applications», V.19, N.3, 2012, pp.258-268. – http://www.foibg.com/ijita/vol19/ijita19-3-p08.pdf

ОНТОЛОГИЧЕСКИЙ ПОДХОД К СОПОСТАВЛЕНИЮ КОМПЕТЕНЦИЙ ПРЕДМЕТНОЙ ОБЛАСТИ ДЛЯ СПЕЦИАЛИСТОВ В НАУЧНО-ИССЛЕДОВАТЕЛЬСКИХ ПРОЕКТАХ

Рогушина Ю.В. , Гладун А.Я.

Предложены объективные методы сопоставления компетенций для разработчиков научно-исследовательских проектов. Эти методы базируются на семантическом сравнении описания проекта с теми документами, которые характеризуют компетенции исследователей в той предметной области, к которой относится проект. Создана специализированная онтология научно-исследовательской деятельности, предназначенная для унифицированного описания терминологии, связанной с вопросами квалификации. Предполагается, что основным источником сведений о компетенция исследователях являются их научные пуликации, представленные в Web, а также их рейтинги в наукометрических базах данных.

Предлагается также извлекать онтологические знания из таких структурированных и естественно-языковых информационных ресурсов, доступных через открытую среду Web, как Wiki-ресурсы, базы данных и знаний, персональные блоги, официальные Web-сайтов учебнх и научных организаций, а также из метаданных и онтологий.