Position papers of the 2014 Federated Conference on Computer Science and Information Systems pp. 271–277
DOI: 10.15439/2014F462
ACSIS, Vol. 3

Information Models and Methods of the University's Scientific Knowledge Life Cycle Support

Zhomartkyzy Gulnaz
D. Serikbayev East Kazakhstan State Technical University,
69 Protozanov A.K.
Ust-Kamenogorsk, Kazakhstan
Email: zhomartkyzyg@gmail.com

Balova Tatiana
D. Serikbayev East Kazakhstan, State Technical University,
69 Protozanov A.K.
Ust-Kamenogorsk, Kazakhstan
Email: tbalova@ektu.kz

Milosz Marek
Lublin University of Technology,
36b Nadbystrzycka,
Lublin, Poland
Email: m.milosz@pollub.pl

***Abstract*—The main aim of this work is to develop methods and technologies of the university's scientific knowledge management. The paper examines the concept of knowledge management and life-cycle processes of the university's scientific knowledge. Text Mining and Semantic Web technologies are used to develop the ontological information model and for information resources processing. The paper describes the developed information model of the university's scientific knowledge, the methods of forming scientific profiles, and the concept of the university's scientific knowledge semantic portal.**

I. Introduction

KNOWLEDGE , intellectual property, and intellectual resources have been understood in recent decades as a major driving force of the economy of the "third wave" (the new economy), the economy based on knowledge of expanded reproduction [1]. Knowledge as intellectual capital is gradually becoming one of the most important factors in the development of economy and the society.

For a modern institution of higher education as an open social and economic self-organizing system the processes of creation, accumulation, and dissemination of knowledge is becoming a key factor for training competitive specialists.

Paraphrasing [2], we can determine that the university's scientific knowledge as the combination of data and information with added opinions, skills and experience of the university's scientists and faculty members is a valuable asset that can provide an advantage in the market of educational services in project activities, scientific and practical work, and innovative activities.

The intellectual capital or intangible assets of the university are the source of new scientific knowledge. Edvinsson developed a hierarchical structure of intellectual capital of a higher education institution - a "Skandia Value Scheme" model [3]. The main components of this model are human assets (the capital) and innovating capital which are tacit knowledge.

Tacit knowledge form the human capital which is embodied in the university staff as a body of knowledge, qualifications, each employee's innovation, as the system of values, culture and philosophy of the institution. It is believed that practical tacit knowledge is the key to decision making and management.

There is a continuous exchange between explicit and implicit knowledge and their transformation. [4]. Nonaka I., Takeuchi H. suggested a cyclical process of knowledge transformation: socialization, externalization, combination, and internalization.

Kantner defines the concept of knowledge management as a strategy for the organization and the process of knowledge transformation [5], [6]. Currently, there is the increasing number of research papers devoted to the issue of the community of practitioners' knowledge management and the implementation of individual processes of knowledge transformation. [7], [8].

The European concept of knowledge management [2] identifies five processes of knowledge life cycle: knowledge identification, creation, storage, distribution and use. Knowledge life cycle actually reflects the methods and technologies of knowledge management at each technological step.

The purpose of the university's knowledge management is to improve human and innovative capital, to accelerate its development and competitiveness in the market of educational services in research, practical and innovative activities.

The effective use of the university's intellectual capital directly depends on the support of possibilities for knowledge creation, storage, dissemination and use. [9], [10].

The importance of developing knowledge management systems (KMS) is due to the fact that the knowledge which is disseminated, acquired and exchanged generates new knowledge [11].

In this regard, it is important to determine what knowledge management system must be created and what transformations must be implemented to use the existing intellectual capital successfully. There arises a need in processes, infrastructure and organizational procedures at a higher education institution that would allow its employees to use its corporate knowledge base. This paper examines some models and methods of the university's scientific knowledge life cycle support.

II. The University's Scientific Knowledge Management System

Knowledge management in an enterprise is the systematic process of identification, use, and transfer of information and knowledge which people can create, improve and apply [12].

© 2014, PTI

It is the process by which the enterprise generates knowledge, accumulates and uses it to gain a competitive advantage [13, 14].

In this paper the university's scientific knowledge management system (SKMS) is considered as an aggregate of information, software, technical means, and organizational solutions aimed at efficient management of the university's available intellectual resources and training specialists who meet the modern requirements. The purpose of SKMS at the university is the formation of a unique ontology-based integrated intellectual environment to improve the competitiveness of the university's science and education. The university's SKMS is the technological component of the university's SKM, which provides the creation, organization and dissemination of scientific knowledge among the university's staff.

The main functions of scientific knowledge management at the university (SKM) are consistent with the university's functions defined in [15]. They are divided into analytical, integration and new knowledge generation.

The analytic functions of SKM include:

- the search of knowledge in the information flow, content filtering;
- identification and classification of existing knowledge according to certain criteria, the formation of the staff's scientific profile, and the monitoring of the university's of scientific schools development.

The SKM integrative function provides:

- the introduction of classified knowledge in the corporate memory and evaluation of its integration with educational programs implementation;
- the extraction of knowledge from the corporate memory by sharing the knowledge between departments, different levels of management, as well as the exchange of expertise and experience of the staff;
- ensuring the accessibility of knowledge for management decisions making, search for ideas, generating ideas, and training.

The function of new knowledge creation provides the fixation of explicit and tacit knowledge in the university's scientific knowledge base.

The life cycle of scientific knowledge is shown in Figure 1.

The implementation of SKMS or its components at the university will allow users:

- to receive the right information at the right place at the right format and in a timely manner with minimal effort;
- to reduce the number of human errors and to increase the quality of decisions;
- to improve communication, reduce the information loss and distortion;
- to stimulate the sharing of knowledge and best practices.

The process-oriented On-To-Knowledge methodology [9] is used as the basis for the university's knowledge management.

There are following approaches to knowledge management: organizational, technological and ecological [16]. The technological approach puts the application of IT-technologies in line

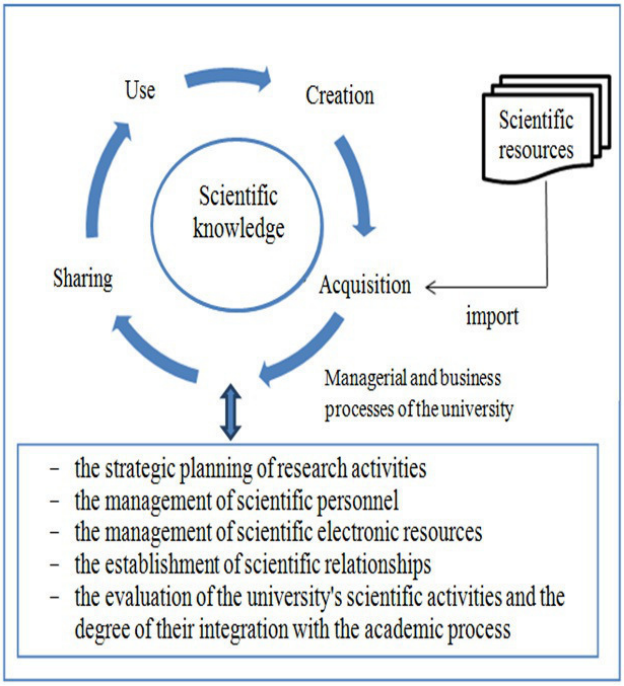

Fig. 1. The life cycle of scientific knowledge

with the organizational measures. The model of technological approach to knowledge management is shown in Figure 2.

The introduction of knowledge management and its related processes of developing and maintaining the SKM at an enterprise usually involve working with unstructured information resources.

The accumulation of knowledge is a complex process involving different work in the cycle of knowledge transformation [5]. The stage of knowledge acquisition includes:

- the knowledge acquisition by analyzing documents (Text Mining) and databases (Data Mining);
- metadata annotating / creation;
- the extraction of the employees' tacit knowledge;
- structuring / classification;
- the formation of organizational memory, knowledge integration and storing.

The use of knowledge suggests that the available knowledge is used by the university staff to perform their jobs more efficiently, and newly created knowledge affects both scientific and educational activities.

The university's SKMS integrates intellectual resources, knowledge management tools and processes of knowledge transformation. The general structure of the university's SKMS is shown in Figure 3.

"The university's intellectual resources" component in the overall structure of SKM determines human and structural resources with their related formalized human capital and innovative capital of the university. The sources of scientific knowledge resources are the electronic version of the university's scientific journal, the electronic version of the conference

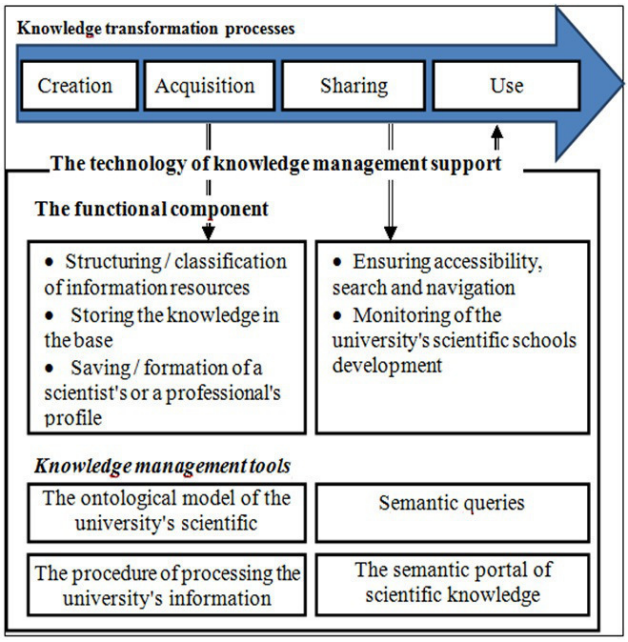

Fig. 2. The model of technological approach to knowledge management

proceedings, a bibliographic database "Irbis".

Knowledge processing and transformation consists of the following processes:

- creation: defining of tacit knowledge (which includes the university staff's knowledge) and of explicit knowledge (in the form of paper or electronic documents or records);
- accumulation, which includes: the ontology as a conceptual framework for describing knowledge resources and a set of methods for the formation of the university's scientific knowledge base;

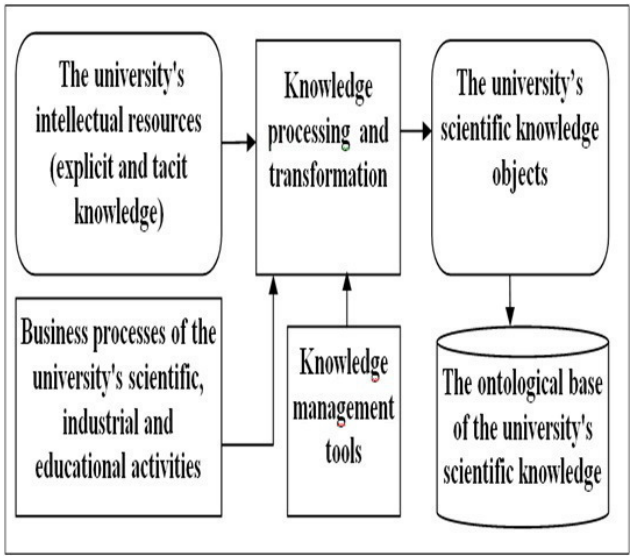

Fig. 3. The general structure of the university's SKMS

- sharing: providing users with opportunities for semantic search and navigation;
- tacit knowledge sharing which is intended for communication in scientific community and new knowledge formation. The participants of tacit knowledge sharing can be individual employees and individual research groups;
- using knowledge to improve the efficiency of research activities at the university.

Business processes at the university include:

- business processes of development: the development of innovative learning technologies; participation in research grants; improving academic qualifications of employees, establishing scientific relationships with companies and enterprises;
- the processes of postgraduate education: library resources management, electronic information resources management;
- the university management processes: personnel management processes and tools for personnel development, strategic planning of the university's research activities organizing knowledge sharing in knowledge networks;
- management of research activities and scientific and production activities;

Knowledge management tools to support the processes of transformation of scientific knowledge are listed below:

- information technology;
- organizational and administrative mechanisms;
- corporate culture;
- technical infrastructure;
- legal aspects.

IT knowledge management tools consist of a set of information technology, providing targeted development and effective functioning of the processes of transformation and networks of scientific knowledge.

III. THE INFORMATION MODEL OF THE UNIVERSITY'S SCIENTIFIC ACTIVITIES MANAGEMENT

Ontology, as a common language in knowledge management, is a conceptual domain model as a system of concepts, their properties and relations [17].

The information model of the university's knowledge can be described as the ontology which includes the basic concepts of the university's scientific activities, such as organizational structure, subjects, the objects of scientific schools and research, information resources, other subdisciplines, etc. [18].

In the ontology of scientific activities in the "Research directions" class the subclasses correspond to the major research areas of the university.

The subclasses correspond to major research areas in the ontology of scientific activity in the "Research directions" class.

The ontology, as a common language in knowledge management, is a conceptual domain model as a system of

concepts, their properties and relations. The use of ontology in knowledge management system makes it possible:

- to integrate the information distributed in various document repositories, databases and knowledge;
- to generalize and systematize the available information, acting as a metamodel;
- to use the automated logical conclusion for better search results, acquiring new knowledge and analyzing information;
- to use more effective mechanisms to receive, visualize and search for knowledge.

The ontological information model of knowledge database supports semantic queries in SPARQL and SPARQL-DL.

Example: semantic query for researchers and information resources in scientific directions can be written as follows:
Example 1:
Person and (peopleHasPublicationIR some (PublHasDivis some TopicsSolidStatePhysics))
Example 2:
Article and (PublHasDivis some TopicsSolidStatePhysics)

Navigation in scientific knowledge and information resources is done by the use of semantic links between the classes of the ontology.

IV. THE PROCEDURE OF THE UNIVERSITY'S INFORMATION RESOURCES PROCESSING WITH THE PURPOSE TO FORM SCIENTIFIC PROFILES

A. The Stages of the University's Information Resources Processing

The main stages of the information resources processing are given below:

1) Extraction of terminological collocations. Pearson criterion is used to detect collocations [19];
2) Feature selection. Mmutual Information method is used as a method for evaluating the importance of terms (Mmutual Information) [20];
3) Classification of texts according to scientific areas. The method of k nearest neighbour (kNN) is used for text classification [20], [21];

Working with the text files in the corpus with the purpose to perform statistical calculations requires the following preliminary steps:

1) To pre-translate to .txt format the files of different formats (pdf, doc,. docx) in the corpus;
2) To delete all hyphenation beforehand;
3) To perform lemmatization of all the text files in the corpus, to delete all punctuation marks, to change all uppercase letters to lowercase letters.

The corpus of documents for processing has been compiled from articles in various fields published in the "Solid State Physics" journal. The founders of the journal are the Russian Academy of Sciences, the Department of General Physics and Astronomy of RAS, Ioffe Physical-Technical Institute of the Russian Academy of Sciences [22].

A detailed description of processing stages is given in the following sections.

B. Collocation Extraction and Feature Selection for the Classification of Scientific Texts

A collocation is regarded as a non-random combination of two or more lexical items common to most scientific texts in a particular scientific field. The set of terminological collocations generated by the specified collection of scientific texts describes a narrow subject area (topics and subtopics) of this collection.

For automatic extraction of terminology collocations from scientific texts a freely distributable Java-library LingPipe interface is used [19]. The array of obtained collocations is ranked in order of importance, where the sequence of lexical tokens is dependent. The significance of the collocations is calculated based on the collocation the Pearson independence statistics. The higher the value of the significance of the collocation, the less the likelihood that the sequence of tokens is independent.

The general scheme of the formation of a software dictionary is shown in Figure 4.

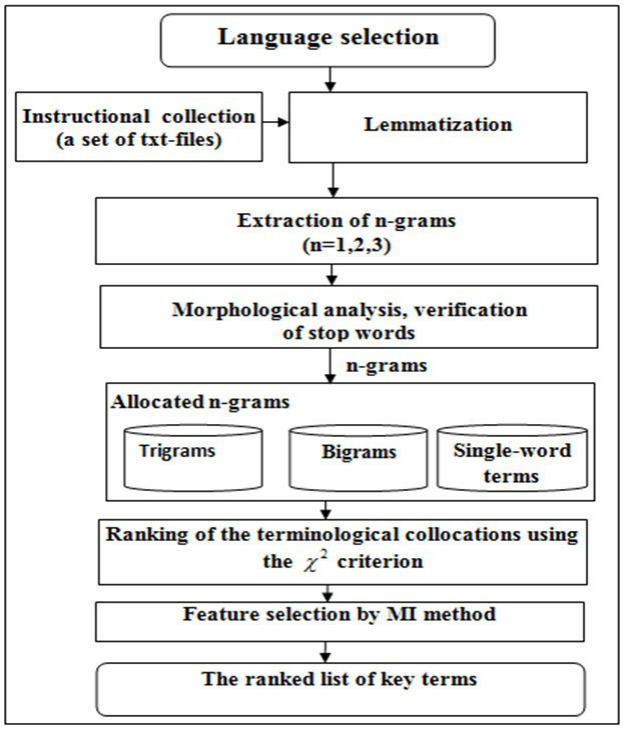

Fig. 4. The extraction and selection of domain terms

The main modification of the method based on the static approach includes the preliminary use of morphological templates of filters [23], [24].

To obtain the list of dominant terms using the χ^2 it is necessary to solve the following tasks:

- the extraction of collocations with the calculated coefficient of significance;
- the determination of the morphological characteristics of each word in the n-gram;

- the removal of stop words and the selection of phrases that match the templates;
- the saving of collocations in a database table.

The following restrictions were set for a bigram and a trigram: the minimum frequency bigram equal to 10, the minimum frequency of trigrams equal to 15.

The thus obtained term-candidates form a list of n-grams (bigrams, trigrams).

Single-word terms are extracted based on a combination of frequency and the inverse document frequency of the term. The weight of a single-word term is calculated by the formula [20]:

$$Tf - Idf_{t,c} = tf_{t,c} \times log\frac{N}{df_t} \quad (1)$$

where $tf_{t,c}$ is term frequency in the collection of the c class; df_t is the number of documents in the collection of the c class which contain the term; N is the number of documents in the collection.

The generated list of terms with weights $Tf - Idf_{t,c}$ is ranked by a certain threshold value, a number of terms are selected which are further recorded in the database table.

Table 1 presents the results of the developed module for extraction and separation of terms according to domains.

Table 1 shows some uninformative words with high critical value of χ^2 (such as the final edition, viewpoints, etc.), as well as some informative words with a lower value χ^2 (a superconducting property, a superconducting parameter).

The further stage of the vocabulary formation is the selection of features to eliminate noise-terms. Feature selection enhances the effectiveness of training the classifier by reducing the size of the vocabulary and the classification accuracy. The measure of utility $A(t, c)$ of each term in the lexicon is calculated for each class, and N terms with the largest value of $A(t, c)$ are selected. All other terms are discarded and are not involved in the classification.

To remove non-informative terms the method of mutual information was chosen [20]. The measure of mutual information estimates how much information about a class in information-theoretic sense the term includes. The measure the usefulness $MI(t_k, c)$ is calculated, and k terms with the highest values of this measure are selected. To select k terms $t_1, ..., t_k$ for a given class, the following formula is used:

$$MI(t_k, c) = log_2\frac{A \times Q}{(A + C) \times (A + B)} \quad (2)$$

where: A is the number of documents which belong to category c and contain term t; B is the number of documents, which do not belong to category c and contain term t; C is the number of documents which belong to category c and do not contain term t; Q is the instructional the training set of documents.

The results of applying the mutual information method for the selection of features obtained in the previous step are shown in Table 1.

As illustrated in Table 1, some terms with low rates of χ^2 (superconducting parameter and superconducting properties) have a high value of MI. At this stage it is necessary to perform a selection of informative terms weighted by their mutual information MI, which are selected in the domain vocabulary and then used for the text classification.

TABLE I
THE COMPARISON OF MUTUAL INFORMATION VALUES AND χ^2 OF TERMS FOR THE "SOLID STATE PHYSICS" DOMAIN

Terms	**Critical value χ^2**	**Value MI**
final version	40462,68	0,093
point of view	23093,13	0,100
superconducting granule	15534,01	1,000
intercrystalline boundary	14328,13	1,000
the first turn	13659,12	0,212
high-temperature superconductor	11518,91	1,000
superconducting transition	6566,12	1,000
phase transformation	4703,25	1,000
the object of study	4413,52	0,415
volume ratio	3584,50	0,263
the discussion of results	3196,66	0,553
at present	3175,12	0,263
ternary alloy	2434,56	1,000
metallic conductivity	2288,77	1,000
mentioned above	2243,87	0,652
amorphous film	1910,57	1,000
maximum value	1832,25	0,049
the system of equations	744,95	0,000
superconducting state	665,31	1,000
electronic spectrum	460,36	1,000
superconducting property	256,72	1,000
superconducting parameter	144,78	1,000

C. The Formation of Scientific Profiles Based on Classifications of Information Resources in Scientific Domains

For the classification of scientific resources kNN- classification is used. The classification task in machine learning is a task to assign an object to one of the predefined classes based on its formal characteristics. kNN method (k method of nearest neighbor) is a vector classification model. kNN classifier assigns the document to the prevailing class of nearest neighbors (Figure 5), where k is the method parameter. The k parameter in kNN method is often selected on the basis of experience or knowledge about the classification task at hand.

As a result of the university's scientific resources processing the documents' profiles are formed [25]. A document profile is defined as the vector of all relevant topics of its ontology:

$$PD(d) = (R_1^d, ..., R_c^d) \quad (3)$$

where: R_c^d are relevant topics c of document d.

Accordingly, the academic profile of a staff member is defined as the profile of all his publications:

$$PD(a) = (R_1^{da}, ..., R_i^{da}) \quad (4)$$

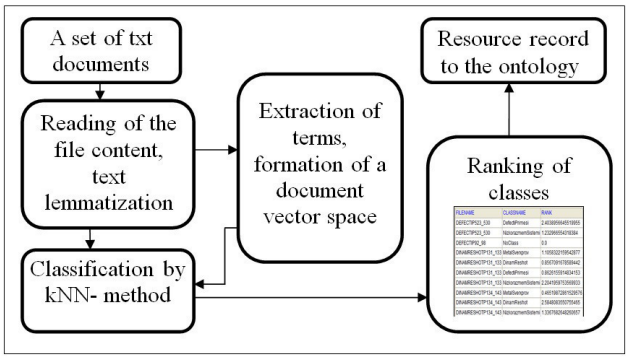

Fig. 5. The scheme of classification algorithm

where: R_i^{da} - are all the documents of the author. The final step of the text classification is the formation of the document's semantic profile by creating the individuals of "Information resources" class in the ontology of scientific research activity.

The classifier is written in Java, such tool sets as LingPipe, Apache Lucene (free Java library for text processing and high-speed full-text search) are used for further text processing. The classification results are $k-$nearest neighbors ranked classes, the parameter k is equal to 5.

V. THE CONCEPT OF THE SEMANTIC PORTAL OF THE UNIVERSITY'S SCIENTIFIC KNOWLEDGE

Previously considered models and methods formed the basis for the semantic portal of the university's scientific knowledge.

Operational intelligent query processing and adaptability to a user's needs is the basic idea of the functioning of the portal which is being developed. Therefore, there is a need to automate a time-consuming search and analysis of data in the process of creating and maintaining the university's scientific knowledge base.

The semantic portal platform architecture is shown in Figure 6.

To ensure the systematization of scientific knowledge and information resources the university's semantic portal supports the following functions:

- the software for navigation through the ontology of the university's scientific knowledge;
- the organization of search queries on the ontology concepts and relations;
- the classification of information resources to determine the development of the university's scientific schools and directions.

For faster data it was decided to load the ontology schema into an intermediate RDF-store of TDB built into Jena solution which supports all the features of work with JenaAPI, including preparation of SPARQL queries. The RDF-store contains the description of scientific knowledge ontology in the form of RDF-triples. This description corresponds to the graph which nodes are the subjects and objects of RDF-predicates, and the ribs are the RDF-predicate itself.

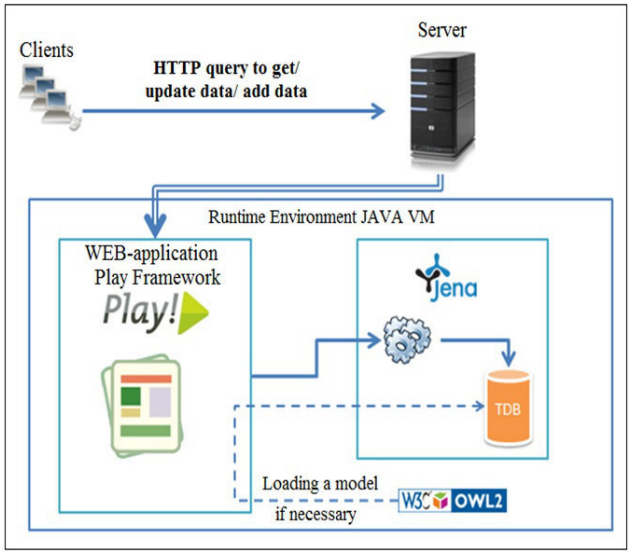

Fig. 6. The architecture of the semantic portal software platform of e-University's scientific knowledge

The main components of the semantic portal of the e-university's scientific knowledge are: the ontology of scientific knowledge, the ontology editor, the module of information resources classification and indexing, the module of navigation and search through the portal content, the database of ontological information.

The portal's architectural components provide the user with a transparent semantic access to the necessary data. User queries are handled by the server applications which are associated with the semantic components. Remote clients work with the portal in all modern browsers using HTTP protocol. Queries are sent to the web application server. In the JavaVM (Virtual Machine) run-time environment stream starts query processing.

When data is obtained using SPARQL or Jena API abstract model of data access, Jena TDB store (RDF-triples store) is accessed; the insert operations (delete, update) are performed, and the answer to the user is generated.

VI. CONCLUSION

This paper examines some models, methods, and technologies to support life cycle processes of the university's scientific knowledge. The ontology of the university's research activities is used as the information model of SKMS. The paper describes the procedure of the university's information resources processing. The thematic classification of documents based on the developed procedures for handling information resources allows forming of employees' scientific profiles and realizing a personalized search engine for the university's scientific knowledge semantic portal. The semantic concept of the university's scientific knowledge semantic portal is described. The software implementation of the semantic portal allows searching for any ontology object according to the following classes: researchers, research areas of the uni-

versity, events, key terms, organizations, departments, subdepartments, university publications. The pilot project of the university's scientific activity semantic portal has allowed us to generate a fragment of the university's scientific knowledge base.

The work was performed under grant "The development of an e-university's ontological knowledge base", state registration number 0213RK00305.

REFERENCES

[1] A. Toffler, *The Third Wave.* Moscow, 2002.
[2] European Guide to good Practice in Knowledge Management, Part 1:Knowledge Management Framework, 2004, http://enil.ceris.cnr.it/Basili/EnIL/gateway/europe/CEN_KM.htm
[3] L. Edvinsson, "Developing intellectual capital at Skandia," *Long Range Planning. J.,* vol. 30(1), 1997, pp. 366 - 373, http://dx.doi.org/10.1016/S0024-6301(97)90248-X
[4] I. Nonaka and H.Takeuchi, *Company - creator of knowledge. Origin and development of innovation in Japanese firms.* Moscow: Olimp-Business, 2003.
[5] H. Zaim, "Performance of Knowledge Management Practices: a causal analysis," *Knowledge Management. J.,* vol. 11, No.6, 2007, pp. 54-67, http://dx.doi.org/10.1108/13673270710832163
[6] J. Kantner, "Knowledge Management, Practically Speaking," *Information System Management. J.,* vol. 16(4), 1999, pp. 7-15, http://dx.doi.org/10.1201/1078/43189.16.4.19990901/31198.2
[7] Yuh-Jen Chen, Yuh-Min Chen, Meng-Sheng Wub, "An empirical knowledge management framework for professional virtual community in knowledge-intensive service industries," *Expert Systems with Applications. J.,* vol. 39, 2012, pp. 13135 - 13147, http://dx.doi.org/10.1016/j.eswa.2012.05.088
[8] K. Stock, T. Stojanovic, F. Reitsma, Yang Oue, M.Bishr, J.Ortmann, A. Robertson, "To ontologise or not to ontologise: An information model for a geospatial knowledge infrastructure," *Computers-Geosciences. J.,* vol. 45, 2012, pp. 98 - 108, http://dx.doi.org/10.1016/j.cageo.2011.10.021
[9] S. Staab, H-P. Schunurr, R. Studer, Y. Sure. "Knowledge processes and ontologies," *IEEE Intelligent Systems. J.,* vol. 16(1), 2001, pp. 26 - 34, http://dx.doi.org/10.1109/5254.912382
[10] D.V. Kudryavtsev, *Knowledge management systems and the use of ontologies.* St. Petersburg. Univ Polytechnic. University Press, 2010.
[11] R. Maier, *Knowledge Management Systems.* Information and Communication Technologies for Knowledge Management, Springer, 2007.
[12] K. Hafeez and H. Abdelmegid, "Dynamics of human resource and knowledge management," *Operational Research Society. J.,* vol. 54, 2003, pp. 153-164, http://www.jstor.org/stable/4101606
[13] T.A. Gavrilova, *Knowledge Engineering. In Innovative development: economy, intellectual resources, knowledge management.* Moscow: INFRA-M, 2009.
[14] M. Alavi and D. E. Leidner, "Knowledge management and knowledge management systems: Conceptual foundations and research issues," *MIS Quarterly. J.,* vol. 25(1), 2001, pp. 107 - 136, http://dx.doi.org/10.2307/3250961
[15] L.A. Trofimova and V.V. Trofimov, *Knowledge Management.* St. Petersburg, 2012.
[16] A.F. Tuzovskii. "Developing knowledge management systems based on a single ontological knowledge base," *In Bulletin of the Tomsk Polytechnic University. J.,* vol. 310(2), 2007, pp. 182 - 185.
[17] D. Allemang and J. Hendler, *Semantic Web for the Working Ontologist.* Burlington, USA, 2011.
[18] Y. Zagorulko. "Information model of scientific knowledge portal," *Information Technology. J.,* vol. 12, 2009, pp. 2 - 7.
[19] Alias LingPipe, http://alias-i.com/lingpipe.
[20] Ch.D. Manning, P. Raghavan and H. Schutze, *Introduction to Information Retrieval.* Cambridge University Press, 2009, http://nlp.stanford.edu/IR-book/
[21] H. Altncay and Z. Erenel, "Analytical evaluation of term weighting schemes for text categorization," *In Proceedings of the Pattern Recognition Letters,* 2010, pp. 1310 - 1323, http://dx.doi.org/10.1016/j.patrec.2010.03.012
[22] Science journal "Solid State Physics", http://journals.ioffe.ru/ftt/.
[23] Multiword Recognition and Extraction, http://www.ilc.cnr.it/EAGLES96/rep2/node38.html.
[24] D.S. Novikov. Automatic allocation of the terms of the texts subject areas and linkages between them. In Information and telecommunication technologies and mathematical modeling of high-tech systems in 2012. RUDN; Russia, http://conf.sci.pfu.edu.ru/index.php/ittmm/2012/paper/view/245l.
[25] K.V. Kryukov, O.P. Kuznetsov and V.S. Suhoverov, "On the notion of formal competence researchers," *In Proceedings of the III International Scientific and Technical Conference - OSTIS-2013,* Minsk, 2013, pp. 143-146, http://www.conf.ostis.net/index.php?title=OSTIS-2013