Semantic Portals in Biomedicine: Case Study

Irina V. Efimenko
Semantic Technologies Department,
Avicomp Services
84/2 Vernadsky Av.
Moscow, Russia 119606
Irina.Efimenko@avicomp.com

Sergey A. Minor
Semantic Technologies Department,
Avicomp Services
84/2 Vernadsky Av.
Moscow, Russia 119606
Sergey.Minor@avicomp.com

Anatoli S. Starostin
R&D Department,
Avicomp Services
84/2 Vernadsky Av.
Moscow, Russia 119606
Anatoli.Starostin @avicomp.com

Vladimir F. Khoroshevsky
Applied Intelligent Systems Department,
Institution of Russian Academy of Sciences Dorodnicyn Computing Centre of RAS
40 Vavilov Str., Moscow, Russia 119333
khor@ccas.ru

Abstract

Case studies for developing and implementing medical portals based on Semantic Technologies and ontological approach in Knowledge Management, Information Extraction and unstructured text processing are presented in the paper.

Keywords
Semantic Technologies, multi-lingual information extraction, medical content gathering, drug descriptions processing, RDF-storage, semantic Wiki, knowledge based analytics.

1. Introduction

Semantic Technologies and the Semantic Web (SW) as the embodiment of know-how for practical usage of these technologies are widely discussed, and it is already clear that semantic content available within knowledge portals shall lead us to a new generation of the Internet and knowledge intensive applications [1].

Medicine should be considered among the top domains for Semantic Web and intelligent applications due to high (and increasing) volumes of health-related information presented in both unstructured and machine readable form [2, 3, 4, 5].

The aim of this paper is to present one particular approach to this task – the Ontos solution for the Semantic Web in the medical domain. Two types of web applications (semantic portals) are examined, as well as the technology which underlies them.

2. Ontos Solution for Semantic Web

2.1 General Remarks

One of the main goals of the Semantic Web is "semantizing" the content which already exists within the classic WWW, and of the new content created each day. Significantly, the semantic representation of processed content should be suitable for usage by program agents oriented at solving customers' tasks. To support customers' activities within the Semantic Web we need common processing platforms with, at least, three key components:

- Knowledge Extractor based on powerful information extraction methods and tools (multilingual, effective and easily scalable).
- Knowledge Warehouse based on the RDF, OWL, SPARQL standards (effective and scalable).
- Set of customer oriented Semantic Services (Semantic Navigation, Semantic Digesting and Summarization, Knowledge-Based Analytics, etc.).

2.2 Ontos Solution: An Overview

2.2.1 Workflow Overview

Semantic Content within the Semantic Web framework can be viewed as a new kind of "raw material", which serves as input for Semantic Services that process it and present the results to customers [6, 7].

The Ontos Service Oriented Architecture (Ontos SOA) and an appropriate software platform were developed to support these aspects of Semantic Technologies within the Ontos solution.

The general workflow within the Ontos Solution is illustrated below.

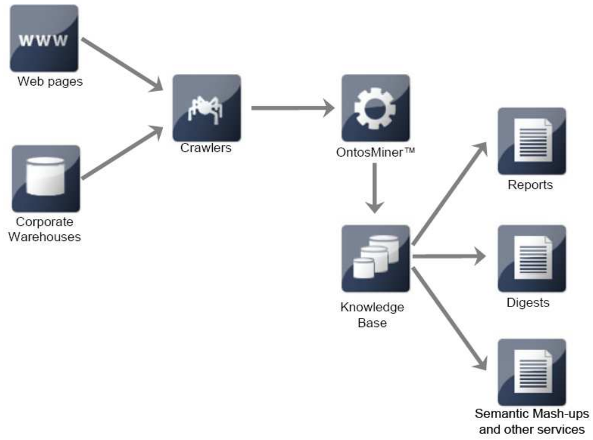


Figure 1. Workflow within the Ontos Solution

The crawler component gathers web-pages from a pre-defined list of resources, and transforms them into plain-text documents. These are then fed as input to the OntosMiner linguistic processors, which are discussed in sections 2.2.2 and 2.2.3. The output of these processors is a semantic graph (in RDF/XML, OWL, Turtle or N3 format) which represents named entities and relations recognized in the input text.

This graph is then stored in the knowledge base, where incoming knowledge is integrated with existing data (see section 2.2.4).

The data in the knowledge base is accessed by various web-oriented semantic applications, which were designed to provide end users with interesting and powerful services based on semantic metadata (see section 3).

2.2.2 Information Extraction with Systems of the OntosMiner Family

Generally speaking, each IE-system of the OntosMiner family takes as input a plain text written in a natural language and returns a set of annotations, which are themselves sets of feature-value correspondences. These output annotations represent the objects and relations which the processor was able to extract from the text.

The basic structure of OntosMiner linguistic processors is based on the well-known GATE architecture [8]. Each OntosMiner linguistic processor consists of a set of specialized modules called 'resources' which are organized into 'resource chains'. In this chain the resources are launched one after another and each subsequent resource has access to the output of previously launched resources.

The first three basic resources are the Tokenizer, the Morphological Analyzer, and the Gazetteer. The Tokenizer determines word boundaries based on formal features of the text, the Morphological Analyzer generates a set of possible variants of morphological analysis for each word, and the Gazetteer annotates key words and key phrases which are later used for the recognition of named entities and relations. These three modules - prepare the input for the two main modules of the system - the Named Entity Extractor and the Relation Extractor.

In the domain of named entity recognition we have adopted the rule-based approach to NLP which means that named entities are identified according to rules defined by developers. Thus, the Named Entity Extractor consists of a set of rules divided into subsets called 'phases' which are applied sequentially to the annotation set. Rules from each subsequent phase have access to the output of rules in previous phases. Each rule consists of a pattern on the annotation set and a sequence of commands which define the action that has to be performed when the pattern is encountered. The pattern is written in the Jape+ language, which is an extended version of the Jape language developed by the Natural Language Processing Group at the University of Sheffield [8]. The action parts of rules are mostly written in Java.

The list of possible keys for named entity recognition includes the key words and phrases annotated by the Gazetteer module, as well as annotations generated by previous phases of the Named Entity Extractor, and even specific features of annotations. For instance, the fact that a word begins with an upper case letter can play a significant role in the recognition of proper names in languages like English and French.

Typically, the system of rules for the recognition of a certain type of named entity comprises several dozens of rules which 'build' the target annotations through a number of intermediate steps.

One of the main difficulties with the rule based approach that we adopt is the emergence of conflicts between different rules. For instance, one set of rules within the Named Entity Extractor can identify a certain text fragment as part of a person's name, while a different set of rules identifies it as part of a company name. We discovered that when the number of rules involved grows beyond one hundred, it becomes increasingly difficult to try to control for such conflicts within the rule system itself. This is why in OntosMiner processors we allow the rules for named entity extraction to apply freely, but complement the Named Entity Extractor with a special module called Minimizer which defines the policies for conflict resolution. The idea is that different rules have a varying measure of reliability and that the developer can evaluate this measure for each rule, stating it as a feature of the annotation created by this rule.

Thus, annotations generated by the Named Entity Extractor come with a feature called 'Weight' which has an integer value ranging from 0 to 100. This feature reflects the probability (as estimated by the developer) that this annotation is correct. The Minimizer resource contains a set of rules which describe different types of conflict and define which annotations should survive and which should be deleted, based on the types of annotations involved in the conflict and their weights. The resulting 'minimized' annotation set is passed on to the Relation Extractor.

Semantic relations are certain facts or situations mentioned in the input text which relate one named entity to another, such as information about a person's employment in a company, or, in the medical domain, the fact that a certain condition can be the side-effect of a certain medicine. The module which is responsible for the recognition of semantic relations in OntosMiner processors is the Relation Extractor. Just like the Named Entity Extractor, the Relation Extractor contains a set of rules written in Jape+ and Java, grouped into a sequence of phases.

Recognition of semantic relations differs from the recognition of named entities in that named entities are expressed by compact word groups, while the keys for semantic relations can be situated quite far apart from each other within one sentence or within the whole text. This is why in developing rules for relation recognition we exploit a different strategy: we reduce the set of annotations which is fed as input to the rules, so that it includes only key words and phrases needed to identify a particular relation, and conversely, 'stop-words' and 'stop-phrases' which should never interfere between the keys. All other annotations are not included into the input and are not 'visible' to the rules.

Another method that we found to be sometimes very effective in relation extraction is to first divide the input texts into meaningful fragments, and then to process each type of fragment with a separate set of rules. This technique proves useful when we are dealing with semi-structured input texts, such as drug descriptions (see below on the MedTrust portal). We can use a different set of rules for each sub-section of the description, which leads to an improvement in both precision and recall.

A distinguished type of relation is the relation 'TheSame' (also called the 'identification relation') which is established between two co-referring occurrences of a named entity within a text. The resource which establishes relations of this type is called OntosCoreferencer. This resource builds a matrix of all the relevant annotations and compares them two by two to establish whether the annotations in each pair can count as two co-referring occurrences or not.

The final resource in the resource chain of every OntosMiner processor is the Triples Converter. This module takes as input the set of annotations created by previous modules and generates an output in the form of an RDF/XML, OWL, Turtle or N3 document. During its work the Triples Converter accesses the OntosMiner Domains Description database (see below) and replaces all the names of annotations generated by the OntosMiner processor with the names of corresponding concepts and relations of the Domain Ontology, using the mapping rules defined in the Mapping Ontology. All the OntosMiner annotations for which mapping rules have not been defined, are excluded from the output.

2.2.3 *Ontological Engineering*

It is well known that the ontological engineering is one of the core processes in the life cycle of semantic-oriented applications. Today there exists a number of methodologies, technologies and tools supporting this activity [9]. An overwhelming majority of them is oriented at creating and maintaining domain ontologies, and doesn't have anything in common with editing linguistic dictionaries or developing natural language processors.

However, on the conceptual level, configuring a linguistic processor or a system of linguistic dictionaries may also be viewed upon as a new domain, which in its turn may be modeled by an ontology or a system of ontologies. The system of ontologies which determines the work of OntosMiner processors is called OntosMiner Domains Description (OMDD). On the physical level OMDD is the data which is uploaded to an RDF based triplestore (OMDD database). Ontological data in the OMDD is stored in a format which is completely compatible with OWL.

Generally speaking, OMDD is a system of ontologies which can be divided into 6 classes:

- Domain ontologies (concepts and relations which are relevant for a certain domain). Domain ontologies are interconnected by relations of inheritance.
- Internal ontologies (sets of annotation types, features and possible feature values used in specific OntosMiner processors).
- Dictionary ontologies (morphological dictionaries and dictionaries of key words and phrases).
- Resource ontologies (sequences of text processing resources which are used by OntosMiner processors).
- Mapping ontologies (mappings which ensure that concepts from the internal ontology are correctly replaced with concepts from the domain ontology).
- Other (auxiliary) ontologies.

The current OMDD contains about 120 ontologies (around 2,5 million triples).

2.2.4 *Ontos Semantic Knowledge Base*

The Ontos Semantic Knowledge Base is one of the core components within the Ontos solution. Its main function is to provide effective storage of the RDF-graph which accumulates all the information extracted from large collections of texts by OntoMiner processors. Data in the Knowledge Base can be accessed via queries in the SPARQL query language.

At the moment, we have two implementation of the Knowledge Base – one based on RDMS Oracle 11g and another one based on Open Source libraries and platforms for the implementation of RDF-stores.

A crucial problem in this regard is the presence of duplicate objects (i.e. objects that represent the same real world entities) within the accumulated RDF graph. The task of merging such instances is performed by algorithms of object identification which take into account the whole set of an object's identifying features, including information about its attributes and relations.

3. Intelligent Applications for the Next Generation Web

The presented Ontos solution presumes two modes of access for external users: either the accumulated semantic content can be accessed via our own implemented semantic applications, or semantic content can be provided for use by third-party applications via an API.

Our own solutions [10, 11] based on semantic content include packages for Nanomedicine and Pharmacology.

3.1 Semantic Portal for Nanomedicine

The main goals of “semantizing” NL-content are related to integrating pieces of information, identifying implicit connections, and providing the possibility to receive an object's profile, to find out trends, etc. All these issues are particularly important for innovative fields such as Nanomedicine.

3.1.1 Information Sources and Domain Models

In order to make it possible to carry out a full-scale analysis of different aspects of any innovative field, one should integrate information from a variety of sources of different structure and content. The relevant information sources can include (but are not limited to) the following ones:

- Patent collections;
- Databases with descriptions of international projects and programmes;
- Conference materials and scientific papers;
- Blogs and forums in the specific domain;
- Regulatory documents;
- Opinion letters of analytical companies;
- Internet portals, news in technology, RSS feeds.

It is also worth mentioning that the most interesting data can be extracted from multilingual document collections, allowing users, above all, to form a view of the situation on an international scale.

The ontological system used for knowledge extraction in the Ontos Nanomedicine solution is based on a combination of ontologies corresponding to specific domains and information sources. This means that each particular ontology contains concepts and relations relevant for the domain and typical for the specific source (e.g. “Inventors” and “Assignees” for Patent analysis). The system of domain models which underlies the portal is presented below in Table 1.

Table 1. System of domain ontologies for Nanomedicine

№	Ontology	Description, Concepts, Relations
1	“Common”	“Basic” concepts and relations relevant for most of the ontologies in the considered domain. It can be viewed as an upper ontology specific for the domain of interest
2	Patents	Inventors, Inventions, Assignees, Agents, Key terms, Fields, etc.
3	Conferences	Events, Participants, Papers, Authors, Co-authors, etc.
4	News (specific for the field)	Mostly coinciding with the ones from the “Common” ontology; Sentiment
5	Projects	Projects, Investments, Programmes, ProgrammeTypes, etc.
6	Finance	Revenue, Shareholders, Producers, Customers, Stock information, Officers, etc.
7	Analytical research	Technology maturity, Producers, Customers, Competence, etc.

All the domain ontologies are language independent. This means that the NLP modules for any language relevant for the project are driven by the same ontologies. Language specificity is taken into consideration at the level of linguistic rules and linguistic (dictionary) ontologies.

3.1.2 Semantic Portal’s Functionalities

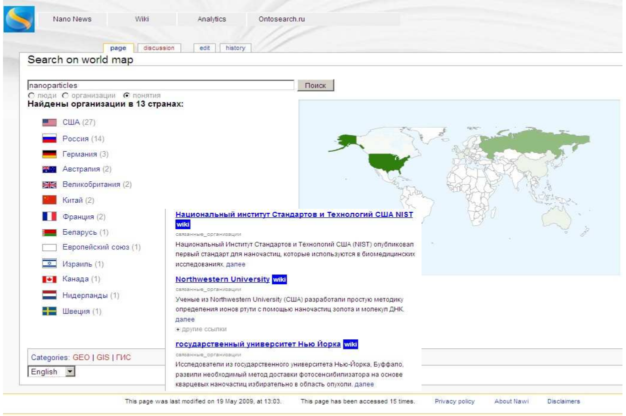
The portal includes the following sections:

News/Monitoring. This page is meant for on-line monitoring of the sources which are considered to be relevant. Objects and relations are extracted which makes it possible to form ratings, illustrate trends, and determine semantic focuses of the processed documents. A multilingual thesaurus is integrated into the page. Filtering by categories, sources, object types, etc. is provided.

Experts. Companies and Institutions. Shadow groups. For the most part, the content for these sections is related to patents, scientific papers, PhD theses, and conference materials. OntosMiner extracts information about inventors, authors and co-authors, assignees, affiliations, etc. This allows users to find experts and leaders in the domain of their interest, as well as to look for shadow groups of people and institutions working in the domain, based on thesauri and objects of interest.

Analytics. “My Objects” analysis. These sections provide BI tools for presenting a variety of views on the data stored in the Knowledge Base. Pie-charts, column diagrams, matrices help users to discover trends, areas of concentration of financial, intellectual and other resources, find out lacunae, etc. Own Ontos tools as well as third-party instruments can be used for presenting information in this section. “My Objects” functionality allows users to form personalized collections of objects, which are stored in user profiles, so that one can monitor their visibility in the media and their public image, compare their ratings, discover the most interesting statements about them, etc.

GIS. Graph Navigation. The GIS section is designed for representing objects and facts from the Knowledge Base on geographic maps (Semantic GIS). Graph Navigation gives access to all objects and relations in the Knowledge Base, allowing users to discover connections between objects starting from an object of interest, with the possibility to filter relations by type, relevance, etc. (Fig. 2).



Search on world map

nanoparticles

○ люди ○ организации ● понятия

Найдены организации в 13 странах:

- США (27)
- Россия (14)
- Германия (3)
- Австралия (2)
- Великобритания (2)
- Китай (2)
- Франция (2)
- Беларусь (1)
- Европейский союз (1)
- Израиль (1)
- Канада (1)
- Нидерланды (1)
- Швеция (1)

Национальный институт Стандартов и Технологий США NIST wiki

связанные_организации

Национальный Институт Стандартов и Технологий США (NIST) опубликовал первый стандарт для наночастиц, которые используются в биомедицинских исследованиях. далее

Northwestern University wiki

связанные_организации

Ученые из Northwestern University (США) разработали простую методику определения ионов ртути с помощью наночастиц золота и молекул ДНК. далее

+ другие ссылки

государственный университет Нью Йорка wiki

связанные_организации

Исследователи из государственного университета Нью-Йорка, Буффало, развили необходимый метод доставки фотосенсибилизатора на основе кварцевых наночастиц избирательно в область опухоли. далее

Categories: GEO | GIS | ГИС

This page was last modified on 19 May 2009, at 13:03. This page has been accessed 15 times. Privacy policy About Nawi Disclaimers

Figure 2. Widgets within the Ontos Semantic Applications

3.1.3 Semantic Wiki, Bookmarking and Navigation

The portal is based on the wiki-engine, since one of its purposes is to create an environment for the community of experts. This functionality is in harmony with the Semantic Wiki approach. Initially the content of wiki-pages is generated automatically based on the accumulated semantic metadata. Later, these data can be supplemented manually by users in the standard wiki fashion by experts with sufficient access rights. Semantic bookmarking tools are also integrated into these wiki-pages [12].

Inside the original information sources (i.e. web pages), one can switch on the so-called Semantic Navigation option by installing a special plug-in which superimposes semantic metadata upon the original content. Superficially, this looks similar to standard hypertext, but the functionality is different. Once the user clicks on a highlighted object a navigation card appears, which delivers accumulated information on the object's features and relations, and provides the possibility to navigate through the semantic graph starting from this object.

3.2 Semantic Portal MedTrust

Another application of Ontos solutions in biomedical domain called MedTrust system was designed for non-professional users looking for impartial information in pharmacology, as well as for physicians prescribing drugs to their patients. This application isn't claimed to be an expert system and is not aimed at giving ready-to-use recommendations. The aim of this application is to provide users with information integrated from a number of trusted sources containing pharmacological data thus giving them opportunity to receive full and detailed information related to their health conditions and prescriptions in one place. This is especially important when a patient has several health problems and is treated by several physicians, each of them prescribing medications according to his or her specialization. The aim of the system is to make users pay attention to possible incompatibilities, contra-indications and side effects taking place as a result of taking several type of medicine at the same time, or taking a certain medicine when having a certain health condition.

3.2.1 Domain Model

The domain model for the MedTrust system is focused on the concept *Medicine (drug, preparation).* It was initially prepared by professional physicians from the Russian State Medical University (RSMU).

The next task was to transform this model into an ontology which would conform to the standards adopted within the Ontos solution.

The main sections within drug descriptions are mostly common for different pharmacological resources and include the following data: Latin name, Composition and Form, Pharmacological Action, Pharmacokinetics, Indications, Pregnancy and lactation, Contraindications, Side effects, Special Notes, Medical Interaction, Overdosage, Doses, Storage conditions, Expiration date, Registration number, Analogues, Active ingredients, Manufacturer, Pharmacological groups, ATC classification, Therapeutic class.

This information is represented either in the form of attributes or in the form of relations in the domain ontology. Types of objects include Preparation, State/Condition, Symptom, Syndrome, Treatment Method, etc.

3.2.2 Information Extraction

A special OntosMiner processor based on this domain ontology was applied to over 10000 texts, including both drug descriptions and unstructured NL-texts about symptoms, syndromes and diseases. The results were accumulated in a knowledge base, which is then accessed via the user interface.

3.2.3 User Interfaces

User interface is implemented as a portal with a variety of sections including a pharmacological guide. There are two types of semantic services integrated into the portal, one is related to semantic navigation (see section 3.1.3), another one is provided in a form of a query interface. In the latter interface, there are several predefined query types, for instance "Which preparations are contra-indicated in case of listed health states/conditions?", and "Are the listed preparations compatible?".

For obtaining complete and detailed information on preparations, one can use search by active ingredients. One can also use search by states' synonyms. Different search methods can be used individually or simultaneously.

The navigation service gives more information on selected preparations and conditions. The provided information is organized according to the domain model. However, the navigation service is not so strictly focused on the specific user needs as the query service, and is thus more suitable for surfing the knowledge base (Fig. 3).

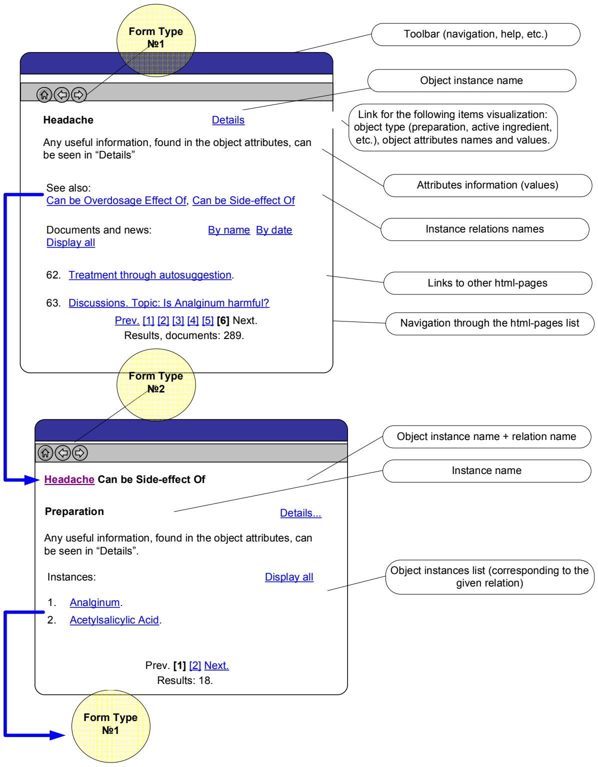

Form Type №1
Toolbar (navigation, help, etc.)
Object instance name
Headache
Details
Link for the following items visualization: object type (preparation, active ingredient, etc.), object attributes names and values.
Any useful information, found in the object attributes, can be seen in "Details"
Attributes information (values)
See also:
Can be Overdosage Effect Of, Can be Side-effect Of
Instance relations names
Documents and news:
By name By date
Display all
62. Treatment through autosuggestion.
Links to other html-pages
63. Discussions. Topic: Is Analginum harmful?
Prev. [1] [2] [3] [4] [5] [6] Next.
Navigation through the html-pages list
Results, documents: 289.
Form Type №2
Object instance name + relation name
Headache Can be Side-effect Of
Instance name
Preparation
Details...
Any useful information, found in the object attributes, can be seen in "Details".
Instances:
Display all
Object instances list (corresponding to the given relation)
1. Analginum.
2. Acetylsalicylic Acid.
Prev. [1] [2] Next.
Results: 18.
Form Type №1

Figure 3. An example of the navigation service interface

4. Conclusion

In this paper we have presented the Ontos solution for the Semantic Web in Nanomedicine and Pharmacology domains. This solution is based on automated processing of large multilingual collections of natural language texts, gathered from Internet resources and corporate databases. This process is controlled by ontological representations of the domains of interest. The output of this analysis is represented in standard formats and stored in an RDF Knowledge Base, where data is merged and accumulated. Finally, we described Semantic Web Applications which are based on this accumulated semantic content.

5. Acknowledgements

We would like to say many thanks to our CEO Victor Klintsov, our Product Director Oleg Ena, to Grigory Drobyazko, Alexander Ren and their teams, and to all members of the OntosMiner team. Particular thanks are due to Prof. Tatiana V. Zarubina and her team who are our wise guides through the tricky science of Medicine.

6. References

[1] V. R. Benjamins, J. Contreras, O. Corcho and A. Gomez-Perez. Six Challenges for the Semantic Web, http://www.cs.man.ac.uk/~ocorcho/documents/KRR2002WS_BenjaminsEtAl.pdf, 2002.

[2] Medline Plus. http://medlineplus.gov/, 2009.

[3] PubMed. http://www.ncbi.nlm.nih.gov/pubmed/, 2009.

[4] D. Gans, J. Kralewski, T. Hammons and B. Dowd. Medical groups' adoption of electronic health records and information systems. *Health affairs (Project Hope)* **24** (5): 1323–1333, 2005.

[5] Health Related Web Resources. http://www.cdph.ca.gov/PROGRAMS/CANCERDETECTION/Pages/HealthRelatedWebResources.aspx, 2009.

[6] I. Efimenko, G. Drobyazko, P. Kananykina, V. Khoroshevsky, et. al.: Ontos Solutions for Semantic Web: Text Mining, Navigation and Analytics. *In Proceedings of the Second International Workshop "Autonomous Intelligent Systems: Agents and Data Mining" (AIS-ADM-07)*. St. Petersburg, Russia, June 3-5, 2007.

[7] V. Khoroshevsky,. Knowledge Spaces in Internet and Semantic Web (Part 1), *Artificial Intelligence & Decision Support, N 1 2008*, p.p. 80-97 (In Russian).

[8] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan. GATE: an Architecture for Development of Robust HLT Applications. *In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, July 2002

[9] A. De Nicola, M. Missikoff and R. Navigli. A Software Engineering Approach to Ontology Building. Information Systems, 34(2), Elsevier, 2009, pp. 258-275.

[10] I. Efimenko, D. Hladky, V. Khoroshevsky and V. Klintsov. Semantic Technologies and Information Integration: Semantic Wine in Media Wine-skin, *In Proceedings of the 2nd European Semantic Technology Conference (ESTC2008)*, Vienna, 2008.

[11] D. Hladky. Ontology Based Text Mining in Temporally Structured Digital Texts. *Proceedings of Semantic Technology Conference 2009*, San Jose, California, 2009.

[12] P. Dudchuk and S. Minor. In Search of Tags Lost: Combining Social Bookmarking and SemWeb Technologies, http://www.semanticuniverse.com/articles-search-tags-lost-combining-social-bookmarking-and-semweb-technologies.html, 2009