I.J. Intelligent Systems and Applications, **2012, 1, 11-20**
Published Online February 2012 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijisa.2012.01.02

Use of Semantic Web Technologies and Multilinguistic Thesauri for Knowledge-Based Access to Biomedical Resources

Anatoly Gladun
International Research and Training Centre of Information Technologies and Systems,
National Academy of Sciences and Ministry of Education of Ukraine, Kyiv, Ukraine
E-mail: glanat@yahoo.com

Julia Rogushina
Institute of Software Systems, National Academy of Sciences of Ukraine,
Kyiv, Ukraine
E-mail: jjj_@ukr.net

***Abstract* — For more relevant informational retrieval and matching of user request with metadata about biomedical informational recourses it is necessary to formulize the user knowledge about this subject domain. We propose to use the ontologies and associated with them thesauri of the appropriate subject domains for representation of biomedicine knowledge. The algorithms of formation and normalization of the multilinguistic thesauruses, and also methods of their comparison are given in this work.**

***Index Terms*— ontology, thesauruses, informational retrieval**

This article consists of ten main parts. In *I.Introduction* we analyze the state of art of knowledge management in actual biomedicine applications. In *II.Semantics of informational resources* we means of representation of semantics of the Web informational resources. *III.Statement of retrieval problem* deals with problems of semantically-based informational retrieval forbiomedicine domain. In *IV. Domain thesauri and ontologies* the modern means and languages of knowledge representation are analysed. In *V.Ontological analysis* we describe the formal model of ontology and propose the classification of ontologies. The part *VI.Use of thesauruses for information resources retrieval* deals with use of thesauri in retrieval procedures. In *VII. Constructing of domain thesaurus* an algorithm of domain thesauri building on base of base of different relevant textual documents. *VIII.Algorithm of domain and IR thesaurusus comparison* proposes the matchmaking method of different thesauri that express the semantics of documents. The part *IX.Intelligent informational retrieval in biomedicine domain* describes the use of methods described above in biomedicine on example of intelligent informational retrieval system MAIPS. In *X.Conclusion* we describe the perspectives of semantic informational retrieval in biomedicine domain.

I. Introduction

The effective retrieval in the Internet becomes difficult and laborious for user that has to process a lot of documents that satisfy to formal request but are not pertinent to his/her real information needs. Efficient informational retrieval has to be semantically oriented and based on knowledge of some subject domain. That's why there is necessary to formulize the model of user interests domain (e.g., as ontology), link all information resources (IR) with some subject domains and then develop the algorithm for matching of IR domains with domain of user interests.

We propose to observe this task in example of biomedicine domain, because now a huge volume of biomedical and genome data is Internet-available. But this data are distributed into heterogeneous biological data sources, with little or even none information organization. Therefore, integration and exchange of data within and among organizations is a universally recognized need in bioinformatics.

Now there is a lot of research works in field of knowledge management and ontological analysis in this domain. The knowledge representation community within computer science has the aim of representing knowledge in a form both understandable by humans and one that is computationally amenable.

Ontology is commonly defined as an explicit and formal specification of a shared conceptualization of a domain of interest. Ontologies formalize the intensional aspects of a domain, whereas the extensional part is provided by a knowledge base that contains assertions about instances of concepts and relations as defined by the ontology. The process of defining and instantiating a knowledge base is referred to as knowledge markup or ontology population, whereas (semi-)automatic support in ontology development is usually referred to as ontology learning.

Ontologies have been broadly used in knowledge management applications, with a recent upsurge around Semantic Web applications and research. In recent years,

ontologies have regained interest also within the NLP community, specifically in the context of such applications as information extraction, text mining, and question answering. However, as ontology development is a tedious and costly process there has been an equally growing interest in the automatic learning or extraction of ontologies. Much of this work has been directed towards extraction from textual data as human language is a primary mode of knowledge transfer.

In this way, textual data provide both a resources for the ontology learning process as well as an application medium for developed ontologies.

For the last years, the World Wide Web Consortium (W3C) Health Care and Life Sciences Interest Group (HCLSIG) [1] has investigated the use of Semantic Web technologies in biomedicine. Ontologies play a central role in the Semantic Web [2], especially in biomedicine for which a large number of ontologies have been developed. This group advocates the use of Semantic Web technologies for supporting translational research [3] and has demonstrated the feasibility of integrating disparate resources in the domain of neurosciences, including Entrez Gene, Gene Ontology Annotations, the Allen Brain Atlas, PubMed/MEDLINE, and MeSH [4].

Other "mashups" (integrative applications) have been developed since (e.g., [5]). Similar approaches have been used to integrate genotype and phenotype information [6], pathway and disease information [7], and to create drug-target networks [8]. Biomedical ontologies are crucial to these integration projects.

There are, however, many obstacles preventing ontologies from being used efficiently for data integration. Despite the existence of repositories such as the National Center for Biomedical Ontology's BioPortal [9] and the Unified Medical Language System (UMLS) [10], not all ontologies can be accessed easily. Furthermore, some ontologies in the UMLS are subject to intellectual property restrictions and the UMLS cannot be used without first signing a license agreement. While OBO and OWL are popular formalisms for representing ontologies, many ontologies are available only in proprietary formats.

The increasing volume and diversity of information in biomedical research is demanding new approaches for data integration in this domain. In [11] Semantic Web technologies and applications can leverage the potential of biomedical information integration and discovery, facing the problem of semantic heterogeneity of biomedical information sources. In such an environment, agent technology can assist users in discovering and invoking the services available on the Internet. In this paper authors present SEMMAS, an ontology-based, domain-independent framework for seamlessly integrating Intelligent Agents and Semantic Web Services. Our approach is backed with a proof-of-concept implementation where the breakthrough and efficiency of integrating disparate biomedical information sources have been tested.

Ontologies are becoming essential for data integration as a result of the increase in the quantities and types of data in the molecular biology domain. Simultaneously, the need to organize, co-ordinate and disseminate ontologies as well as coherent ontology development methods is now accepted and is evidenced by the funding of the National Center for Biomedical Ontology (NCBO). Though the need to use ontologies is widely appreciated, the right manner in which they should be developed and applied is not well understood. Researchers still resort to ad hoc methods in developing and using ontologies, resulting in lost opportunities for integration and cross-disciplinary communication, and creation of obstacles to cross-domain reasoning.

Ontologies are used in biomedicine:

- As a controlled vocabulary to annotate genes and gene products (e.g The Gene Ontology);
- As a data exchange format and for data integration (e.g. MGED, SBML and BioPax);
- To define a knowledgebase schema (e.g. BioCyc and Reactome);
- For driving natural language processing (e.g. Textpresso and Geneways);
- For semantically rich querying of federated databases (e.g. TAMBIS);
- Creating formal representations of biological processes for hypothesis evaluation (e.g. HyBrow).

The emergence of information and communication technologies has drastically changed biomedical scientific processes. Experimental data and results today are easy to share and repurpose thanks to the Web and public application programming interfaces (APIs) enabling connection to databases containing such information.

Biomedical researchers have turned to ontologies and terminology to describe their data and turn it into structured and formalized knowledge. For instance, the Gene Ontology2 (GO) is widely used to describe the molecular functions, cellular location and biological processes of gene products as well as integrate these descriptions across several databases.

The bio-ontology community falls into two camps: first we have biology domain experts, who actually hold the knowledge we wish to capture in ontologies; second, we have ontology specialists, who hold knowledge about techniques and best practice on ontology development. In the bio-ontology domain, these two camps have often come into conflict, especially where pragmatism comes into conflict with perceived best practice.

II. SEMANTICS OF INFORMATIONAL RESOURCES

Informational resources (IR) represented in the Internet can be classify on textual and multimedia ones, static and dynamic, structures and not structured etc., but every IR has some semantics and is concerned with some subject domain. In process of information retrieval is very important to discover IR concerned with the domain interested to user.

Structures textual information in the Internet is mainly given in HTML and XML formats. The subject domain of textual IR can be define by two ways:

1) analyzing of IR textual content and

2) considering metadata of these IR.

There is a great deal of the widespread formats for a storing of audio and video information, 3D-scripts and images. The multimedia resources are accessible for indexation much worse than textual information. Therefore for multimedia IRs only the second way is efficient. Metadata contains machine-readable information about the document that can be automatically processed by computer. Now the most perspective and common metadata model is RDF (Resource Description Framework) based on XML.

However, most publicly available biomedical data are unstructured and rarely described with ontology concepts available in the domains.

The challenge is to create consistent terminology labels for each element in the public resources that would allow the identification of all elements that relate to the same type at a given level of granularity.

Metadata can be built in IR or be stored and updated independently of resources. With the help of RDF one can describe the structure of IR and connect it with appropriate domain. RDF describes IR in a form of oriented marked graph - each IR can have properties that also can be IR or their collections. Most widespread set of elements for metadata specification of the Internet IR is Dublin Core Metadata Elements.

Initially World Wide Web technology was focused on work with static IR represented in the Internet. Now a lot of sites offer to the clients not only the documents, but also service (for example, sites of e-commerce). They use application servers that are able to process the data entered by the user (queries, completed form etc.) and dynamically generate new IR depending on the parameters, specified by the user. Such dynamic component of the Internet grows much faster then static one and requires application of more complex information technologies. In this connection it is possible to consider a separate class of IR - Web-services.

Web-service is a set of logically connected and program-accessible through the Internet functions that are based on three basic Web-standards. SOAP (Simple Object Access Protocol) - the protocol for sending of messages by the HTTP and other Internets protocols; WSDL (Web Services Description Language) - language for the description of program interfaces of Web-services; UDDI (Universal Description, Discovery and Integration) - indexing standard of Web-services.

III. Statement of Retrieval Problem

Efficient informational retrieval in biomedical domain has to be semantically oriented and based on knowledge of subject domain. Though we have to develop the algorithms that can use domain ontological information and domain rules in the information retrieval process. That`s why there is necessary to formalyse the model of user interests domain (for example, as ontology), link all IR with some subject domains and then develop the algorithm for matching of IR domains with domain of user interests. The parts of this work are:

- Creating an ontological proection of IR (semantic markup of natural language by ontological tterms) and automatical generation of IR metadata (in RDF format);
- Creation of information model of user as a intersection of user ontology with domain ontology;
- Matching of IR ontology with user ontology dependly to domain ontology.

IV. Domain Thesauri and Ontologies

By definition, "thesaurus" is the study of term usage in given domains associated to a human activity. Now a lot of thesauri exist for medical and biomedicine domain, mathematics, computer science, etc. A term is a sequence of words used in a given domain and which makes sense in this domain.

It is important to understand that terms can be in synonymous relation in some subject domain but not in the general usage. Therefore, thesaurus is on the domain knowledge side and it is used for domain description.

A thesaurus is a sort of terminological base: it is a collection of terms, plus a set of relations among them. In some ways a thesaurus can be a bridge from a terminological base to document indexing. It can be used as a normalization of indexing terms.

Terms of a thesaurus are used to describe a domain terms of a thesaurus are used to describe a domain Manual thesaurus building is a hard task but in this way, one can guarantee a good quality of the collected terms.

Manual thesaurus building is a hard task but in this way, one can guarantee a good quality of the collected terms. Automatic thesaurus building is not guarantee the quality. It relies on the content of document sources and also on the Natural Language treatment implemented.

Thesaurus is extracted from natural language text by means of linguistic analysis.

The structure of thesauri is controlled by international standards that are among the most influential ever developed for the library and information field. The main three standards define the relations to be used between terms in monolingual thesauri (ISO 2788:1986), the additional relations for multilingual thesauri (ISO 5964:1985), and methods for examining documents, determining their subjects, and selecting index terms (ISO 5963:1985). ISO 2788 contains separate sections covering indexing terms, compound terms, basic relationships in a thesaurus, display of terms and their relationships, and management aspects of thesaurus construction. The general principles in ISO 2788 are considered language- and culture-independent. As a result, ISO 5964:1985, refers to ISO 2788 and uses it as a point of departure for dealing with the specific requirements that emerge when a single thesaurus attempts to express "conceptual equivalencies" among terms selected from more than one natural language [12].

Every domain has phenomena that people allocate as conceptual or physical objects, connections and situations. With the help of various language mechanisms such phenomena contacts to the certain descriptors (for example, names, noun phrases).

At present the usefulness of domain ontologies is generally recognized and is caused by their widely use. But the elements and the structure of domain ontologies are not defined identically in different applications.

Now three main approaches to defining of domain ontology exist. They are connected with the ways of ontological analysis application and deal with different sciences. The first one – humanitarian approach – suggests definitions in terms understood intuitively but can't be used for solving of technical problems.

The second one – computer approach – is based on some computer languages (such as OWL, DAML+OIL) for representation of domain ontology and applied software that realized the processing of knowledge represented on these languages.

The OWL (Web Ontology Language) is being designed by the W3C Web Ontology Working Group as a revision of the DAML+OIL web ontology language. This description of OWL contains a high-level, abstract syntax for both OWL and OWL Lite, a subset of OWL. This syntax serves as part of a high-level specification for the formalism. A mapping from the abstract syntax to the OWL exchange syntax is also provided.

The description of OWL here abstracts from concrete syntax and thus facilities access to and evaluation of the language. A high-level syntax is used to make the language features easier to see. This particular syntax has a frame-like style, where a collection of information about a class or property is given in one large syntactic construct, instead of being divided into a number of atomic chunks (as in most Description Logics) or even being divided into even more triples, again for ease of readability. The syntax used here is rather informal, even for an abstract syntax - in general the arguments of a construct should be considered to be unordered whereever the order would not affect the meaning of the construct.

OWL ontology is a sequence of axioms and facts, plus inclusion references to other ontologies, which are considered to be included in the ontology. All OWL ontologies are web documents, and can be referenced by means of a URI. Ontologies also have a non-logical component (not yet specified) that can be used to record authorship, and other non-logical information associated with a ontology.

The third one – mathematical approach – defines the domain ontologies in mathematical terms or by mathematical constructions.

OWL-DL [13] is an ontology language based on description logics (DLs), which are a family of logic-based knowledge representation formalisms describing "objects", "classes" and the "relationships" between them. Most DLs are fragments of standard first order logic. Originally, they were designed to give a unified logical basis to various well-known traditions of knowledge representation like frame-based systems and semantic networks; they have found various applications in conceptual modeling and as a logical underpinning of ontology languages. OWL-DL is based on an expressive DL, i. e., it provides a wealth of constructors to describe complex class expressions from atomic classes and relationships. In this section, we will only use a small portion of OWL-DL's expressiveness to highlight its core features. The semantics of OWL-DL is best understood when talking about "objects" that are "instances" of "classes", and that are related to other objects via "relations".

An object can be an instance of a class, and a class can be a sub-class of another class. For example, the object Robert is an instance of the class Man which, in turn, is a subclass of Person. The meaning of the sub-class relationship is that all instances of the sub-class, Man, are also instances of its super class(es), Person. In OWL-DL, to describe a class, we can describe it in terms of other classes (e.g., saying that Man are "Person and not Woman") and of properties of its instances.

We can consider that at first step of domain ontology building the humanitarian approach is used, then the mathematical model of ontology is constructed, and at last it`s software realization is developed.

Till now no generally accepted universal definition of domain ontology has been suggested. In [14] different definitions are analyzed. On the meaningful level domain ontology will be understood as a set of agreements (domain term definitions, their commentary, statements restricting a possible meaning of these terms, and also a commentary of these statements). Domain ontology is:

- the part of domain knowledge that is not changed;
- the part of domain knowledge that restricts the meanings of domain terms;
- a set of agreements about the domain;
- an external approximation represented explicitly of a conceptualization given implicitly as a subset of the set of all the situations that can be represented.

We consider that a professional activity is a characteristic of a domain. This activity consists in solving different tasks. Task solving needs special knowledge, the same for all the tasks that can be represented verbally. Therefore we can speak about special vocabulary of every domain that is used for specification of tasks and their solutions in this domain. A domain is considered as a set of the tasks, which are solved by specialists of this domain. When solving a task, a person uses a finite set of objects and relations among them. These agreements are a result of understanding among members of the domain community.

V. ONTOLOGICAL ANALYSIS

Every domain has phenomena that people allocate as conceptual or physical objects, connections and situations. With the help of various language mechanisms such phenomena contacts to the certain descriptors (e.g., names, noun phrases).

Professional activity is a characteristic of a domain. This activity consists in solving different tasks. Task solving needs special knowledge, the same for all the tasks that can be represented verbally. Therefore we can speak about special vocabulary of every domain that is used for specification of tasks and their solutions in this

domain. A domain is considered as a set of the tasks, which are solved by specialists of this domain. A domain ontology is the part of domain knowledge that restricts the meanings of domain terms, a set of agreements about the domain.

The formal model of domain ontology O is an ordered triple $O = <X,R,F>$ *(1)*,

where

- X - finite set of subject domain concepts that represents ontology O;
- R - finite set of the relations between concepts of the given subject domain;
- F - finite set of interpretation functions of given on concepts and relations of ontology O.

An ontology is a specification of a conceptualization. The word "ontology" seems to generate a lot of controversy in discussions about AI. It has a long history in philosophy, in which it refers to the subject of existence. It is also often confused with epistemology, which is about knowledge and knowing.

In the context of knowledge sharing, I use the term ontology to mean a *specification of a conceptualization*. That is, an ontology is a description (like a formal specification of a program) of the concepts and relationships that can exist for an agent or a community of agents. This definition is consistent with the usage of ontology as set-of-concept-definitions, but more general. And it is certainly a different sense of the word than its use in philosophy.

The thesaurus can be considered as a special case of ontology. A thesaurus is a networked collection of controlled vocabulary terms. This means that a thesaurus uses associative relationships in addition to parent-child relationships. The expressiveness of the associative relationships in a thesaurus varies and can be as simple as "related to term" as in term A is related to term B [15]. The formal model of thesaurus based on (1) is a pair

$Th = <T, R>$ (2),

where T - finite set of the terms; and R - finite set of the relations between these terms.

A formal definition of a thesaurus designed for indexing is:

- a list of important terms (single-word or multi-word) in a given domain of knowledge; and
- a set of related terms for each term in the list.

Terms are the basic semantic units for conveying concepts. They are usually single-word nouns, since nouns are the most concrete part of speech. Term relationships are links between terms that often describe synonyms, near-synonyms, or hierarchical relations.

All ontologies can be classified as:

- high level, general ontologies;
- domain ontologies;
- individual ontologies (user ontologies, task ontologies).

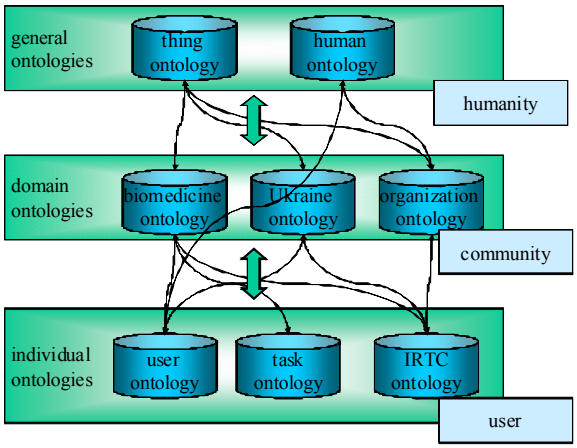

Figure1. General hierarchy of ontologies

General ontologies contains terms that are used by all people and formalize the use of these terms (synonyms, hierarchical, mereological and taxonomic relations).

Domain ontologies contains terms that characterize specific concepts of different sciences, industries, countries and spheres. They are connected with terms from general ontologies but include their subclasses. For example, general term "human" can be concretize and supported by different properties in various ways for domain ontologies for medicine, economy and telecommunications.

Level of individual ontologies is characterized by big number of instances of classes. For example, for class "human" user ontology can contain information about some real people (with their names, addresses etc.).

VI. Use of Thesauruses for Information Resources Retrieval

At thesaurus construction it is necessary to use ontologies of the appropriate areas (with higher level in comparison with user domain to normalize the multilingual thesauruses). Normallization procedure is similar to stemming and provides for integrated processing of words in different morphologic forms and multilingual representations. Normalysed thesaurus contains relation between equivalent terms in different languages. As every thesaurus is constructed from the user point of view (which is reflected in user domain ontology), therefore it`s forming is the user task.

For taking into account semantics of area of user interests in process of retrieval of IR that satisfy his/her informational need it is necessary (fig. 2):

1. to generate the domain thesaurus corresponding to information needs of the user (by analysis of IR that this user considers relevant to this domain [16];
2. to construct the thesaurus for every IR known to IRS (simple dictionary without stop-words);
3. to compare the thesauruses of IR relevant to user query to IRS with the domain thesaurus and to find those ones that contain the maximum number of words in intersection.

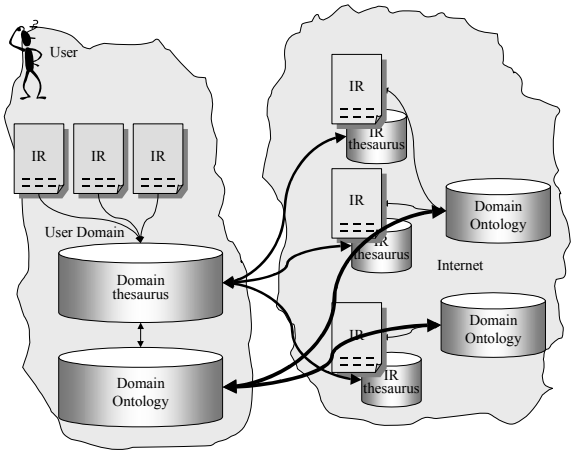

Figure 2 – Informational retrieval on base of domain thesauruses

VII. Constructing of domain thesaurus

At first user should select the set of IR that he/she considers relevant to domain of his/her interests. Every IR is described by not empty set of the textual documents connected with this IR - text of content, metadescriptions, results of indexing etc. The domain thesaurus is formed as a result of the automated analysis of these documents (the user actions are reduced to constructing of semantic bunches - by linking of each word of the formed thesaurus with some term of domain ontology). Algorithm of domain thesaurus construction consists from the following steps:

1. **Formation of initial set** of the textual documents relevant to domain. At the input of algorithm the set A of the textual documents describing chosen IR comes (documents from A can have the coefficients of importance and the coefficients of IR relevance that allows to define differently weight of words from these documents for the IR description).

2. **Creation of domain information space**. For every document from A, $a_i \in A, i = \overline{1,n}$, the IR thesaurus $T(a_i)$ - dictionary that contains all words occurred in the document a_i - is constructed. The IR thesaurus is formed as union of the thesauruses a_i: $T_{IR} = \bigcup_{i=1}^{n} T(a_i)$, and domain thesaurus - as association of the IR thesauruses.

3. **Clearing of the thesauri**. User should specify dictionary for every $a_i \in A, i = \overline{1,n}$ containing a stop-words $s_j, s_j \in Voc$. It is necessary to remove words contained in $s_j, s_j \in Voc$ from the thesauri. Then all service information (e.g., marking tags) is rejected. The cleared thesauri $T`(a_i), \forall p \in T(a_i) \Rightarrow p \in T`(a_i) \vee p \in s_j$, $T`(a_i) \cap s_j = \varnothing$ thus are formed. The cleared IR thesaurus is constructed as association of the cleared thesauruses $a_i$: $T_{IR} = \bigcup_{i=1}^{n} T(a_i)\ T`_{IR} = \bigcup_{i=1}^{n} T`_{IR}(a_i)$, and cleared domain thesaurus - as association of the IR thesauri.

4. **Linking of thesaurus with domain ontology.** To integrate processing of words with equivalent semantics (e.g., synonyms, translations of the term on different languages, various kinds of a spelling) the domain thesaurus is associated with some domain ontology (the user can form it himself, use some ready ontology, modify it or construct it himself).

Each word from the thesaurus it is necessary to link with one of the ontological terms. User has to do it manually on base of his own experience and knowledge in appropriated subject domain, e.g. to link word combinations "Lada de Mandraka" (there is my dog – J.Rogushina) and "Staffordshir terrier" with ontological term "Dog".

For each word in the list of thesaurus terms user defines the ontology name, then selects some one from the list of ontology terms and confirms the link between them.

If the relation is lacking the word is considered as a stop-word or mark-up element (e.g., HTML tag) for domain described in ontology O and should be rejected. $\forall p \in T`(a_i) \exists t = Term(p, O) \in T_O$. If word is significant for domain then go to step for extend the domain ontology.

The group of the IR thesaurus words connected with one ontological term named the **semantic bunch** $R_j, j = \overline{1,n}$ is considered as a single unit, $\forall p \in T`_{IR} \exists R_j = \{r : r \in T`_{IR}, Term(p, O) = Term(r, O)\}$ It allows to integrate processing of semantics of the documents written on various languages and, thus, to ensure the multilinguistic analysis of the Internet IR.

5. **Extension of ontology.** If the IR thesaurus contains words that can`t be linked with ontological terms but user considers that these words are significant than it is necessary to add the appropriate terms to domain ontology, specify their connection with other terms of ontology and return to step 4.

We use Protégé to process the ontologies in OWL. This instrumental tool supports the extension of ontology by new classes and instances.

The Protégé project has come a long way since 1987 when M.Musen first built the Protégé tool for knowledge-based systems [17, 18]. Protégé can be run on a variety of platforms, supports customized user-interface extensions, incorporates the Open Knowledge Base Connectivity (OKBC) knowledge model, interacts with standard storage formats such as relational databases, XML, and RDF, and has been used by hundreds of individuals and research groups.

The original goal of Protégé was to reduce the knowledge-acquisition bottleneck by minimizing the role of the knowledge engineer in KB constructing. Now

Protégé is a general-purpose environment for knowledge modeling.

Protégé allows the developers to build inference mechanisms in an entirely separate component, a problem-solving method, which could be developed independently from the knowledge base. These problem-solving methods (PSMs) were generic algorithms that could be used with different knowledge bases to solve different real-world tasks. Protégé extended the original two-step process—generating a knowledge-acquisition tool and using it to instantiating a knowledge base—with additional steps that dealt with the problem-solving method. This methodology consisted of:

1) developing or reusing a problem-solving method,
2) defining an appropriate domain ontology,
3) generating a knowledge-acquisition tool,
4) building a knowledge base using the tool, and
5) integrating these components into a knowledge-based system by defining mappings between problem-solving methods and specific knowledge bases.

The OntoViz tab plug-in used to give an alternative visualization for the Protégé knowledge base.

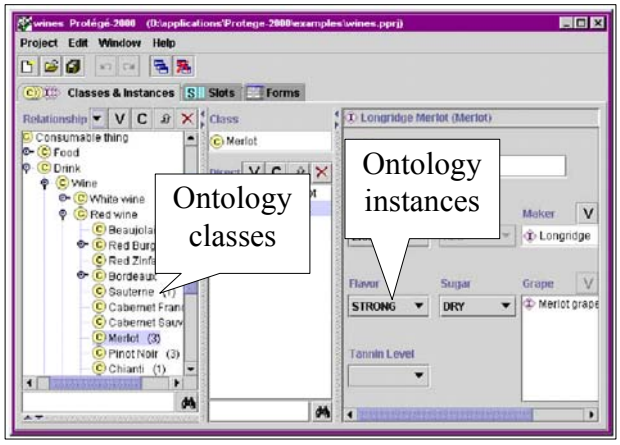

Figure 3 – The default user interface for Protégé

On base of Protégé user can create his own ontologies on base of existing ones that reflects his individual believes about subject domain (fig.3). Such ontologies are not global and widely used but they represent the personalized knowledge of user and normalize his own domain thesaurus.

Relations between ontological terms and words from thesaurus are individual for every user or user's group. They reflect informational interests of user and represent his ability to information processing that is a function of his educational, cultural characteristics and experience etc.

6. **Construction of the normalized domain thesaurus**, i.e. association of all terms of domain ontology that are connected with words from the normalized IR thesaurus (Fig.4).

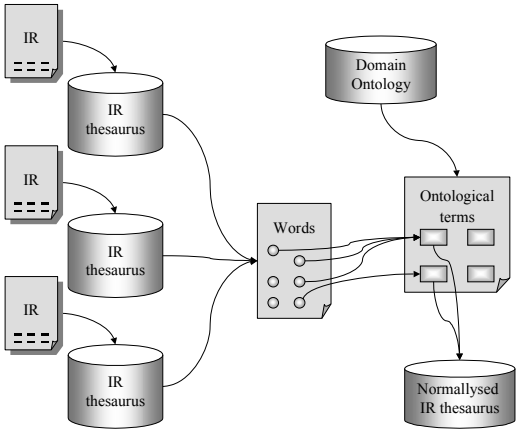

Figure 4 – Building of normalized IR thesaurus

The normalized thesaurus is a projection of set of the IR thesaurus words on set of the domain ontology terms. $L_{IR} = \left\{ t : p \in T\left(a_i\right), i = \overline{1,n}, t = Term\left(p, O\right) \in T_O \right\}$ (1), and normalized domain thesaurus is a union of the normalized IR thesauruses (Fig.4). Informational retrieval systems (IRS) can use this set for representation of subject domain relevant with textual IR (fig.5).

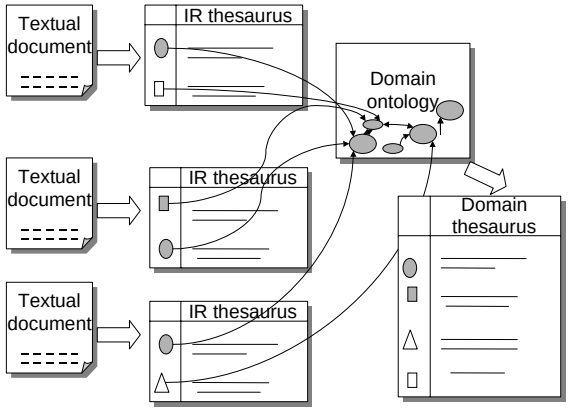

Figure 5– Building of domain thesaurus

As result of the user query execution IRS finds a set of IR. The thesaurus of such IR is simple a dictionary that does not contain the relations between words (discovery of such connections from the text is rather difficult and in this case is not justified). IRS builds this dictionary automatically by IR content processing.

The algorithm of the IR thesaurus building consists of the following steps:

1. Formation of the initial IR set U, $U = \left\{ IR_j, j = \overline{1,m} \right\}$ (2).
2. Formation of the IR thesauri from U from (2). For each IR a thesaurus is formed and cleared.
3. Construction of the normalized IR thesauruses: for normalization the semantic bunches generated by the user during formation of the domain thesaurus are used.

VIII. Algorithm of Domain and IR Thesaurusus Comparison

The normalized IR thesauri L_{IR} and domain thesaurus L_{domain} are the subsets of the domain ontology terms O chosen by the user: $L_{IR} \subseteq Term(O)$, $L_{domain} \subseteq Term(O)$.

If IR description contains more words linked with terms of domain interest for user (that is reflected in the normalized domain thesaurus) then it is possible to suppose that this IR can satisfy informational needs of the user with higher probability than other IR relevant to same formal query. Thus, it is necessary to find IR q satisfied the conditionst $f(q, L_{domain}) = \max f(L_{IR}, L_{domain})$ where the function f is defined as number of elements in crossing of sets L_{IR} and L_{domain}: $f(A,B) = |A \cap B|$. If the various terms of the normalized thesauruses have for the user different importance it is possible to use the appropriate weight coefficients w_j that take into account their importance. In that case the criterion function is

$$f(A,B) = \sum_{j=1}^{z} y(t_j) \quad (3),$$

where the function y is determined for all terms of domain ontology and thesauri.

$$y(t_j) = \begin{cases} 0, t_j \notin A \vee t_j \notin B \\ w_j, t_j \in A \wedge t_j \in B \end{cases} \quad (4).$$

IX. Intelligent Informational Retrieval in Biomedicine Domain

Text is the predominant medium for information exchange among experts. The volume of biomedical literature is inreasing at such a rate making it difficult to efficiently locate, retrieve and manage relevant information without the use of text mining (TM) applications. In order to share the vast amounts of biomedical knowledge effectively, textual evidence needs to be linked to ontologies as the main repositories of formally represented knowledge.

Ontologies are conceptual models that aim to support consistent and unambiguous knowledge sharing and that provide a framework for knowledge integration. Ontology links concept labels to their interpretations, i.e. specifications of their meanings including concept definitions and relations to other concepts. Apart from relations such as “is-a” and “part-of”, generally present in almost any domain, ontologies also model domain-specific relations, e.g. “has-location”, “clinically-associated-with” and “has-manifestation” are relations specific for the biomedical domain.

Therefore, ontologies reflect the structure of the domain and constrain the potential interpretations of terms. As such, ontologies can be used to support automatics semantic interpretation of textual information (Fig. 6), and thus provide a basis for sophisticated TM. Fig.6 lists some popular biomedical ontologies. Many such ontologies exhibit differing degrees of overlap, e x-haustivity and specificity and indeed differing views over conceptual space. Therefore, TM applications that rely on multiple ontologies also need to include methods for mapping between such on-tologies.

These methods, together with other biomedical applications (including TM) that rely on the use of ontologies, would benefit from a standard ontology language (e.g. using standard initiatives such as RDFa and OWLb). Still, even when a single standardised ontology is used, it is not always straightforward to link textual information with ontology due to the inherent properties of language. Two major obstacles are: (1) in -consistent and imprecise practice in the naming of biomedical concepts (terminology), and (2) incomplete ontologies as a result of rapid knowledge expansion.

IR is extensively used by biomedical experts to locate relevant information (most often in the form of relevant publications) on the Internet. Apart from general-purpose search engines such as GoogleTM, many IR tools have been designed specifically to query the databases of biomedical publications such as PubMed [19.20,21].

It is particularly important in biomedicine not to restrict IR to exact matching of query terms, because term ambiguity and variation phenomena may cause irrelevant information to be retrieved (low precision) and relevant information to be overlooked (low recall). Some biomedi-cal ontologies (e.g. UMLS) explicitly store such terminological information (though not always complete). In addition, the hierarchical organisation of ontologies and relations between the described concepts (and through them the corresponding terms) can be used to constrain or relax a search query and to navigate the user through huge volumes of published information. For example, Suarez et al. [22] utilized UMLS for this purpose. Similarly, TIMS30 uses ontology to perform a sophisticated search, which enables users to access implicitly stated relevant information through hierarchical query expansion. More recently, Textpresso [23] is an IR system operating at the sentence level. It uses a specifically designed ontology to query a corpus for information on specific classes of biological concepts (e.g. gene, allele, cell, etc.) and their relations (e.g. association, regulation, etc.).

Domain ontologies are interoperable and can be used in non-specialized intelligent informational retrieval systems. Use of normalized thesauruses linked with domain ontologies is realized in original intelligent IRS system MAIPS [24]. These results can be used for knowledge management in biomedicine domain.

If user want to use MAIPS for informational retrieval in some sphere where he has some stabil informational demand, for example, in biomedicine, he has to make some steps:

- Registration in MAIPS for creation of user profile;

- Choice of domain ontology (for example, a lot of biomedicine ontologies are proposed on Protégé site) (Fig.6);
- Creation of task thesaurus (by set-theoretic operations on sets of ontological terms and natural language analysis) by MAIPS means;
- Formulation of stabil informational quari (with explicide choise of desireble and undesirable informational recources);
- Execution of this query.

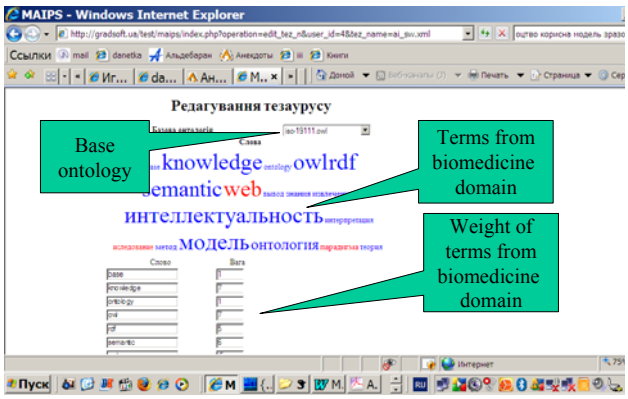

Figure 6 List some popular biomedical ontologies

Results of retrieval by external IRS are filtered by individual user thesauruses built on base of domain ontologies, corresponded to IR and sequence of logical operations on thesauri (Fig. 7).

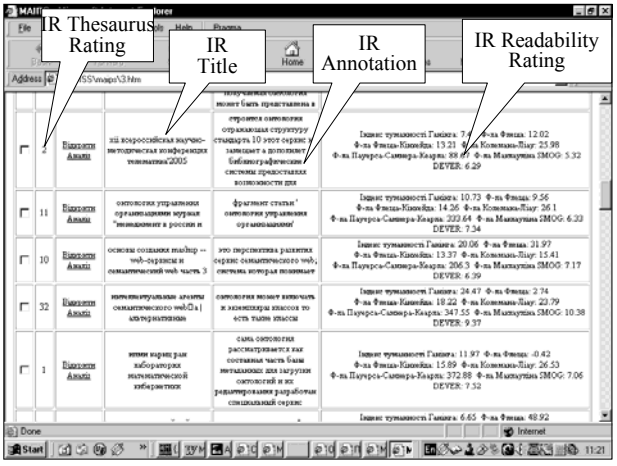

Figure 7 – MAIPS user interface.

Domain knowledge is represented by OWL ontologies

There is currently a huge volume of biomedical and genomic data Internet-available [25]. However, data are distributed into heterogeneous biological data sources, with little or even none information organisation. Therefore, integration and exchange of data within and among organisations is a universally recognised need in bioinformatics [26,27]. One of the major obstacles for integration efforts in bioinformatics is that relevant information is widely distributed, both across the Internet and within individual organisations. Besides, it can be found in a variety of storage formats, including structured and semi-structured ones.

X. Conclusion

We analyze the modern means of knowledge representation of the Web informational resources adequate for special purposes of biomedicine subject domain. The main objectives of ontological approach are an interoperability of knowledge representation, explicit semantics suitable for machine processing, high expressive power and availability of relevant languages, standards and software tools. Thesauri as a special case of ontologies are easier for processing and understanding.

The proposed approach to use of domain ontology for creation and normalization of the IR thesauri allows fulfilling informational retrieval at a semantic level abstracting from language of the IR description. The application of thesaurus measure of the information allows to offer to the user only understandable to him/her items of information that provides pertinence of information retrieval.

In future we plan to construct a repository of biomedicine ontologies and thesauri accompanied with a set of Web services for knowledge management.

References

[1] Health Care and Life Sciences Interest Group.- http://www.w3.org/2001/sw/hcls/

[2] Schroeder, M., Neumann, E.: Semantic web for life sciences. Web Semantics: Science, Services and Agents on the World Wide Web, N.4, 2006, pp. 167-169.

[3] Ruttenberg, A., Clark, T., Bug, W., Samwald, M., Bodenreider, O., Chen, H., Doherty, D., Forsberg, K., Gao, Y., Kashyap, V., Kinoshita, J., Luciano, J., Marshall, M.S., Ogbuji, C, Rees, J., Stephens, S., Wong, G.T., Wu, E., Zaccagnini, D., Hongsermeier, T., Neumann, E., Herman, I., Cheung, K.H.: Advancing translational research with the Semantic Web. BMC Bioinformatics 8, Suppl. 3, S2 (2007).

[4] HCLS Banff 2007 demo.- http://esw.w3.org/topic/HCLS/Banff2007Demo

[5] Sahoo, S.S., Bodenreider, O., Rutter, J.L., Skinner, K.J., Sheth, A.P.: An ontology-driven semantic mash-up of gene and biological pathway information: Application to the domain of nicotine dependence. Journal of Biomedical Informatics (2008), doi: 10.1016/ j.jbi.2008.1002.1006.

[6] Butte, A.J., Kohane, I.S.: Creation and implications of a phenome-genome network. Nat. Biotechnol. 24, pp. 55-62 (2006).

[7] Chabalier, J., Mosser, J., Burgun, A.: Integrating biological pathways in disease ontologies. Medinfo. 12, pp. 791-795 (2007).

[8] Yildirim, M.A., Goh, K.I., Cusick, M.E., Barabasi, A.L., Vidal, M.: Drug-target network. Nat. Biotechnol. 25, pp. 1119-1126 (2007).

[9] BioPortal.- http://www.bioontology.org/tools/portal/bioportal.html

[10] Bodenreider, O.: The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 32, pp. 267-270 (2004).

[11] Garcia-Sanchez F., Fernandez-Breisa J.T, Valencia-Garcia R., Gomez J.M., Martinez-Bejar R. Combining Semantic Web technologies with Multi-Agent Systems for integrated access to biological resources. Journal of Biomedical Informatics (2008), Volume: 41, Issue: 5, pp. 848-859.
[12] B.M. Matthews, K. Miller, M.D. Wilson "A Thesaurus Interchange Format in RDF". – http://www.limber.rl.ac.uk/External/SW_conf_thes_paper.htm
[13] Horrocks I., Patel-Schneider P.F., van Harmelen F.: From SHIQ and RDF to OWL: The Making of a Web Ontology Language. Journal of Web Semantics 2003, 1, pp.7-26.
[14] Gladun A.J., Rogushina J., "Semantic Search of Internet Information Resources on Base of Ontologies and Multilinguistic Thesauruses" // International Journal «Information Theories and Applications», vol.14, 2006.- P.117-129.
[15] Kleshchev A., Artemjeva I. "A Structure of Domain Ontologies and their Mathematical Models". – http://www.iacp.dvo.ru/es/
[16] Gladun A.J., Rogushina J., and Shtonda V. "Ontological Approach to Domain Knowledge Representation for Informational Retrieval in Multiagent Systems", in Information Theories and Applications, V.13, N.4, 2006, pp.354-362.
[17] The differences between a vocabulary, taxonomy, thesaurus, ontology, and meta-model. – http://www.metamodel.com/article.php?story=20030115211223271.
[18] Gennari J.H., Musen M.A., Fergerson R.W., Grosso W.E., Crubezy M., Eriksson H., Noy N.F., and Tu S.W. "The Evolution of Protege: An Environment for Knowledge-Based Systems Development" - http://smi.stanford.edu/smi-web/reports/SMI-2002-0943.pdf.
[19] Srinivasan, P. (2001) 'MeshMap: A text mining tool for Medline', in Proc AMIA Symp, pp. 642-646.
[20] Perez-Iratxeta, C., Pérez, A., Bork, P. and Andrade, M. (2003) 'Update on XplorMed: A web server for exploring scientific literature', Nucleic Acids Res, Vol. 31 (13), pp. 3866-3868.
[21] Fisk, J., Mutalik, P., Levin, F., Er-dos, J., Taylor, C. and Nadkarni, P. (2003) 'Integrating query of relational and textual data in clinical databases: A case study', J Am Med Inform Assoc, Vol. 10 (1), pp. 21-38.
[22] Suarez, H., Hao, X. and Chang, I. (1997) 'Searching for information on the internet using the UMLS and medical world search', in 'Proc 1997 Annual AMIA Fall Symposium', Masys, D., Ed., pp. 824-828.
[23] Müller, H., Kenny, E. and Sternberg, P. (2004) 'Textpresso: An ontology-based information retrieval and extraction system for biological literature', PLoS Biol, Vol. 2 (11), pp. e309.
[24] Intelligent IRS system MAIPS.- http://progproblems.gradsoft.ua/maips-2006/
[25] Gladun A.J., Rogushina J., Garcia-Sanchez F., Martinez-Bejar R. and Fernandez-Breis J.T. "An Application of Intelligent Techniques and Semantic Web Technologies in e-Learning Environments"// Journal of Expert Systems with Applications, ELSEVIER, 2008. - Vol.23.-pp. 72-83.
[26] Gladun A.J., Rogushina J., "Mereological aspects of ontological analysis for thesauri constructing" // in Book "Building and Environment", 2009 Nova Scientific Publishing, pp.198-212. - (www.novapublishers.com).
[27] Gladun A.J., Rogushina J. "Knowledge Management in the Clinical Multiagent E-learning Systems"//Proceedings of Intern. Conf. "Advanced Information and Telemedicine Technologies for Health" , AITTH'2005, Minsk, Belarus, 2005.- P.212-225

Dr.Anatoly Gladun was born in Rivne, Ukraine in 1961. He received the B.Sc. and M.Sc. degrees from Technical University in Lviv, Ukraine in 1984. He holds a PhD in Department of Computer Sciences at the Electrotechnical University (Saint-Petersburg, Russia). He is Head of Department of Intelligent Systems at the International Research and Training Centre of Information Technologies and Systems (National Academy of Sciences). He has been involved in several national and internal research projects (FP5), for example, INCO-Copernicus Project 960114 – EXPERNET "A distributed Expert System for the Management of the National Network", Grant NATO NIG 971779 - "National Telecommunication Networks for Scientific and Educational Institutions" – URAN, ATM-Sat Project "ATM-Based Multimedia Communication" (in GMD FOKUS, Berlin, 2000-2001) etc. Several national research projects, for example, "Research of intellectualization means for multi-agent information retrieval systems". He is the author of more than a 150 publications in conferences, journals and books. His research interests include the development and application of knowledge technologies to different fields such as e-Medicine, e-Commerce, e-Learning, Retrieval Systems, Semantic Web, Network Management, Intelligent Software Agents (models, architectures, methodologies of development) and their Application; Semantic Web Services, Ontologies, Wireless Networks. He is an Associate Professor at the Department of Computer Science (half-time) at University "Kiev-Mogyla Academy", (www.ukma.kiev.ua) and at European university in Kiev, (http://e-u.in.ua/eng/).

Dr.Julia Rogushina was born in Kyiv in 1967. She received the B.Sc. from Kyiv Taras Shevchenko State University in 1989. Her PhD degree in Computer Science she received in Glushkov`s Institute of Cybernetics, Kyiv, in 1995. She is a senior researcher at the Institute of Software Systems, National Academy of Sciences of Ukraine.

Her research interests include the development and application of intelligent information systems; theory of software agents behavior, inductive knowledge acquisition, intelligent information retrieval, ontological analysis, Semantic Web technologies. She has published more than 140 publications in scientific journals and conferences. She is the coauthor of monograph "Agent technologies" and several textbooks. Julia Rogushina has been involved in several national research projects, for example, "Research of intellectualization means for multiagent informaion retrieval systems".

She is an Associate Professor at the Department of Information Systems of Kyiv Slavistic University where gives the courses "Modern Internet technologies", "Systems of Artificial Intelligence", "Data Mining".

