UDC 002.53:004.89

# FORMATION OF SEMANTIC METADATA FOR THE OBJECTS OF KNOWLEDGE CONTROL SYSTEM

A.F. Tuzovskiy

Institute TPU «Cybernetic center»
RAS SD Tomsk scientific center
E-mail: TuzovskyAF@kms.cctpu.edu.ru

*The methods of forming semantic metadata for different elements of knowledge management system have been proposed. The method of manual annotation of different objects in knowledge management system using metadata editor is considered. For the documents the semiautomatic annotation method involving surface linguistic analysis is suggested.*

## Introduction

According to ontological-semantic approach to the development of organization knowledge management system (KMS) [1] all the objects (of documents, specialists, departments, data bases etc.) containing knowledge are described using different metadata [1]. Metadata are the data describing context and content of objects. Context metadata describe the connection of the object with other objects of the system and content metadata describe the content of the object (that is knowledge being in the object). Using metadata, especially content (semantic) one, allows solving efficiently such problems of working with knowledge as search, categorization and recommendation of knowledge. Annotation, the process of metadata creation, may be performed by a man (there where necessary) and without him – automatically. However, in the connection of the fact that the task of text understanding in natural language is still completely unsolved it is not possible to compose qualitative content metadata without a man. At the best, this process is semiautomatic. When programs propose variants of approval for content metadata a man analyses them and either accept them or edits and rejects.

Analyzing content of modern organization information shows that the main part of it is contained in the form of texts in natural language – more than 80 %, in paper-based and electronic form. In this connection one of the most difficult tasks in KMS structure is to develop the methods of composing rather accurate content metadata for text documents.

## Ontological approach to solving the problem of documents annotation

Ontological approach assumes using ontology elements as the content of metadata [2]. Content (semantic) metadata $M_c=\{s_1, s_2, ..., s_m\}$ are the sets of semantic statements (triplets) $s_i$ which have the form $s_i=(c, r, o, v)$ where $c$ is the subject of the statement (a notion or a sample are context metadata of a certain concept) $o$ is the object (a sample is context metadata of a certain concept) and $r$ is the ratio between a subject and an object, and $v$ is the weight ratio which estimates the value of the given statement for describing the object of the knowledge. In this case the notions, ratios should be described in ontology $O$ and samples are described by context metadata of knowledge ontological base. Without using weight coefficients the examples of the statements are the following triads $<C, R, C>$, $<I, R, I>$, $<C, R, I>$, $<I, R, V>$, $<C, R, NULL>$, $<C, NULL, NULL>$, $<I, R, NULL>$, $<I, NULL, NULL>$, where $C$ is the notion; $I$ is the notion sample; $R$ is the connection; $A$ is the attribute; $V$ is the attribute value (text or numerical).

The task of annotation is in semantic metadata creation, that is, in forming a set of statements (triplets) on the basis of a certain ontology and data base corresponding to it. Manual and semiautomatic variants of this problem solving are possible.

Manual variant of realization is in creating metadata editor which allows a user to select the elements of the statement using special interface and ontology of a certain knowledge domain and his knowledge about annotated knowledge object (document, specialist etc.). The main task of interface is to give an opportunity of metadata construction with simultaneous ontology navigation including interactive rending of its segments.

Semiautomatic variant of realization presupposes the creation of subsystem which analyses the object of knowledge having text content and after that it gives to the user the «initial variant» of semantic metadescription which may be edited by the user. In this case specialist time for acquaintance with object content is saved.

Semantic metadata are applied for describing objects of knowledge management system [2] and used in techniques of information semantic processing. Objects may either have text description or have no one. Depending on this fact the formation of semantic metadata is performed in different ways. In the given research the method of forming semantic metadata which defines choice laws of predicates and objects from ontology as well as algorithm of searching notions and samples in the text was developed.

Semantic metadata of portal object should be formed by a man. He should define the elements of semantic metadata in accordance with the matter of description object. The elements represent either triplets with the structure «subject – predicate – object» or certain notions or samples from ontology which we will call «subject». Creating the element of semantic metadata a man should obligatory point «subject». After that he may additionally point the «predicate» and «object».

If the subject is pointed by a man to reflect the matter of description object then additional constraints are imposed on predicate and object choice which result from the laws of forming descriptive logic statements.

The set of possible predicates in a triplet is limited by a chosen triplet subject. After predicate choosing a man should obligatory point the object of a triplet. The set of possible objects depends on chosen predicate. Possible predicate values are determined either by the area of attribute concrete values or by the ratio value area.

If semantic metadata are formed on the basis of text description of an object then the *algorithm of searching notions and samples in the text* is used in addition to the choice laws of predicates and objects. It allows partially automating the process of subject choosing from ontology. For this purpose text description is analyzed on a presence of notions and samples which may serve as subjects in the elements of semantic metadata.

A man forming semantic metadata should edit the obtained set of notions and samples:

- to delete elements not reflecting the matter of object description;
- to remove multiple meaning if the set contains the elements with the same lexical labels;
- to complement the set with notions and samples have not been found by algorithm.

After that elements of the set may be used for triplets forming according to the choice laws of predicates and objects described before.

When semantic portal operating [3] the examined technique is used for forming semantic metadata of various types objects. For example, the algorithm of searching notions and samples in the text is not used in the process of semantic describing specialist knowledge as there is no proper text description of his knowledge. For documents, semantic metadata are developed on the basis of their text content that allows using the algorithm of searching notions and samples.

To compose semantic metadata the set of programs for carrying out manual and semiautomatic annotation was developed: an editor of context metadata for manual annotation of objects and a component of semiautomatic annotation.

### Realization of manual semantic documents annotation

Carrying out objects manual semantic annotation included in semantic portal is realized with the help of an editor of context (semantic) metadata. The given

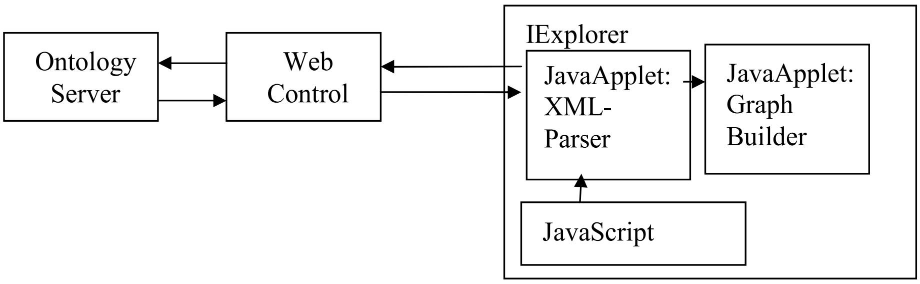
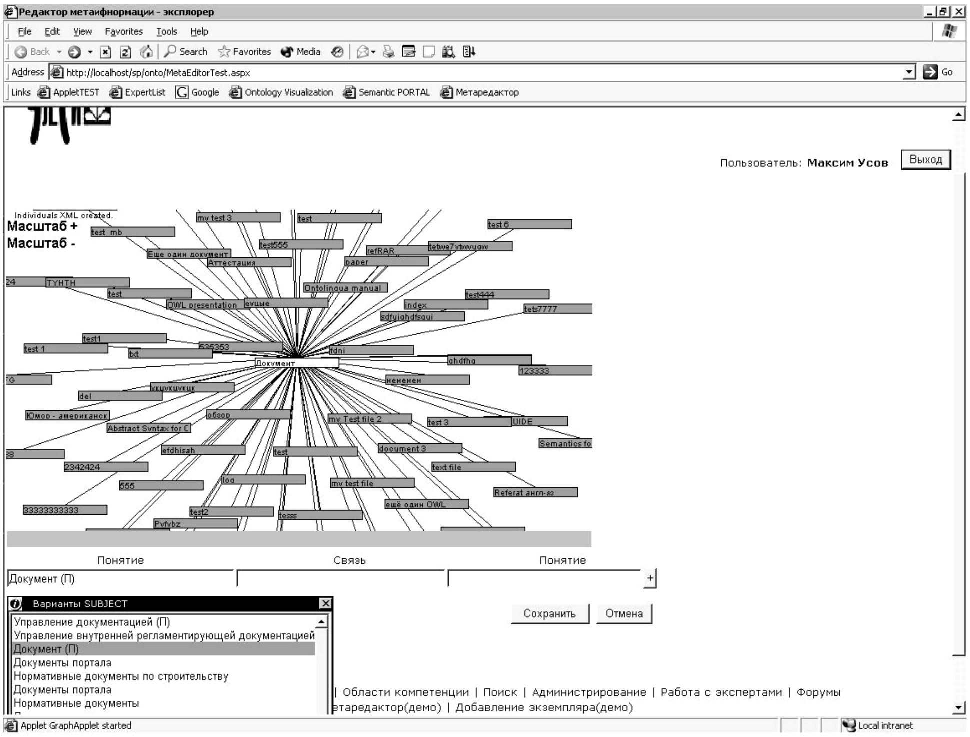


***Fig. 1.*** *Interaction of elements of the component «Ontology navigator»*



Редактор метаифнормации - эксплорер

Address http://localhost/sp/onto/MetaEditorTest.aspx

Пользователь: Максим Усов

Individuals XML created.

Понятие: Документ (П)

Связь:

Понятие:

Варианты SUBJECT

Управление документацией (П)
Управление внутренней регламентирующей документацией
Документ (П)
Документы портала
Нормативные документы по строительству
Документы портала
Нормативные документы

| Области компетенции | Поиск | Администрирование | Работа с экспертами | Форумы
етаредактор(демо) | Добавление экземпляра(демо)

Applet GraphApplet started

***Fig. 2.*** *Interface of semantic metadata editor: image of samples of chosen notion*

editor consists of two parts: a component of interface support (*WebControl*) and HTML page connected with it which also includes data supplier (*DataSource*) and component of imaging ontology structure (Ontology navigation).

The component of interface support is in charge of editor visualization, carrying out of editor operation logic and realized in the *JavaScript* programming language. The program in *JavaScript* language is used both for control and data analysis from data source *DataSource*. Data supplier *DataSource* having obtained the data (ontology description) passes them in *XML* format to the script supporting interface. Typical demands to the data source are: obtaining all notions and samples of ontology on a certain lexical label; obtaining all possible properties (ratios and attributes) for the specified notion or sample; obtaining domain name for ratio (domain may be complex). The scheme of operating the component «Ontology navigator» is shown in Fig. 1.

Element *WebControl* is in charge of *JavaApplet* loading, passing calls to it as well as assigning program interface to *JavaApplet*. Element *JavaApplet* is in charge of drawing function and solving topology tasks. It sends a demand to its data source, receives XML data, disassembles them (grammatical analyzer (*napcep*) *Nano* for mobile devices and telephones chosen due to small size is used ) and visualizes data, has onto function of a rather high tree but images appropriately not more than 3–4 level.

At this stage the data source generates first stage nesting for saving space on a screen (Fig. 2).

Data supplier *DataSource* having received the demand from *JavaApplet* of the given page calls to ontology and forms XML for applet in which the information about ontology hierarchical structure is contained.

User interface consists of lines including three fields. In the first field the chosen notion of ontology (subject) is pointed, in the second one the chosen ratio (predicate) is pointed and in the third one – the specified value. It may be unlimited quantity of such lines. The editor is provided with on-line help which is activated after the moment when user stops changing the text during some time.

### Semiautomatic semantic document annotation

Within the bounds of investigations on semiautomatic annotation the approach based on using linguistic methods of morphological, syntactic and surface semantic (general descriptive) analyses of documents texts in natural language was developed [4]. The result of surface semantic analysis of a sentence in natural language is a connected graph (in some cases a set of graphs) which complements a sentence with a set of connections from a fixed dictionary being, in its turn, a prototype of metainformation for the sentence. Using text analysis in algorithm as initial data of surface semantic analysis result allows escaping a number of interpretation ambiguities connected with linguistic peculiarities of a language.

The next step to composing semantic information is the stage of juxtaposition of semantic graph parts to the ontology elements. Actually the projection of one graph to another here occurs. The projection is possible only at appropriate rules occurrence. Theses rules should represent a certain set of linguistic patterns.

The final task of text analysis is a detection of surface (abstract) meaning expressed in a set of statements. Level-based processing of a text is used for its fulfilling. In this case output data of each next level are the output data of the previous one.

The main levels of processing are: *Grapheme analysis* (marks the text on sentences, words, figures, letters and names); *Morphological analysis* (detects a type of parts of speech of each word, form, declension, number etc.); *Syntactic analysis* (defines syntactic link in a sentence between its parts, detailed syntactic analysis) and *Surface semantic analysis* (it is inherently a modified syntactic analysis).

In contrast to syntactic analysis where links are constructed by fixed algorithms on the basis of morphological data here a set of heuristics and data about traditions of sentence formation in language are used. The result is the established abstract links between the parts. It should be mentioned that this type of links is already inherently a part of high level ontology. However, at this stage there is no notion ontology itself and domain model is not used but there is instead a number of rules and information about traditions of language using (example base).

This type of text analysis is supported by one of Russian design «RML 2006» of AOT company [5]. RML is the product with open code. Its modules are realized in the form of dynamic library and support component technology ActiveX that allows connecting rather easily the library to the supplement in medium .Net. AOT system has a qualitative syntactic analyzer which may be used for generating semantic metadata.

### Method of text analysis

**Preparation.** First of all text is divided into sentences. Sentences are transmitted to AOT syntactic analyzer input. Final and intermediate results of analysis are entered in a hash-table. It is for more rapid operation of a main algorithm which requires multiple normalization and analyzing each word in a sentence and the whole sentence. After that the input text is normalized (cast to initial forms). Ontology is processed in the same way – lexical marks of all notions are also normalized and entered into hash-table.

**Terms abstraction.** Normalized lexical marks of notions are searched among normalized word sequence of the text. Found or similar terms (including compound ones) are separated and put into the list of triplets. Incomplete triplet $<C, , >$ corresponds to each found term where $C$ is the found ontology notion.

**Sample searching.** This stage consists of two parts. The first part is similar to the phase of terms abstraction but in this case only the samples of notions which are already in

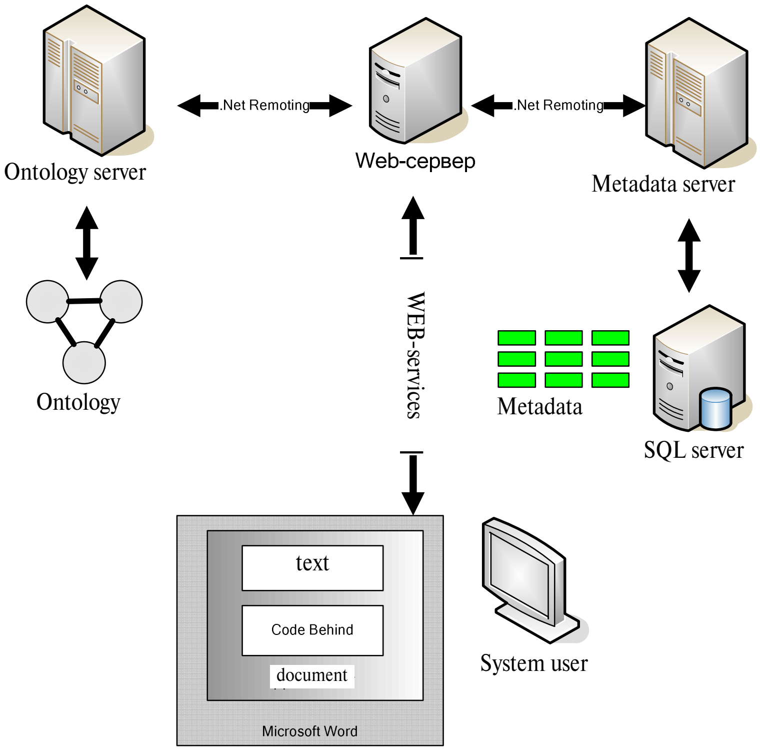


***Fig. 3.*** *Order of interaction of metadata composing system with knowledge management system*

ontology are searched and incomplete triplet <*I*, , >, where I is the found sample of ontology notion, is associated with each found sample. The second part is based on a set of heuristic rules which fulfill the preliminary abstraction of applicants for being samples. After that the check of additional rules is started, for example, if the applicant:

- is close to the notion of object domain then it is a sample and there is a strong possibility that this is a sample of this very notion;
- begins with a capital letter or is written in quotes then it is also definitely a sample of some notion.

### Realizations of semiautomatic annotation system

The basis of the created system of semiautomatic semantic metadata composition is a component of text analysis [4]. This component is realized on the basis of Microsoft Framework 2.0 platform using «Code-behind» technology in respect to the documents created with the use of text editor Microsoft Word 2003. Such approach presupposes that documents of organization are created on the basis of specially developed document pattern. This pattern includes a reference to the executed code of a component, analyzing text, and presented in the form of dynamic library which is loaded on client side with document pattern. Thus, there is a possibility to broaden the functions of usual office facility (in this case it is Microsoft Word) to the possibility of semiautomatic composition of semantic metadata on the basis of text analysis and work with ontology.

Using the component connected with document pattern allows carrying out efficiently semantic analysis directly from text editor Microsoft Word and as a result obtaining semantic metadata in the form of triplets which may be edited and saved in server. The order of interaction of metadata composing program with knowledge management system is shown in Fig. 3.

Connection with a web-server of semantic portal in knowledge management system is carried out through Web-services by SOAP protocol based on XML. Web-server has an assess to ontology through ontology server and to metadata archive where metadata for all portal objects as well as information about their belonging to the objects are saved.

### REFERENCES

1. Davenport T., Prusak L. Working Knowledge: how organizations manage what they know. – Boston: Harvard Business School Press, 1998. – 200 p.
2. Tuzovskiy A.F. The development of knowledge management system on the bases of unified ontological model // Bulletin of the Tomsk Polytechnic University. – 2007. – V. 310. – № 2. – P. 182–185.
3. Tuzovskiy A.F. Architecture of semantic Web-portal // Bulletin of the Tomsk Polytechnic University. – 2006. – V. 309. – № 7. – P. 142–145.
4. Tuzovskiy A.F., Usov M.V. Semantic annotation of documents // Information and system technologies in industry, education and science: Scientific studies of International symposium. – Karagandaz, 2006. – P. 240–242.
5. Documenation package to software product RML // [On line resource]. – 2006. – Assess mode: [http://www.aot.ru/technology.html].

*Received on 21.11.2006*