PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 133
Pozyskiwanie wiedzy i zarządzanie wiedzą 2010

**Jerzy Gołuchowski, Barbara Knapik**
University of Economics in Katowice

# AUTOMATION OF SEMANTIC ANNOTATIONS ON THE EXAMPLE OF ONTEA SYSTEM

**Summary:** One of the most onerous daily activities in organisation is searching and scanning documents. Indexing of documents used in the Document Management Systems is characterized by many limitations. Semantic description of documents, called semantic annotation, allows its users to automate the process of documents' searching. By using automatic or half-automatic annotation, user cannot only benefit from the already created annotations but develop new ones, which may be re-used in the future. The key objective of the article is to present the concept of annotation as one of the approaches towards semantic description of documents, and automatic creation of semantic annotations. Ontea – one of the most popular tools for creating semantic annotation – is presented.

**Keywords:** semantic annotation, ontology, semantic web, taxonomy, identifier of object, knowledge base, document indexing.

## 1. Introduction

Organisations store more and more documents in electronic form. Searching out and searching them is one of the most troublesome activities of everyday life. For making process of searching out the documents more efficient, the indexation in the Document Management Systems is used, which leads to creation of knowledge base gathering data about content of index-linked documents. However, traditional description of documents, used in indexation, is characterized by many limitations.

Semantic description of documents, called annotation of documents, allows users to automate the process of documents' description. Thanks to automatic or semi-automatic description of documents, user can take advantage of annotations created earlier, and make new ones, which can be used to semantic description of documents in the future.

The aim of the paper is to present the idea of annotations, treated as one of the ways of semantic description of documents and possibility to automate description. To illustrate possibilities of semantic annotation, there are many examples, based on one of the most popular tools for creating semantic annotations – Ontea.

## 2. Deficiencies in searching out the documents and necessity of semantic description of documents

Along with the increase in amount of documents, there is more difficulties in searching them out. Traditional methods of indexation and searching out the documents, based on indexes, so far turned out to be insufficient. Retrieval systems can confront values of these features. However, not always there is a possibility to define a value of some feature (e.g. colour). What is more, algorithms which realize searching out, with the aid of classical methods, use only keywords responsible for identifying right object. Search engines cannot read the semantic meaning of word showed by user. That is why during searching out the documents, strictly established values are taken into consideration. The situation is similar, when it comes to synonyms and pointed words. Classical methods of searching out cannot retrieve by synonyms of specific keyword and also cannot "guess" the meaning of the keyword on the basis of the context. For example the term "computer scientist" can take meanings like "programmer" or "administrator". Similar issue applies to the change of the word (e.g. declension).

There have been actions undertaken to introduce new methods of searching out, based on analysis of meaning layer of text, and representing documents by vectors with components from set of semantic elements, instead of set of entries. In this way vectors' row is lower than row in primary model, based on keywords [1].

Formally, semantic description of a document presents specific way of metadata sorting, which guarantees references to individual descriptions on documents or unambiguous identifiers [6]. With regard to semantic description of documents, the term "annotations" is used.

## 3. The essence of semantic annotations in documents

The term "annotate" means providing with footnotes [9]. "Annotation", according to WordNet, has two meanings. The first one defines action as adding comments and description. The second one specifies the result of an action: annotation, description, note or comment, usually added to the text document. Annotation is a way of semantic description of documents, added by users.

In a document, annotations can take the form of comments, hypotheses, explanations and other external kinds in relation to the document. They can also be used as elements enclosed to the whole document or the part of it, without the necessity to interfere in the document itself [4]. Annotations can emerge as passwords or hierarchically described bookmarks, which allow to create classifications and opinion pools, and make finding documents, group work, saving paths of search etc. much more easier.

Annotations are represented by a set of metadata and can be placed on special annotation servers. To use them, specialized tools are required. One of such tools will be discussed in the further part of the paper.

Documents included in semantic web are searched out with the aid of object, opinion and value, assigned to the document (web-of-trust), but not by containing sequence of signs from asked question. This allows to search out the documents more carefully and precisely.

## 4. Creating semantic annotations in documents

Notes are being made with the aid of specific description languages. Usually they are stored on annotation server. Users of this solution can [4]:

– use existing annotations,
– create new annotations,
– modify existing annotations,
– delete annotations.

The way the annotations are being used for description of documents is presented on Figure 1.

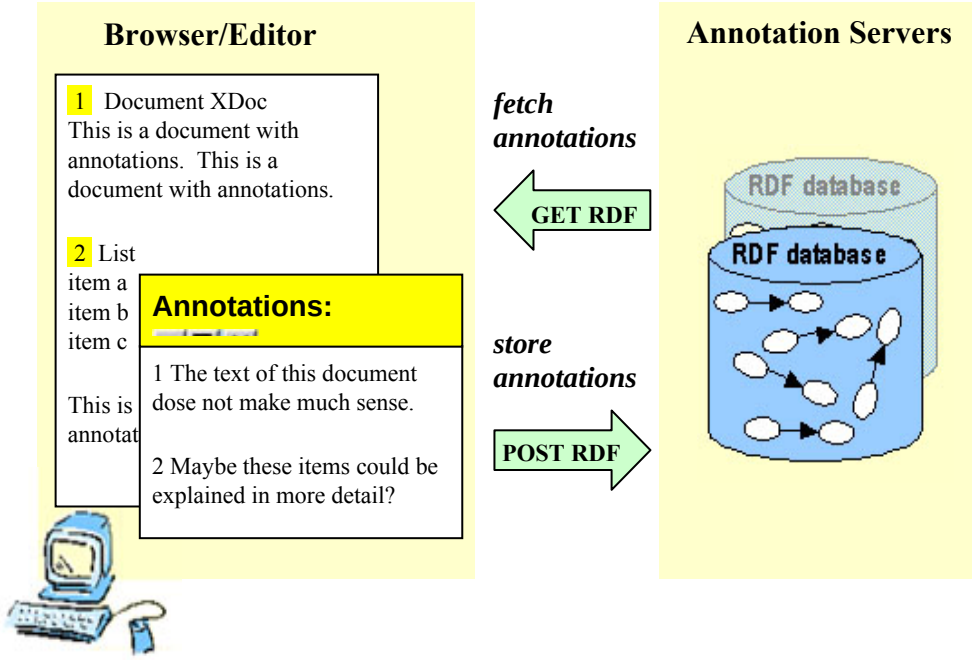


**Figure 1.** Using annotations for description of documents

Source: [4].

There are many alternatives as far as making notes goes. For instance, there are languages such as HTML, SGML or XML, but they are quite faraway from description methods and they have annotations separated from the text, not embedded in the text context. These languages will be discussed in the further part of the paper.

The content of a document can be described semantically, in many different ways. One of the best solutions, with the exclusion of the above-mentioned method, connects descriptions of particular elements, and creates semantic annotations, which allows to manage the document by semantic indexing and searching out, creating hyperlinks, advanced visualization and navigation. Figure 2 shows general diagram of semantic description.

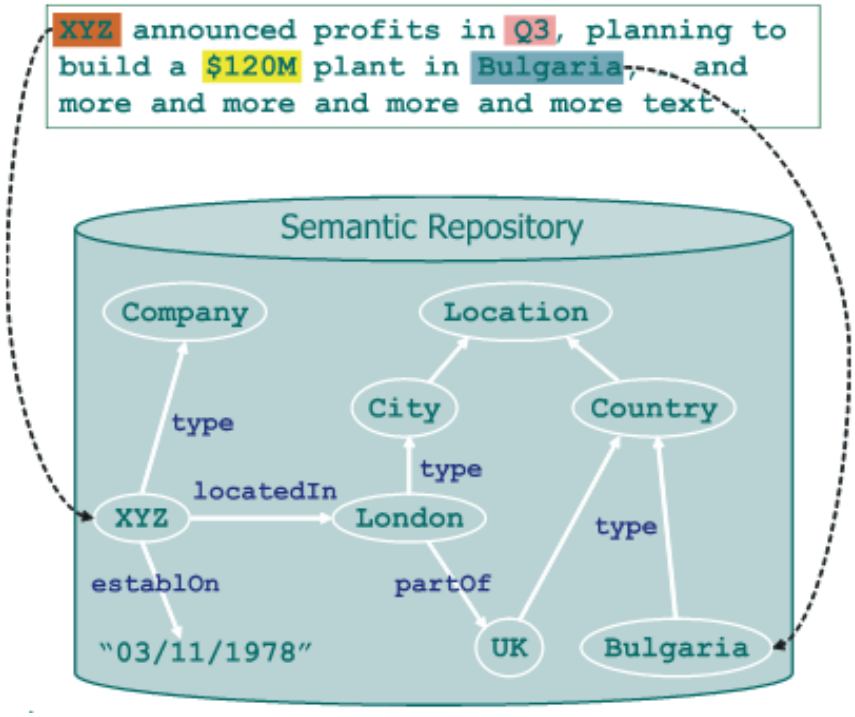


**Figure 2.** Diagram of semantic description for text document

Source: [6].

Important feature of semantic description is a possibility to use it not only for description of text documents, but also for creating special annotation specifications for Web Services (e.g. description of WSDL). For this purpose special instruments, like WSMO Studio, are used [6]. Annotations can be implemented in different ways. The most popular methods of description are ontologies (taxonomies) which define classes of objects. They should give a possibility to refer to

created classes. The second method of description are identifiers of objects, which should be distinguished and transformed into links to their semantic descriptions. The third method are knowledge bases with descriptions of objects.

Foundation of semantic description of documents are ontologies. That is why generally annotations should be created in accordance with the ontology of specific object description language, but there is no uniform rules, but recommendations from creators of systems only.

When it comes to application of annotation building rules, important is popularity of description system in which notes are made [5]. Documents described in this manner compose semantic web which merges documents not only literally, but also semantically.

Tools which use solutions of semantic description of documents can be divided into two groups (in regard of way of notes making):

– manual annotation,
– semi-automatic annotation.

Choice of model is contingent upon concrete annotation use. Particular tools follow annotation idea, and guarantee visually useful instruments for manual annotation, navigation through Web sites, reading and looking through semantic descriptions. Furthermore, they provide infrastructure for protocols, in order to make manual marking of documents, with the use of semantic descriptions, like Annotea or RDF.

Semi-automatic annotation is characterized by creating semantic metadata for future processes, thanks to semantic data in knowledge management and semantic applications in organizations. Semi-automatic annotation is based on the analysis of the document. Some of them are using Google API for automatic annotation. This algorithm seems to be slower, when there is more documents, required to knowledge management and semantic applications in organizations. It is difficult to evaluate the execution, but description of algorithm and its present connections with Google API does not seem to be too fast.

## 5. Ontea system – creating semantic annotations in documents

Automatic annotation of Web documents is main challenge for developing area of Semantic Web. Web documents are structured, but their structure is comprehensible only for a human and that is the biggest and the most difficult barrier for semantic Web to break down. One of the most frequent document description tools – Ontea – is an attempt to solve this problem.

Functioning of Ontea means analysis of document or text, which uses fixed, regular terms and expressions based on pattern, and detecting of equivalent, semantic components, according to specific ontology domain. Assuming that the text, which belongs to a specific field solution with ontology model defined, was ana-

lyzed, there is a chance to create semantic version of it, which can be a basis for future computer analyses, where the structure formalizes required version of document. This can be very helpful when it comes to categorization, document visualization and retrieving, gathering and understanding of knowledge. That is why, as main purposes of Ontea, are mentioned the following [8]:

– detecting/creating metadata from text,
– improving of data structure for later computer processes,
– organizing of data based on ontology model of application.

Ontea, as a tool for document indexing and retrieving, uses RFTS (Rich full-text search) which means full-text searching. RFTS functionality is especially used during creating new ontology units and judges association between newly created cases. Working of Ontea editor is based on analysis of document or text, with the use of regular expressions and discovering equivalent semantic elements, according to defined domain of ontology. Miscellaneous joined expressions have been already defined, but to achieve better results and outcomes, new patterns and expressions should be defined for all recently created applications. Additionally, Ontea builds new ontologies for defined classes individually and assigns discovered elements of ontology as properties for newly created class of ontology. On the other hand, new domain of ontology needs including of additional internal ontologies used in Ontea. To clear up, there is an example of ontology pattern, on Figure 3, with several classes taken from NAZOU project.

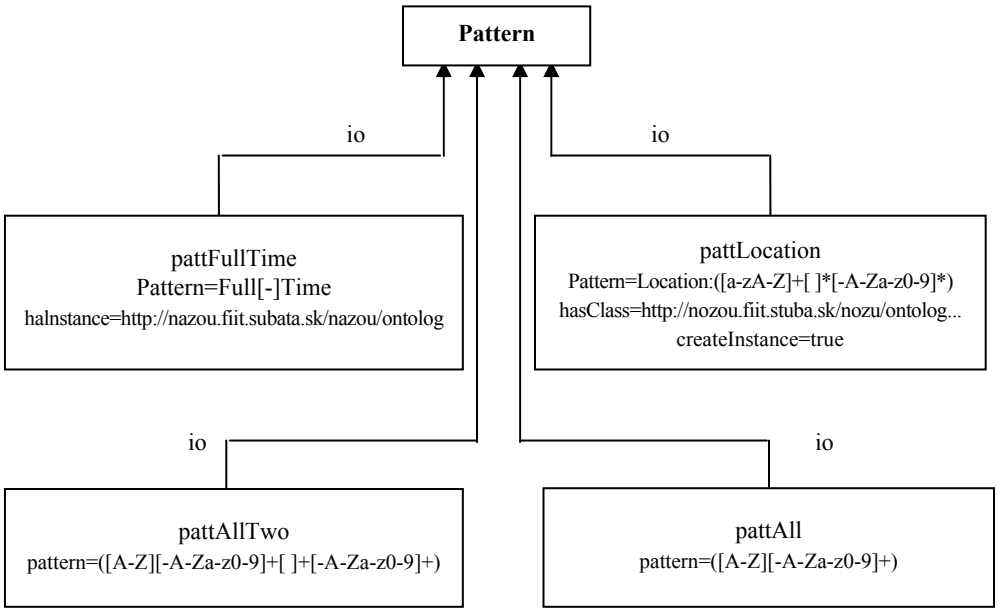


**Figure 3.** Ontology pattern with several units from ontology domain of NAZOU project

Source: [8].

"Pattern" class presents regular expressions which have been used in order to make notes for simple text with elements of ontology. Class {Pattern } – pattern has been evaluated as semantic algorithm of annotation. Figure 3 presents several simple patterns which are able to discover unit of ontology, thanks to connection of properties, from particular units. "Pattern" class has the following properties: has-Class.Pattern, hasInstance.Pattern, pattern.Pattern, pattern.createInstance.

$$
\begin{aligned}
&\text{Pattern} \subseteq \\
&\text{hasClass.Pattern(Thing)} \cap \\
&\text{hasIstance.Pattern(\{Thing\})} \cap \\
&\text{pattern.Pattern(String)} \cap \\
&\text{pattern.createInstance(boolean)} \\
&\{\text{pattern}\} \in \text{Pattern}
\end{aligned}
$$

**Figure 4.** Definition of {Pattern} class

Source: [8].

Elements of {Pattern} class have been pointed out on Figure 4, to define and identify relation between text/document and its semantic version, in accordance with domain of ontology, where "Pattern" properties contain regular expressions which describe text representation of relevant elements of ontology. Looking through a document/text is considered in terms of all regular expressions for each pattern. For example, if hasInstance property is not empty, then unit which contains this property is added to the set of discovered elements of ontology. Moreover, if hasClass property exists in "Pattern", then question is created (RDQL or SeRQL) and, to find unit, conversion is taking place. Unit meets the following conditions:

– unit is hasClass class,
– properties of unit contain matching words.

When createIndividual property is set on "True" and suitable unit, with recovered words in metadata of ontology, has not been found, hasClass unit is built.

Operating of Ontea system is described by the following steps [8].

1. Loading the text from the document.

2. Extracting regular expressions from the text. If appropriate unit of ontology is not recovered, according to the pattern of property, it will be added to the set of discovered units of ontology.

3. If no unit is recovered for specified patterns and createInstance property is established, then simple unit of class type, included in hasClass property, will be created with property rdf only: label containing matching text.

4. The process is repeated for all regular expressions and the result is a set of found units.

5. Empty unit, representing class of extracting text, is created and all possible properties of every class of ontology are discovered thanks to definition of classes.

6. Discovered unit is compared with appropriate type, and if the property of the type is the same as the type (class) unit, then the unit will be attributed for this property.

7. Comparison is ended for all properties of a new unit, which responds to text/document as well as all discovered units.

Algorithm also uses concluding, in order to facilitate task of recovery of units with appropriate properties, also when the type of concluding of found unit is the same as the type of property. The weak point of the above-mentioned algorithm is difficulty with allocating and attribution of element in case, when it answers to recovered text, which contains different properties of the same type. This problem can become resolved, if the algorithm is used only in the course of creation of units of different types (properties).

Architecture of Ontea system is approximated, in relation to its elements, to Oneta algorithm described earlier. Text stocks (such as HTML documents, e-mail or simple text) are entrance data which we want to provide with annotations and domain ontology. Exit data are a new unit of ontology, which answer described text. According to defined patterns, properties of this unit are fulfilled with discovered units of ontology. Ontea architecture is presented on Figure 5.

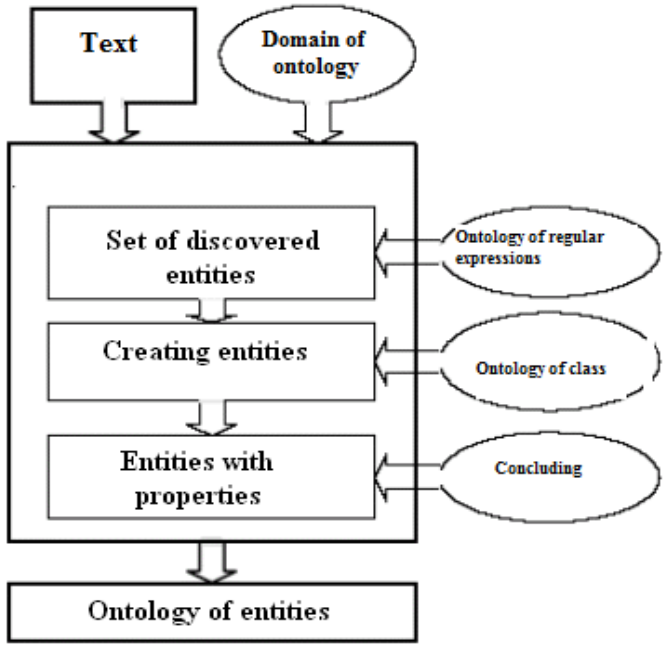


**Figure 5.** Architecture of Ontea tool

Source: [8].

Ontea uses OWL and RDF languages for creation of ontology. It is implemented in Java at use of Jena Semantic Web Library [3] or Sesame library [10]. Concluding is used to achieve better results in both applications.

## 6. Example of using Ontea for semantic description of document

Patterns of regular expressions are fundamental elements of Ontea. Pattern is a sequence of symbols which express mathematical or physical laws [2]. However, usually for each domain of problem, there is requirement to create new, specific for the problem pattern which is suitable for elements of ontology. Nonetheless, there are several exemplars:

– matching one word, which starts with capital letter: ([A-Z][-A-Za-z0-9]+),
– two-verbal pattern: ([A-Z][-A-Za-z0-9]+[\s]+ ([A-Z][-A-Za-z0-9]+),
– similarly to three- and four-verbal patterns.

If units in sphere of ontology include labels with clean text describing units, then they can be discovered by Ontea system. Even with so simple exemplars, achievement of satisfying results is possible. If reference of ontology includes big amount of units with coherent labels, then results of notes are satisfying.

Ontology of localization is good example, it contains such notions as region, state, settlement, mountains, rivers or lakes. Therefore, it allows to create ontology with concrete objects of cities, settlements, mountains or rivers. It is possible to find such data with facility in internet stocks [7]. When such keywords as “Vistula ” or “Warsaw ” appear in the text, they are discovered. “Vistula” – as object of class of river, and „Warsaw” – as capital and subunit of city and settlement. Such discovering can refer to each created entity.

As it has been mentioned earlier, Ontea not only discovers, but also builds objects. For example, on many websites with job offers, localization of work-place is given in the following way: “Localization: city or name of province ”. When websites become converted into clean text, regular patterns of expressions can be recovered with facility: Location [\s]*([A-Z][-a-zA-Z]+[\s]*[A-Za-z0-9]*). Such expression informs that one or two words, which define localization, can be placed after the name of localization. Words as formulation will be treated as a sequence of signs. If, for example, formulation will include “Localization: New York” [7] and words “New York” will be recovered in no object and ontology reference, new simple object with the type of region (localization) can be created. In hasClass properties, this localization is set up for pattern as rdfs:label “New York”.

Thanks to the creation of new object, in the future the words “New York” recovered in the next document will be marked. However, it is noteworthy that if we create the object “New York City ”, this object is surely not region, but rather subclass – city in our ontology of localization. Besides, if any object is changed into object of city class, it will be modernized in all discovered places automatically.

Ontea has been developed in the course of the projects: NAZOU and K-Wf Grid. Semantic description of a document has a big meaning in both projects. In the project K-Wf Grid, Ontea translates and links together input text from user with elements from domain of ontology. Ontea has been created in order to:

– define problem by user with the aid of typical free text. Ontea discovers relevant elements of ontology and creates semantic version of problem, which is understandable for future computer processes,
– use text annotations for knowledge sharing; records are presented to user in an appropriate context.

As far as Slovakian Project called NAZOU, developed in 2004, is concerned, Ontea has been worked out as a specialized tool. To exemplify, a fragment of a job offer (Job Offer Application) has been characterized. Ontea has been used for creation of ontology of metadata for HTML documents. Example of job offer, placed in Internet, is shown on Figure 5. Afterwards, ontology is processed by other instruments of NAZOU. Search engine of job offers is main application, where instruments are used for searching, downloading, categorizing, describing, finding and displaying job proposals for job seeker. Main components of ontology of proposals of work are: category of work, workplace, skills required, which are recognized by Ontea algorithm.

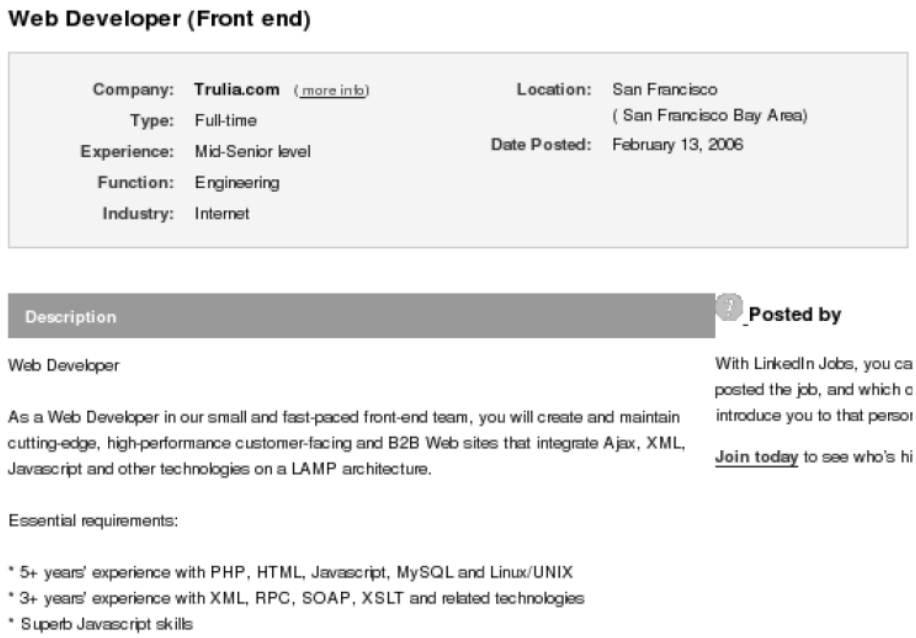


**Figure 6.** Job offer placed on website

Source: [7].

Figure 7 presents a unit of job proposal, based on semantic annotation of work proposal document, which is also placed on the figure. Simple, regular expressions, where units can be mainly discovered by titles of properties (skillSQL, skillPHP), are used for this. In this example, localization of offer proposal (New York and USA) is identified by regular expressions, such as "([A-Za-z]+)" and "([-A-Za-z0-9]+[]+[-A-Za-z0-9]+)", because locNY has property: title "New York" and locUS has property: title "USA". Other elements of ontology are discovered very similarly.

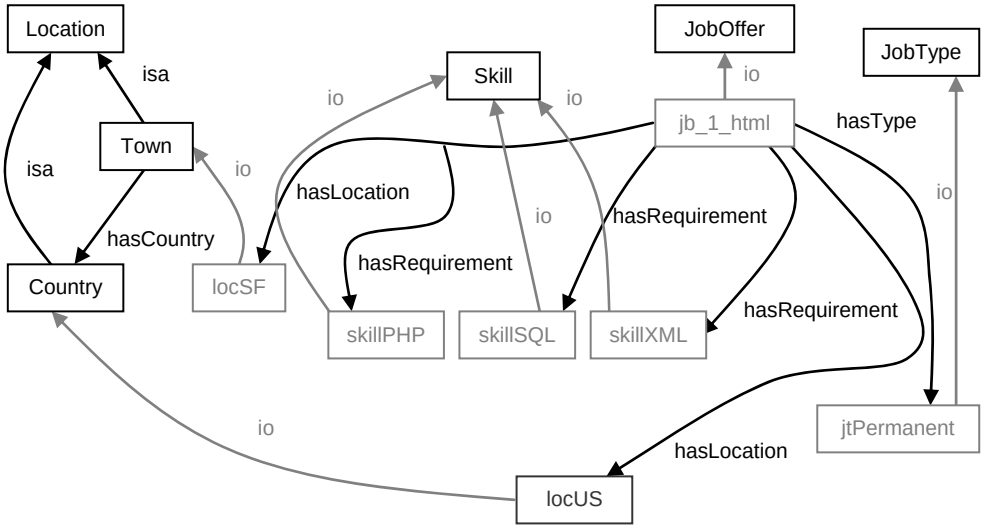


**Figure 7.** Unit of work proposal created in Ontea

Source: [7].

Moreover, discovered units of ontology are allocated as properties of work proposal in this way, so the example of ontology is created beyond its representation of text, in pilot version of NAZOU project. To sum up, it can be emphasized that system discovers elements of ontology, based on domain of ontology.

## 7. Conclusions

Usually semantic description of document, called annotation, is added by users. These can be all kinds of comments, hypotheses, explanations and other external types, affixed to whole document or its part, but without a necessity to interfere in the document itself. It is complicated process, therefore, there are works lasting over automation of creation of semantic annotation.

Semantic annotations are built according to ontology of concrete object description language, however, there is no homogeneous rules, but recommendations

from creators of systems only. The way of creation of semantic description of documents, introduced in the paper, is valuable solution in many systems of documents. However, full utilization of capabilities of introduced tool meets many implementation difficulties. That is why it requires additional research and design-software works.

## References

[1] Deerwester S., Dumais S., Furnas G., Landauer T., Harshman R., *Indexing by Latent Semantic Analysis*, http://www.stat.cmu.edu/~cshalizi/350/readings/Deerwester-et-al.pdf.
[2] *Internetowy słownik języka polskiego*, Wydawnictwo Naukowe PWN, http://sjp.pwn.pl/lista.php?co=wz%F3r.
[3] Jena – A Semantic Web Framework for Java, http://jena.sourceforge.net/.
[4] Kahan J., Koivunen M.-R., Prud'Hommeaux E., Swick R.R., *Annotea: An Open RDF Infrastructure for Shared Web Annotations*, http://www10.org/cdrom/papers/488/.
[5] Kamiński A., *Connotea – nowy typ serwisów informacyjnych*, Ośrodek Dokumentacji Krajowej Szkoły Administracji Publicznej, http://www.ebib.info/2006/77/kaminski.php.
[6] *KIM: Semantic Annotation Platform*, http://www.ontotext.com/kim/semanticannotation.html.
[7] Laclavík M., Seleng M., Babík M., *OnTeA: Semi-automatic Ontology-based Text Annotation Method*, http://nazou.fiit.stuba.sk/home/misc/itat_nazou_ontea.pdf.
[8] Laclavik M., Seleng M., Gatial E., Balogh Z., Hluchy L., *Ontology-based Text Annotation – OnTeA*, http://laclavik.net/publications/P626_ios_press.pdf.
[9] *Multimedialny słownik PWN-OXFORD wersja 1.0*, Wydawnictwo Naukowe PWN SA and Oxford University Press, Warszawa 2005.
[10] *Open RDF*, http://www.openrdf.org/.

## AUTOMATYZACJA SEMANTYCZNYCH ADNOTACJI NA PRZYKŁADZIE SYSTEMU ONTEA

**Streszczenie:** jedną z najbardziej uciążliwych codziennych czynności w organizacji jest wyszukiwanie i przeszukiwanie dokumentów. Ich indeksowanie w systemach zarządzania dokumentami cechuje się wieloma ograniczeniami. Semantyczny opis dokumentów, zwany adnotacją dokumentów, pozwala użytkownikom na zautomatyzowanie procesu opisu dokumentów. Korzystając z automatycznego lub półautomatycznego opisu dokumentów, użytkownik może wykorzystać stworzone już wcześniej adnotacje, a ponadto stworzyć nowe, które mogą być wykorzystane ponownie w przyszłości. Głównym celem artykułu jest przedstawienie koncepcji adnotacji jako jednego ze sposobów semantycznego opisu dokumentów oraz automatyzacji opisu. W tym celu zaprezentowane zostało jedno z najbardziej popularnych narzędzi do tworzenia semantycznych adnotacji – Ontea.