

# **Theory of K-representations as a Source of an Advanced Language Platform for Semantic Web of a New Generation**

Vladimir A. Fomichov,  
Professor of Computer Science, Ph.D., Dr.Sci.

Faculty of Business Informatics  
State University – Higher School of Economics  
Kirpichnaya str. 33, 105679 Moscow, Russia  
E-mail: vdrfom@aha.ru and vfomichov@hse.ru

## **1. Introduction**

The immense social significance of the Web has caused the awareness of the necessity to create the Web Science. One of the central problems of Web Science is to elaborate the theory of transforming the existing Web into a Semantic Web [1].

It has been possible to observe the permanent expansion in the scientific literature of the following opinion: a Semantic Web satisfying the initial goal of this project will be created in an evolutionary way as a result of the efforts of many research groups in various fields. This point of view is expressed, in particular, in [2]. In this paper, the e-science international community is indicated as a community playing now one of the most important roles in quick generation of semantic content in a number of fields. The activity of this community seems to give a sign of future success of Semantic Web project.

In [2], the authors ground the use of RDF as the basic language of the Semantic Web project with the help of the principle of least power: "the less expressive the language, the more reusable the data".

However, it seems that the stormy progress of, first of all, e-science urges us to find a new interpretation of this principle in the context of the challenges faced nowadays by the Semantic Web project. E-science (in particular, bioinformatics) needs to store on the Web the semantic content of the definitions of numerous notions, the content of scientific articles, technical reports, etc. The similar requirements are associated with semantics-oriented computer processing of the documents pertaining to economy, law, politics. In particular, it is necessary to store the semantic content of the articles from newspapers, of TV-presentations, etc.

That is why it can be conjectured that, in the context of the Semantic Web project, the following new interpretation of the principle of least power is reasonable: an advanced language platform for Semantic Web is to allow for modeling a system of operations on conceptual structures enabling us to build semantic representations (SRs) of practically arbitrary texts in Natural Language (NL) pertaining to arbitrary field of professional activity.

## **2. Conceptual Operations Introduced by the Theory of K-representations**

The question immediately emerges what this system of operations might look like. A possible answer to this question is given by the theory of K-representations (knowledge

representations) stated in numerous publications of the author in English and Russian, in particular, in [3 – 8]. The basic mathematical model of this theory describes a system consisting of 10 partial operations on conceptual structures [4, 6, 8]. The model determines a new class of formal languages for building SRs of sentences and complex discourses in NL – the class of SK-languages (standard knowledge languages). An early version of this model set forth in [3] determines the class of RSK-languages (restricted standard knowledge languages).

Let's consider the central ideas of determining the class of SK-languages. At the first step (consisting of a rather long sequence of auxiliary steps), a class of formal objects called *conceptual bases* (*c.b.*) is defined. Each c.b.  $B$  is equivalent to a system of the form  $(c_1, \dots, c_{15})$  with the components  $c_1, \dots, c_{15}$  being mainly finite or countable sets of symbols and distinguished elements of such sets. In particular,  $c_1 = St$  is a finite set of symbols called sorts and designating the most general considered notions (concepts);  $c_2 = P$  is a distinguished sort "meaning of proposition";  $c_4 = X$  is a countable set of strings used as elementary blocks for building knowledge modules and semantic representations (SRs) of texts;  $X$  is called a primary informational universe;  $c_5 = V$  is a countable set of variables;  $c_7 = F$  is a subset of  $X$  whose elements are called functional symbols.

Each c.b.  $B$  determines three classes of formulas, the first class  $Ls(B)$  being considered as the principal one and being called *the SK-language (standard knowledge language) in the basis  $B$* . Its strings (they are called K-strings) are convenient for building SRs of NL-texts. We'll consider below only the formulas from the first class  $Ls(B)$ .

In order to determine for arbitrary c.b.  $B$  three classes of formulas, a collection of inference rules  $P[0], P[1], \dots, P[10]$  is defined. The rule  $P[0]$  provides an initial stock of formulas from the first class. E.g., there is such c.b.  $B_1$  that, according to  $P[0]$ ,  $Ls(B_1)$  includes the elements

*house1, green, city, set, China, 7, all, any,*  
*Height, Distance, Staff, Suppliers, Quantity, x1, x2, P1, P7.*

For arbitrary c.b.  $B$ , let  $Degr(B)$  be the union of all Cartesian  $m$ -degrees of  $Ls(B)$ , where  $m \in \mathbb{N}$ . Then the meaning of the rules of constructing well-formed formulas  $P[0], P[1], \dots, P[10]$  can be explained as follows: for each  $k$  from 1 to 10, the rule  $P[k]$  determines a partial unary operation  $Op[k]$  on the set  $Degr(B)$  with the value being an element of  $Ls(B)$ .

For instance, there is such conceptual basis  $B$  that the value of the partial operation  $Op[7]$  (it governs the use of logical connectives AND and OR) on the four-tuple  $\langle \check{U}, Austria, France, Germany \rangle$  is the K-string  $(Austria \check{U} France \check{U} Germany)$ .

Thus, the essence of the basic model of the theory of SK-languages is as follows: this model determines a partial algebra of the form

$$(Degr(B), Operations(B)),$$

where  $Degr(B)$  is the carrier of the partial algebra,  $Operations(B)$  is the set consisting of the partial unary operations  $Op[1], \dots, Op[10]$  on  $Degr(B)$ .

The volume of complete descriptions in [4, 6, 8] of the mathematical model introducing, in essence, the operations  $Op[1], \dots, Op[10]$  on  $Degr(B)$  and, as a consequence, determining the class of SK-languages considerably exceeds the volume of this paper. That is why, due to objective reasons, this model can't be included in this paper. So let's only regard (ignoring many details) the structure of strings that can be obtained by applying any of the rules P[1],..., P[10] at the last step of inferring the formulas.

The rule P[1] enables us to build K-strings of the form  $Quant Conc$  where  $Quant$  is a semantic item corresponding to the meanings of such words and expressions as "a certain", "any", "arbitrary", "each", "all", "several", etc. (such semantic items will be called *intensional quantifiers*), and  $Conc$  is a designation (simple or compound) of a concept. The examples of K-strings for P[1] as the last applied rule are as follows:

*certn house1, all house1, certn consignment,*  
*certn box1 \* (Content1, ceramics),*

where the last expression is built with the help of both the rules P[0], P[1] and the rule with the number 4, the symbol '*certn*' is to be interpreted as the informational item corresponding to the expression "a certain".

The rule P[2] allows for constructing the strings of the form  $f(a_1, \dots, a_n)$ , where  $f$  is a designation of a function,  $n1$ ,  $a_1, \dots, a_n$  are K-strings built with the help of any rules from the list P[0],..., P[10]. The examples of K-strings built with the help of P[2]:

*Distance(Paris, Moscow),*  
*Weight(certn box1 \* (Colour, green)(Content1, ceramics)).*

Using the rule P[3], we can build the strings of the form  $(a1 \circ\circ a2)$ , where  $a1$  and  $a2$  are K-strings formed with the help of any rules from P[0],..., P[10], and  $a1$  and  $a2$  represent the entities being homogeneous in some sense. The examples of K-strings for P[3] are as follows:

*(Distance(Paris, Moscow) \circ y3 ), (y2 \circ y5) ,*  
*( Height(certn house1) \circ 36/m) .*

The rule P[4] is destined, in particular, for constructing K-strings of the form  $rel(a_1, \dots, a_n)$ , where  $rel$  is a designation of  $n$ -ary relation,  $n1$ ,  $a_1, \dots, a_n$  are the K-strings formed with the aid of some rules from P[0], ..., P[10]. The examples of K-strings for P[4]:

*Belong(Tomsk, Cities(Russia)),*  
*Subset(certn series1 \* (Name-origin,zinnat), all antibiotic).*

The rule P[5] enables us to construct the K-strings of the form  $Expr : v$ , where  $Expr$  is a K-string not including  $v$ ,  $v$  is a variable, and some other conditions are satisfied. Using P[5], one can mark by variables in the semantic representation of any NL-text: (a) the descriptions of diverse entities mentioned in the text (physical objects, events, concepts, etc.), (b) the SRs of sentences and of larger texts' fragments to which a reference is given in any part of a text. Examples of K-strings for P[5]: *certn house1 : x3*, *Higher(certn house1 : x3, certn house2 : x5) : P1*. The rule P[5] provides the possibility to form SRs of texts in such a manner that these SRs reflect the referential structure of NL-texts.

The rule P[6] provides the possibility to build the K-strings of the form  $\neg Expr$ , where  $Expr$  is a K-string satisfying a number of conditions. The examples of K-strings for P[6]:  $\emptyset$  *antibiotic*,

$\emptyset$  *Belong(penicillin, certn series1 \* (Name-origin, tetracyclin))*.

Using the rule P[7], one can build the K-strings of the forms  $(a_1 \dot{\cup} a_2 \dot{\cup} \dots \dot{\cup} a_n)$  or  $(a_1 \dot{\cup} a_2 \dot{\cup} \dots \dot{\cup} a_n)$ , where  $n > 1$ ,  $a_1, \dots, a_n$  are the K-strings designating the entities which are homogeneous in a certain sense. In particular,  $a_1, \dots, a_n$  may be SRs of assertions (or propositions), descriptions of physical things, descriptions of sets consisting of things of the same kind, descriptions of concepts. The following strings are the examples of K-strings for P[7]:

*(streptococcus \dot{\cup} staphylococcus),*  
*(Belong((Bonn \dot{\cup} Heidelberg \dot{\cup} Stuttgart), Cities(Germany)) \dot{\cup}*  
*\emptyset Belong(Bonn, Cities((Finland \dot{\cup} Norway \dot{\cup} Sweden))))*.

The rule P[8] allows us to build, in particular, K-strings of the form  $c * (rel_1, val_1), \dots, (rel_n, val_n)$ , where  $c$  is an informational item from the primary universe  $X$  designating a concept, for  $i=1, \dots, n$ ,  $rel_i$  is the name of a function with one argument or of a binary relation,  $val_i$  designates a possible value of  $rel_i$  for objects characterized by the concept  $c$ . The following expressions are the examples of K-strings for P[8]:

*box1 \* (Content1, ceramics),*  
*consignment \* (Quantity, 12)(Compos1, box1 \* (Content1, ceramics)).*

The rule P[9] allows for building, in particular, the K-strings of the forms  $" \nu (conc) D$  and  $\$ \nu (conc) D$ , where  $"$  is the universal quantifier,  $\$$  is the existential quantifier,  $conc$  and  $D$  are K-strings,  $conc$  is a designation of a primary concept ("person", "city", "integer", etc.) or of a compound concept ("integer greater than 200", etc.).  $D$  may be interpreted as a SR of an assertion with the variable  $\nu$  about any entity qualified by the concept  $conc$ . The examples of K-strings for P[9] are as follows:

*" n1 (integer) \$n2 (integer) Less(n1, n2),*  
*\\$y (country \* (Location, Europe)) Greater(Quantity(Cities(y)), 15).*

The rule P[10] is destined for constructing, in particular, the K-strings of the form  $\langle a_1, \dots, a_n \rangle$ , where  $n > 1$ ,  $a_1, \dots, a_n$  are K-strings. The strings obtained with the help of P[10] at the last step of inference are interpreted as designations of  $n$ -tuples. The components of such  $n$ -tuples may be not only designations of numbers, things, but also SRs of assertions, designations of sets, concepts, etc.

### 3. Conclusions

There are weighty grounds to believe that, combining ten partial operations determined by this model we are able to construct (and it is convenient to do) a semantic representation of arbitrarily complex NL-text pertaining to arbitrary field of professional

activity, in particular, complex definitions of the notions, business contracts, the descriptions of technologies, etc.

The analysis carried out in [4 – 8] shows that the theory of K-representations is a convenient tool for (a) building semantic annotations of Web-sources and Web-services, (b) semantic data integration in the field of e-science and e-health, (c) representing the results of semantic-syntactic processing of NL-sentences and NL-discourses, (d) constructing formal representations of the contents of arbitrary messages sent by computer intelligent agents (CIAs), (e) describing communicative acts, (f) building formal representations of the contracts concluded by CIAs solving the problems of e-commerce, (g) reflecting metadata about the resources, (h) building high-level conceptual representations of pictures.

That is why it seems that the theory of K-representations may be called a Universal Resources and Agents Framework and may be considered as an appropriate framework for developing an advanced language platform for Semantic Web of a new generation and for semantic integration of data accumulated in e-fields.

### References

1. Hendler, J., Shadbolt, N., Hall, W., Berners-Lee, T., Weitzner, D. : Web science: an interdisciplinary approach to understanding the web. *Communications of the ACM*, 2008, Vol. 51, No. 7. P. 60-69.
2. Shadbolt, N., Hall, W., Berners-Lee, T.: *Semantic Web Revisited*. *IEEE Intelligent Systems*, 2006, Vol. 21, No. 3.
3. Fomichov, V.A.: A mathematical model for describing structured items of conceptual level. *Informatica. An Intern. J. of Computing and Informatics (Slovenia)*, 1996, Vol. 20, No. 1. P. 5-32.
4. Fomichov, V.A.: *The Formalization of Designing Natural Language Processing Systems*. Moscow, MAX Press, 2005 (in Russian). 368 p.
5. Fomichov, V.A.: Standard K-Languages as a Powerful and Flexible Tool for Building Contracts and Representing Contents of Arbitrary E-Negotiations. In: Bauknecht K., Proell B., Werthner H. (eds.), *The 6th Intern. Conf. on Electronic Commerce and Web Technologies "EC-Web 2005"*, Copenhagen, Denmark, Aug. 23 - 26, *Proceedings. Lecture Notes in Computer Science*, Vol. 3590, Springer Verlag, 2005. P. 138-147.
6. Fomichov, V.A.: *Mathematical Foundations of Representing the Content of Messages Sent by Computer Intelligent Agents*. Moscow, State University – Higher School of Economics, The Publishing House "TEIS", 2007 (in Russian).
7. Fomichov, V.A.: A comprehensive mathematical framework for bridging a gap between two approaches to creating a meaning-understanding Web. *International Journal of Intelligent Computing and Cybernetics (Emerald Group Publishing Limited, UK)*. 2008, Vol. 1, No. 1. P. 143-163.
8. Fomichov, V.A.: *Semantics-Oriented Natural Language Processing: Mathematical Models and Algorithms*. New York, Berlin, Heidelberg, Springer US, 2009.