FORMAL LANGUAGE FOR NATURAL SCIENCES

Alexander Shkotin,

State Geological Museum, Russian Academy of Sciences, Moscow, Russian Federation

Email ashkotin@acm.org

Mathematical representation of knowledge allows one to perform its automated logical processing. At present, the body of natural-scientific knowledge is presented as texts (monographs, papers, reports and the like), databases and mathematical models.

We believe that controlled natural language (CNL) [1] is enough to record textual and database knowledge. The crucial point herewith is the choice of the formal language, in which the knowledge should be translated from CNL.

At first approximation, following the experience of mathematics, let us assume that the knowledge in particular natural-scientific domain can be represented as an axiomatic theory. Then the texts written in natural language would be translated into formulae and conclusions of the theory [2]. For CNL, this would be done automatically. To find out what properties the formal language of science (FLoS) should possess to fit the given domain, perhaps the simplest solution is to examine the record of this domain (e.g., petrography) in CNL. In our case, originally there were two CNLs, Russian and English (ECNL), and in the present work, due to the language of writing, only ECNL will be used.

The principal statements include those of rock samples taken by different researchers throughout the world and rock definitions in different classifications. This can be exemplified by the Proba database [3]:

1. ECNL statement:
Issue "1(2)" of journal "ACTA NAT. ISLAND." of 1946 has paper B5633 at pages 1-15.
This is formalized in three formulae.

Paper(B5633).
Title(B5633,"A CONTRIBUTION TO THE GEOLOGY OF THE KERLINGARFJELL").
∀x:body issue_of_journal(x) →
(title_of_journal(x,"ACTA NAT. ISLAND.") ∧ year_of_publication(x,1946) ∧
issue_of_journal(x,"1(2)") → read(x,"1,15")=B5633),

where
- B5633 is a variable of data sort,
- “paper” is unary predicate with argument of data sort, which identifies whether the addressed data is paper,
- x:body means that variable x belongs to body sort, i.e., a body in physical sense, normally solid (as we used to think),
- “issue_of_journal” is unary predicate with argument of body sort (indeed, for any solid body a question whether it is a journal issue can be definitely answered),

- “title”, “title_of_journal”, “issue_of_journal” are binary predicates, in which the first arguments belongs to body sort, and the other one is String,
- year_of_publication is binary predicate, the first argument of which is of body sort, and the other is number.
- “read” is a function of two arguments, the first, of the body sort (i.e., physical body, which should be examined for data), and the other, string (saying which part of the body should be examined). The returned value belongs to data sort.

2. ECNL statement:
Weight percent of SIO2 in sample S32994 is 73.95.
It can be translated into a formula with an operator as follows.

WPC(SIO2)(S32994)=73.95.

The operator WPC for each predicate of substance (for ex. SIO2) gives a function, which assigns number to a physical body, i.e. being applied to any given body this function (for ex. WPC(SIO2)) returns a number (mostly 0).

In the report [4] an experience of formalization of rock type definitions is given.

3. harzburgite(sb1) =
plutonic(sb1) and not (pyroclastic(sb1) or kimberlite(sb1) or lamproite(sb1) or lamprophyre(sb1) or charnockite(sb1))
and WPC_carbonates(sb1)≤50 and WPC_melilite(sb1)≤10 and WPC_M(sb1) ≥ 90
and WPC_kalsilite(sb1)=0 and WPC_leucite(sb1)=0 and WPC_hornblende(sb1)=0
and 0.4*WPC_OOC(sb1)≤WPC_Ol(sb1)≤0.9*WPC_OOC(sb1) and
WPC_Cpx(sb1)<0.05*WPC_OOC(sb1)

The above experimentation gives ground to suggest some properties the FLoS must possess to fit the requirements of petrography.
- FLoS must include decimal values and strings which represent “data”. Hence, FLoS must be multi-sort.
- FLoS must include operators.
- FLoS must ensure an opportunity for defining newly introduced constants, predicates, functions and operators.

Building an axiomatic theory of a particular science, one should focus on identification of primary predicates and look for definition of others via these. The resulting theory should include finite models that handle the state of a given domain.

Importantly, Web Ontology Language 2 (OWL 2) [5] or, at least, its DL part provides operators that top the first-order predicate calculus language and therefore may readily enter FLoS. Nevertheless, not all FLoS formulae will be translatable into OWL 2, and operator-based formulae are not well readable and, consequently, may have limited use. These can be compared in some sense with the relational algebra formulae. Still, the interpretability of some FLoS formulae in OWL 2 brings all the benefits of DL inference engine.

In conclusion, the main instrument for modeling the world is mathematical modeling of different complexity. FLoS is a means of interaction of scientist with mathematical modeling. It provides self-consistent data to models and enables logical inference. Furthermore, it can be used for verification of model algorithms.

References

[1] Wikipedia(en), Controlled natural language, http://en.wikipedia.org/wiki/Controlled_natural_language, 2012

[2] Davis, E. Representations of commonsense knowledge, 1990, Morgan Kaufmann Publishers, Inc, San Mateo, California. 515 p.

See also Ernest Davis, Guide to Expressing Facts in a First-Order Language, February 17, 2006. http://cs.nyu.edu/davise/pubs.html

[3] Shkotin, A., and Ryahkovsky, V. Proba DB. Ontology of RDB, Report (ru), 2008-2012 https://sites.google.com/site/alex0shkotin/formalnaa-geologia/bd-proba-ontologia

[4] Shkotin, A., Ryahkovsky, V., and Kudryavtsev, D. Classification algorithm of igneous rocks, and the formal definition of rock type, Report (ru), 2010

https://sites.google.com/site/alex0shkotin/formalnaa-geologia/acmgp-september

[5] W3C, OWL 2 Web Ontology Language. Document Overview. W3C Recommendation 27 October 2009, http://www.w3.org/TR/owl2-overview/