

INIS Multilingual Thesaurus and Nuclear Knowledge Management

V. Koupriyanov

TSNII Atominform
MINATOM
Moscow, Russian Federation
E-mail: kvm@ainf.ru

Multilingual Approach to a Nuclear Knowledge

INIS System has celebrated recently the 30 years jubilee. This system is unique by set of characteristics. One of them is initial idea about necessity to use the different language information on the nuclear science and technology under a unity field area.

It was proposed to take the multilingual dictionary of the terms as a base of this unity and the thesaurus and classificatory designed on this base.

Currently this idea is not finally full accepted. In really good condition the only English-language tools for classification and finding are to be for society. However, the efforts to realize the multilingual system are in progress now by INIS section. There are the follow synchrony versions for multilingual dictionary: English, Spanish, French, Russian, German, Chinese are in preparing, and Arabic is in plan.

The principal technological decision for the following efforts on a knowledge management is the pair of combined tools. They are multi language dictionaries of terms and thesaurus, which shows the reference system on the terms of the dictionary.

The INIS abstracts base initially was developed as international bibliographic system. It was the simplest decision to create the base using the only English language. However, historically, there were several science schools for knowledge formulation in the nuclear science and technology beginning on the early stages of a nuclear project (end of 40s of the last century). There are English-American, Russian, French, German, and some others among them. The special directions were formed in each system by confidential character of investigations. In the same time the specific science schools as well as a terminology and of cause there were used terms and notes of the different national cultures. In the Cold War condition these differences were revealed as opposition to English-language and Russian –language schools.

Finally, speaking of the knowledge systems in nuclear science and technology, in my mind, there are some versions of this knowledge. This is knowledge that have developed on the base by expert experience of different cultures, namely English-language, French-language, German-language, Russian-language, Japan-language ones.

More significantly these differences are revealed not in terms but rather in joint notions and methods. For example, in Russian-language literature the name “polinoms of Chebyshev” are disseminated, and they are almost non-used in English-language science vocabulary. They use actually the term “mass crossing” in parallel with term “corrosion”, and so on. Technological isolation of the former USSR countries in COCOM conditions has led, for example, to the following situation: the experts of Russian-language science schools for their nuclear physics accounts have developed the analytical methods for complicated dynamic systems. They did it

through simplicity, approximation, and one-measured approaches and so on. Numerical decisions by high-speed computers like it were done by USA and Japan experts were not provided. This was revealed through multi group constants' using, and in development of libraries with data on cross-sections and so on. There are enough the same examples.

At the creating the bibliographic INIS system there were not necessity to consider the specific different schools and terminology because of the result of finding in the data base was presented as a special abstract. This abstract was considered as keeping item, namely essence, object.

However, the specific aspects are manifested it if we work with joint notions, namely with knowledge.

In this connection at the knowledge keeping and management task formulating it is necessity from the initial stage to provide the corresponding lingual tools, which could allow considering above noted differences. The practice of using by modern finding Internet machines (GOOGLE, YAHOO and so on) shows that INIS approach (identification through thesaurus words as indicators but not the words of article finding) is only acceptable in case if we find the component (for example, to select the information on water corrosion for manganese steels).

In addition, principally new point in the knowledge management system in comparison of any библиографической system is necessity to describe and to find the formulas, numeric tables, diagrams and other graphic images.

The noted differences in classified approaches national investigations as it was before are displaying during the student' education in the universities. Without details it can be established that the term finding in the full text information systems, but not words, requires deep knowledge of a language by which these term are formulated. The situation begins to be more complicated by non- comparison of polisemies in different languages. For example, there are names of colors to describe the rainbow in the sky are differing in English and in Russian. Obviously, the index problems at the formal description of the field area (knowledge formalization), and the following finding of the essences the lingual aspects will appear more and more. The modern version of the multi language INIS dictionary is a tool we need to create a multi language INIS thesaurus because the hierarchy system of an English language thesaurus is constructed on the formal reference using to the words of Basic English dictionary by their unique number in a fixed list. The installation of this reference system into the Russian language part of a multi language dictionary gives a possibility automatically to create an authentic thesaurus of INIS terms for Russian language by the words Russian language part of a dictionary.

The same situation could be applied to any other language, for example to Georgian.

Obviously, at this approach to thesaurus development the system of keeping objects' index is totally storied (in the present time — abstract the only) at the crossing from English-language tools of description to the Russian-language ones.

It allows, first of all, the possibility to organize a finding of objects in the keeping base by any language for query none depended on the language for created index.

Namely, query formulated by Russian-language user to the INIS data base (in English) will be automatically realize in full volume if a simple translator will placed under the words from English-language part of the dictionary instead Russian ones. The special number in the dictionary could realize it.

However, historically it was the situation where the Russian-language dictionary and thesaurus of the Information System for Russian National INIS Center are in Russian (SARI – The System of Computer Information Distribution). Minatom created this system in the 1980s years

of the last age. It was developed without synchronize to English-language version of thesaurus, and in fact, was improving separately up to 2000. It has driven to an absence of a real possibility to use the Russian- language terms for finding in the English-language INIS data base.

The efforts for developing the synchronize version of English-Russian- language dictionary was initiated by IAEA INIS Section' experts at the Russian National INIS Center in 2001. In parallel it was created the Georgian-Russian dictionary of nuclear safety terms under frame of Russian and Georgian experts' collaboration. At the present time, there are efforts on Georgian-English-Russian thesaurus development.

The multi lingual problems at the knowledge management presentation were discussed jointly with INIS representatives and Members of the Special Committee included the NIS countries during the 6th Meeting of NIS Committee on peaceful using of atomic energy (April 14, 2004). The Meeting Order pointed the importance to activate the efforts on access to the nuclear-technological knowledge for the national experts. In the present time, the Workshop Team is established. This Team is in progress to provide the Report on abilities and needs have countries-participants for nuclear-technological knowledge management.

The main challenge here is non-ability to financial supporting of these efforts in a full volume.

Above pointed lingual and national characteristics are displayed at the using of INIS bibliographic data base. At the developing of the knowledge management system it will be displayed more and more. On this base the direction and selection of the priorities for these efforts should consider the lingual problems of the potential users.

In the simple case it could be strong rules for classification and catalogs creation of the field area (as it was done early by Karl Linney) on a base of hierarchical thesaurus model. It is necessity the developing of the object-oriented knowledge presentation system in the full volume (in form as aggregate of meta-descriptions for significant components: text, graphics, formulas and so on).

In reality, the existing abstracts INIS data base system particularly covers the first stage of knowledge management task, namely, it gives possibility to get an accordance between finding image and the document' title where this image is. It could be realized through results of finding by key words. However, the full task of knowledge management proposes that a user will get not only reference to the original information as a result of query for the specific date base but in addition he will get some essence. In other words the data base should be not only catalog with abstracts but rather encyclopedia.

At the present time, the content of those objects and tools for manipulation for them is widely discussing by the printed sources. In particularly, as base tool for funding and access they propose the XML language (eXtensible Markup Language) that allows to locate the texts and their fragments (objects) by such manner that a query to a system would bring a queried data for user. (<http://www.w3.org>).

The special language tools for several field areas as mathematics — MathML (<http://www.w3.org/math/>), chemistry — CML (<http://www.xml-cml.org>), biological information science — BSML (<http://www.visualgenoms.com>) are developed to the moment.

Specially, we pay attention to the efforts of American Standard and Technology Institute (NIST USA) on the development of material properties description language - MatML (Sturrock, C.P., Begley, E.F., and Kaufman, J.G., "NISTIR 6785 MatML - Materials Markup Language Workshop Report," National Institute of Standards and Technology, Gaithersburg, MD, August 2001).

Conceptually, Workshop Team on meta-data (OCLC/NCSA Metadata Workshop, 1995) creates approaches to macro-description tools' development (meta-data). These approaches are good studies and concentrated in so-called Dublin Core. The Core includes 13 key notions that should be placed in the document, which describes the knowledge.

In the current October the next international DC conference will be hosted in China.

<http://dc2004.library.sh.cn/english/prog/index.htm>).

It is expediently, taking account above, to develop anagogic requirements to the tools for description and presentation of knowledge for the tasks of knowledge management in the nuclear science and technology area by IAEA. At the initial stage we propose to describe restrictions and requirements for the multi lingual presentations to provide the active using these data by experts – IAEA speakers. Also it is useful to take as basic system of the field area the INIS classificatory and thesaurus, and, naturally, to take abstracts data base of INIS.

It is necessity to complete the term structure of thesaurus for more effective knowledge management. It is necessity to add for term the system of horizontal relations displaying the synonyms (multi lingual, the only), associative relations, and to describe the commentary (interpretation) for every term. We think that it is more difficult to form the synonyms lists because the synonyms of different languages could be differ. For example, it is known that to describe the big man in Russian they actually use synonym “brow” and in English they use the synonym “nose”.

Obviously that to present the knowledge in nuclear science and technology area it is necessity to make structures for all field area as notes (knowledge elements), to develop the similar requirements for these elements' description, and to form the models of finding these elements in the bases. We should organize the finding with using of the tools that are constructed considering the national and lingual characteristics of knowledge.

Access Restrictions for Nuclear Knowledge

Nuclear knowledge in a signified step has a corporate character. Obviously, the access to them should be regulated by owner of data base, and, first of all, they should be open for representatives of countries that created these data base. Also, the principals fixed by Nonproliferation Nuclear Weapons Orders and by other documents should work in this connection.

Commercial Aspect of Nuclear Knowledge

One of the most difficult tasks where it is necessity to take a decision at the efforts for keeping the nuclear knowledge is commercial value of them. At the present time, in knowledge complex it could be separated obviously several objects that are used effectively by different commercial communications.

First of all it is data base with properties of matters and materials. (The example of the effective using is STN). Also, it should be displayed computer applications and their algorithm descriptions (best practice). There is FAAE system of standard and reference data that includes information about properties of matters and materials in the nuclear science and technology area in Russia. This system structure includes 17 special Centers supporting the data bases about different properties of matters (as well as International Center of Nuclear Data).

Obviously, the next getting the same data is practically impossible because the investigations were completed. Obviously also, that these data are significant part of discussed knowledge. The analysis of structure for these data allows showing that we can allocate as commercial part as common science part. The significance of this or that characteristic can not be considered as

commercial secret because it can not directly to be in commercial using. The significance makes to be considered as an object for sale in that case only if its reliability and error are known. However, for the science purpose the significance is the subject of interest because allows to construct the models, to bring hypothesis and so on.

It is simple to separate a commercial value and science value in almost any object of a technological knowledge. In the same time, it is methodologically important to provide the knowledge verification, namely, to manage of only that knowledge which we could trust by opinion of experts, and they could not deceit of potential user. We should identify accurately the goals and tasks of activity on the knowledge management, and so we can avoid conflicts with author' rights because the using of knowledge are considering with scientific and educational purposes.

Purposes and tasks

At a conclusion the purposes and tasks of Russian National INIS Center under the knowledge management task in collaboration with IAEA could be formulated as the following:

- (1) A participation in the IAEA Workshop Team on creation and synchronization with IAEA' INIS
 - of lingual tools for queries to knowledge data providing,
 - of developing the requirements for the structure and formats for knowledge presentation (multi lingual thesaurus, dictionary of interpretations, new classificatory and so on).
- (2) A forming the Project under the NIS Committee on peaceful using of atomic energy IAEA to provide the base knowledge and to provide the corporate access to them.
- (3) A participation in IAEA Projects connected with knowledge using, in particularly, under the INPRO Project, first of all to participate in the knowledge forming about fast reactor technology (Russia, Kazakhstan, Georgia, Ukraine, and Belarus).