Improving compatibility of terminological collections with a bridging classification of data categories

IGOR KUDASHEV
University of Helsinki

INTRODUCTION

In spite of the existence of a standardized set of data categories (ISO 12620:1999) and standardized methods of terminology work (ISO 704:2000), it is still hard to find two terminological databases created in two different organizations which could be easily merged without loss of data or distortion of the original principles of compilation of each of the databases. The most obvious reasons for that are as follows:

- national languages and traditions differ;
- LSP domains differ;
- backgrounds and approaches of the compilers differ;
- technical solutions differ.

At the same time, one of the trends in the management of terminological resources and other kinds of data has been towards the aggregation of scattered resources into bigger portals or services, usually without merging them physically. One of the examples is the EuroTermBank portal (http://www.eurotermbank.com) which provides search in multiple internal and external terminological databases and can make a compilation of the results.

As users' demands and volumes of the data grow, it becomes necessary to provide advanced search which covers all fields of the entry and not just headwords and to tailor entries according to users' preferences. In this article, we propose a classification of data categories which can serve as a bridge between terminological collections. The classification addresses the following problems related to the management of terminology collections with different sets of data categories:

- aggregation and merging of terminological resources;
- organization of search which covers all fields of the entries;
- tailoring the entries from multiple collections in accordance with users' preferences.

Terminological database may contain different types of data in addition to terminological data proper, such as information about sources, users and terminology management transactions. In this article we focus on the classification of data categories related to the description of LSP expressions.

WHAT IS A DATA CATEGORY?

Different types of data are placed in different fields in terminological databases. Data category is the result of specification of a given data field (ISO 1087-2:2000: 13). Terminologists use different metaphors in order to explain the concept of a data category. One metaphor is that of a wardrobe where different drawers are used to store different types of clothes. Another metaphor is a shopping basket where each kind of goods is wrapped in its own package so that they do not get mixed.

TYPICAL MISMATCHES BETWEEN DATA CATEGORIES

Data categories are the result of data classification. Data can be classified in many different ways depending on the views of the classifier and the needs of the users. Using the wardrobe metaphor one can say that different wardrobes come with different number of drawers which may be of different shape, size, etc.

At least six types of mismatching are possible between data categories in terminological databases:

- mismatch of the names of data categories;
- mismatch of the "sizes" of data categories, i.e. different "granulation";
- mismatch of the "places" of data categories, i.e. their location in the classification scheme;
- mismatch of the classification principles (overlapping);
- mismatch of the contents of data categories;
- mixed cases.

Below are a few examples which illustrate some common mismatches.

Case 1: data categories are named differently.

Example: data category is called *note* in one database, *comment* in another, *remark* in the third one and *NB* in the fourth one.

Case 2: the name of a data category is used in different senses.

Example: data category *synonym* may correspond to "full synonym", "near-synonym" or "full or near-synonym".

Case 3: names of data categories are "false friends".

Example: English *abbreviation* and *acronym* as defined in the ISO standard (ISO 12620:1999: 6–7) and Russian *аббревиатура* and *акроним* correspond to each other crosswise (Kudashev & Hajutin 2003: 104–105).

Case 4: granulation of data categories is different.

Example: data category *abbreviated form of term* is split into five subclasses (*abbreviation*, *short form of the term*, *initialism*, *acronym* and *clipped term*) in ISO 12620:1999, but there is no such division in ISO 12616:2002 ("Translation-oriented terminography").

Case 5: overlapping of data categories.

Example: data categories *example* and *context* overlap. Some examples are contexts and some contexts may serve as examples, but the two categories are not identical.

Case 6: the same data category is put under different broader categories.

Example: data category *context* is considered concept-related data in ISO 12620:1999, apparently because contexts may provide additional information on the concept. However, a more common function of contexts is to provide information about term usage and collocations, so "many databases classify *context* as a term-related category" (ISO 12620:1999: 25).

Case 7: language, sign system or notation of the data differ.

Example: a part of speech may be coded as *noun* in one database, *n.* in another, *subst.* in the third one and as a graphical symbol in the forth one.

Case 8: identical or similar values of data categories are used in different meanings and different connections.

Example: value *neologism* in the ISO 12620:1999 *term provenance* data category sounds like a chronological label while in fact it refers to the methodology employed in creating the term.

Case 9: values of semi-closed classes differ due to different classification of linguistic phenomena in different languages and different traditions.

Example: in ISO 12620:1999 labels for describing LSP expressions belonging to the "lower style" include *slang register* and *vulgar register*. They roughly correspond to labels *профессиональный жаргонизм* (professional slang) and *профессиональное просторечие* (professional colloquialism) in Russian. In Finnish, however, there is only one category, *ammattislangi* (professional slang; see Sanastotyön käsikirja 1989: 12).

It should also be noted that some data may be expressed implicitly in terminological collections.

Example 1: the presence of the label *obsolete* in some cases and its absence in others in a terminological collection implies that terms which are not marked with the label, belong to the active LSP stock.

Example 2: if two or more terms are put in the same entry, this usually implies that they are synonyms or equivalents.

Data exchange or aggregation of terminological resources may require that such implicit data is made explicit. For example, if data is stored in the form of ontology rather than static entries, then implicit information about synonyms and equivalents mentioned above has to be made explicit and saved during the "dissembling" of the entries.

ISO INVENTORIES AND CLASSIFICATIONS OF DATA CATEGORIES

ISO 12620:1999 specifies a set of data categories for recording terminological information. It does not prescribe what data categories should be used but rather serves as an inventory. This set of data categories is being broadened in the *ISOcat* project (www.isocat.org) which documents widely accepted linguistic data categories.

While the *ISOcat* project is undoubtedly useful for linguists who get a chance to better categorize and define linguistic concepts and for designers of term banks who can pick ready data categories from a vast inventory, it is unlikely to bring relief to the problem of data exchange, aggregation and full-entry search in multiple databases.

Firstly, it can't prevent compilers of terminological products from

classifying data in a different way and from using their own data categories. Secondly, the more data categories are used in terminological collections, the more diverse they become, making data exchange even more challenging.

The list of data categories is infinite in theory and quite vast in practice. For example, the list of data categories related to terminology in the *ISOcat* inventory already exceeds 500 items. One of the practices used in the *ISOcat* project can increase this number many times. We refer to presenting values of closed and semi-closed classes as individual data categories.

For example, values of the data category *register*, such as *vulgarRegister*, *slangRegister*, etc. are now presented as data categories in their own right. Likewise, data category *reliabilityCode* is split into *reliabilityCode1*, *reliabilityCode2*, etc. There can be different opinions whether this is reasonable from the practical point of view, but data can be organized in many different ways, so there are no formal restrictions for splitting data categories all the way to the primitive classes which can only accept values *yes* and *no*. It is clear, however, that this realm of data categories needs proper structuring and classification in order to remain manageable and well-organized.

Meanwhile, classification of data categories proposed in ISO 12620:1999 is problematic in several respects. To start with, there are some inconsistencies concerning the principal division of the data. According to section 6.2 (*Typology of data categories*), data categories are divided into three major groups: term and term-related information, descriptive data and administrative data. However, in *Annex D (Systematic listing of data categories)* the second group is called *Data categories related to concept description*. At the same time, this group contains subgroup *Note* which "stands alone because it can be associated with any one of the other categories and therefore cannot be subordinated to any other specific subgroup" (ISO 12620:1999: 4).

If we assume that the intended division included four groups: term and term-related information, concept-related information, administrative data and *Note*, this classification still raises many questions. To name a few:

- Why *examples* and *contexts* are concept-related data and not term-related data? Cf. the description of the *context* data category: "A text or part of a text in which a term occurs" (ISO 12620:1999: 25).
- Why *synonym* and *equivalence* are term-related data while otherwise everything related to the meaning is concept-related data? Cf. description of the data category *degree of equivalence*: "The extent to which the intensions of two or more concepts overlap" (ISO 12620:1999: 21).

- Why *antonym* and *homograph* are administrative data and not term- or concept-related data?
- Why *audio*, *video*, etc. are (just) concept-related data?
- What is the exact definition of administrative data? Why this class includes such heterogeneous categories?

Division of terminological data into concept-related and term-related may be useful from the technical point of view because it supports the concept-oriented approach which reduces the number of relations between terms by linking synonymous terms to the same concept. However, such division in general and its implementation in ISO 12620:1999 in particular may present a challenge for common users of terminological databases – translators and domain experts. Indeed, it is not easy to comprehend why *synonyms* should be searched in term-related data, but *antonyms* in administrative data and *examples* in concept-related data.

Common users work with terms – words and word combinations, so for them it is more natural to speak about the meaning of a term, synonyms of a term, examples of term usage, etc. This implies that classification of data categories aimed at providing full-entry search in multiple collections should be term-oriented, intuitively clear and based on common linguistic categories.

BRIDGING MISMATCHES BETWEEN DATA CATEGORIES

The problem of mismatches between data categories can be solved with the help of mapping. If mismatches are nominal, for example, if names of data categories differ or the same contents are presented differently, direct mapping between data categories can be used. Once the direct mapping is applied, the data in the aggregated resource can be searched with the same precision as in the original databases.

More substantial differences between data categories require finding a common denominator through the mapping between *classifications* of data category sets. The use of a common denominator somewhat decreases the precision of the search, but this is the only way to provide common search for terminological resources with different structure. If the user is not ready to sacrifice the precision of the search, he has to perform separate search in each individual collection.

As with data exchange in general, using an intermediate format, i.e. some bridging classification of data categories, is more effective in the long

run than multiple mappings between different classifications. The general principles of the classification of data categories intended for serving as a common interface between other classifications are as follows:

1. Since classes of the classification are supposed to serve as common denominators, they have to be on a higher level of abstraction than most "primitive" data categories which are not subdivided further. At the same time the level of abstraction may not be too high because the users would not be interested in a classification which is too general. In practice a classification with two levels of abstraction is sufficient.

2. The classification should cover all types of LSP expressions which are typically described in terminological databases, including term elements, proper names, nomenclature, set phrases, etc.

3. The classification should contain only those data categories which are directly related to the description of LSP expressions. Description of technical and administrative aspects (e.g. medium, encodings, sources, reliability, etc.) is a different task.

4. The classification should be hierarchical, but several bases of division should be allowed on the same level of abstraction.

5. The classification should be extensible, i.e. it should contain category "other" on each level of the hierarchy.

These principles have to be combined with the requirement of user-friendliness mentioned above. In our opinion classification based on linguistic functions of the data is a particularly strong candidate in the latter respect.

In the bridging classification based on the linguistic functions the focus is shifted from the names of data categories to the function or functions of the data they contain. A comparison which can be made here is a library with its alphabetic and subject directories.

Search by the names of data categories is similar to the search with the help of an alphabetical directory. The title gives some clue about the contents of the book, but sometimes it can be misleading. Besides, users cannot guess all the possible titles which cover a certain topic. Alphabetic directory is useful in cases when the users know for sure what item(s) they want to locate. Most library users, however, start their search with the subject directory because they do not know beforehand what items cover the topic they are interested in.

Terminological portal which aggregates multiple terminological collec-

tions with different sets of data categories poses the same problem to users as the realm of books on the shelves of a library. Users know what kind of information they are interested in (grammar, meaning, usage, etc.), but they do not necessarily know in what data categories this information can be found. In different terminological collections similar types of data may reside in different data fields. For example, information about the areal status of an LSP expression can be found in such fields as *usage*, *regional label*, *language symbol*, etc.

If the search is based on the linguistic functions of the data, users don't need to care about the exact names of data categories. They just specify that they are looking for information about areal status, meaning or synonyms, and the terminology management system locates and displays full or abridged entries in different databases which contain the specified type of information. Search by the function may and should be complemented by search by the name of the fields for those users who know exactly in what fields they want to search.

Data fields, like library items, may contain different kinds of linguistic information. For instance, *example* is a typical multifunctional field which may contain information about form, meaning and usage of a term, in different proportions. Each data category may be described as having one or several functions. It should also be possible to specify the degree to which a particular data category reflects certain functions. This characteristic may be verbal (e.g. primary and secondary function) or numerical (e.g. 0–100%).

The easiest way to do the mapping of a particular data category set to the intermediate classification is to consider the contents of data fields of a particular type uniform and make a simple table of correspondence. A more precise but also more complicated mapping would allow the compilers to specify deviations of individual data fields from the values used in the global table of correspondence. For example, contexts usually provide information about term usage and meaning. It is reasonable to include these types of information in the global table of correspondence as they pertain to every context field. However, some contexts may also contain encyclopedic information. This occasional use of encyclopedic information may be marked locally, on the field level.

CLASSIFICATION OF DATA CATEGORIES BASED ON LINGUISTIC FUNCTIONS OF THE DATA

Since LSP expressions are linguistic signs, information about them can be divided into information about their form, meaning, usage, relations to other units, origin and development. Below each of these classes is described in detail and some examples of the data belonging to them are provided. Please see *Appendix 1* for the compact version of the classification.

Data related to form

LSP expressions have two forms – written and oral. Besides, data related to the form of an LSP expression may be subdivided into three classes:

- data related to the canonical form;
- data related to the formation of the unit;
- data related to the inflection of the unit.

DATA RELATED TO CANONICAL FORM

Canonical form is the form in which the headword is given in the database. It serves as "representative" for other forms. For example, in most European languages the canonical form for nouns is nominative singular and for verbs infinitive presence. However, rules for choosing canonical form vary in different languages and lexicographical traditions.

Here are a few examples of the data related to written canonical form:

- type of expression by its written form (e.g. full form, abbreviated form, symbol, formula, etc.);
- spelling of the form;
- spelling variants of the form;
- hyphenation.

Examples of data related to oral canonical form:

- type of expression by its oral form (e.g. initialism, acronym);
- pronunciation;
- pronunciation variants of the form;
- syllabification.

As one can see, some data categories may relate to both written and oral forms. For example, indicating that a particular expression is an initialism, i.e. an abbreviated form made of initial letters of the full term, in which these letters are pronounced individually (e.g. United Nations – UN), provides information about both written and oral forms. Besides, it may

be said to provide information about term formation, origin and development.

DATA RELATED TO TERM FORMATION

Examples of the data related to the formation of LSP expressions:

- term components and term elements;
- morpheme structure;
- syntactical model;
- model of term formation;
- method(s) of term formation;
- morphological variants;
- derivatives.

DATA RELATED TO INFLECTION

Examples of the data related to the inflection of LSP expressions:

- grammatical parameters (number, gender, animacy, etc.);
- complete or partial paradigm of the forms;
- models of inflection, conjugation, etc.

Data related to meaning

This part of the classification starts with a rather marginal category in order to keep in line with other sections. LSP expressions may be divided into several classes depending on how much their inner form reflects their meaning. For example, a set expression is a unit, the meaning of which cannot be deduced from the combined sense of the words making up the expression.

Apart from that, data related to the meaning of LSP expressions can be divided into two big categories. The first one is *logical meaning* – description of the logical concept denoted by an LSP expression. This type of information is usually provided in terminological definitions and definition-like descriptions.

The second category includes rather heterogeneous components of meaning which can be called *induced meaning*. These components include, for example:

- different connotations, i.e. evaluative components of the meaning;
- inner form of the expression (its "literal", morpheme-by-morpheme meaning);

- other LSP or LGP meanings of the same expression;
- components of meaning resulting from antonymous, synonymous, paronymous and other systematic relations of the expression;
- different kinds of associations;
- components of meaning resulting from consonance, rhymes, etc.

We call these components of meaning "induced" because they result from the attitude of language users towards the objects denoted by LSP expressions, from associations of the users or from relations of an LSP expression with other language units. In this sense additional components of meaning are "induced" on the LSP expression by language users, other language units or both. For more information about the presentation of induced meaning in LSP dictionaries see (Kudashev 2006; Kudashev 2007: 254–258).

Induced components are welcome and even cultivated on purpose when they create positive associations and connotations or allow users to express their attitude to the subject in informal communication. However, in most cases they only distract the users' attention from the logical meaning which is supposed to be at the core of LSP communication. This is probably one of the reasons why components of induced meaning have to a great extent been neglected in terminology theory. However, taking them into consideration is an important prerequisite for successful terminological nomination and effective LSP communication.

Data related to encyclopedic description

Encyclopedic description provides information about objects denoted by LSP expressions, and in most cases this information is extralinguistic. However, sometimes it is not easy to draw the line between the description of the concept and the description of the object denoted by an LSP expression. Many terminological databases already contain information which accounts as encyclopedic description, and the share of such information is expected to grow in the future as different types of reference products tend to draw closer to each other (cf. Hartmann 2001: 5). To acknowledge this fact we have decided to include encyclopedic description into the classification of the data categories related to the description of LSP expressions.

Data related to usage

Information related to usage appears to be the most asked-for type of

terminological data (Kudashev 2007: 207). It can be subdivided into two big categories:

- restrictions in usage;
- frequency of use.

DATA RELATED TO RESTRICTIONS IN USAGE

Usage of any LSP expression is restricted to at least some national language, domain and chronological period. In addition to this usage may be restricted to certain geographical area, professional group, organization, register, etc. Below is the list of the most frequent restrictions of usage with a few examples in brackets:

- national language (en, fi, ru);
- domain and subdomain (physics – atomic physics – high energy physics);
- scientific school/theory (Newton's physics, Einstein's physics; Danish structuralism in linguistics);
- chronological restriction (obsolete, neologism, used during WWII);
- geographical restriction (en-US, en-GB, dialect expression);
- organizational restriction (term used in/by Nokia, Microsoft, UN, WHO);
- proprietary restriction (trade mark, trade name);
- register restrictions (official term, informal term, professional slang);
- professional group restrictions (physicians, nurses, medical assistants);
- combinatory restrictions (to carry out/conduct/make/launch an investigation);
- compliance restrictions (standardized, preferred, recommended, non-recommended term).

The two latter types of restrictions probably need to be commented on in more detail.

Data related to combinatory restrictions

Combinatory power is the ability of linguistic units to form bigger units. Combinatory power can be divided into semantic, lexical and syntactical. Semantic combinatory power suggests that expressions do not have controversial components in their meaning. Lexical combinatory power manifests itself in the ability of expressions to combine with certain other lexical means. Syntactical combinatory power is the ability of an expres-

sion to combine with certain grammatical forms of other expressions and auxiliary words. Combinatory restrictions are usually described with the help of syntactical models, examples and contexts.

Data related to compliance restrictions

Terminological products tend to be more or less prescriptive in nature. This normativeness may range from recommendations by the compilers or domain experts whom they have consulted to normative authorization. In any case the question is about the compliance of an LSP expression with a "good" or "correct" style (from the compilers' point of view), and in the case of standards – also with some normative document.

DATA RELATED TO FREQUENCY OF USE

Information about the frequency of use may be based on corpus evidence and expressed numerically or it may be a more or less subjective estimate expressed verbally (e.g. commonly used – infrequently used – rarely used).

Data related to systematic relations

By systematic relations we mean ontological relations (relations containing knowledge about the world), systematic lexical relations and cross-language equivalence relations. Below are examples of the most common systematic relations:

- synonymous relations;
- antonymous relations;
- homonymous relations;
- paronymous relations;
- generic relations;
- partitive relations;
- non-hierarchical ontological relations (associative, sequential, temporal, causal, etc);
- (cross-language) equivalence relations.

Relations pertaining to word formation, inflexion and combinatory power of expressions do not belong to this category because they are not systematic *lexical* relations.

Data related to origin and development

This type of data is close to etymological data but not limited to it. It may include, for example:

- information on the forms from which a specified LSP expression is believed to originate;
- information on the stages of development of a specified LSP expression;
- information on earlier form(s), meaning(s), usage, etc. of a specified LSP expression.

CONCLUSIONS

In this article we have discussed the principles of data classification aimed at bridging structural mismatches between data categories in different terminological collections. Intended applications of the classification include aggregation and merging of terminological data, organization of full-entry search in multiple terminological collections and tailoring the entries in accordance with users' preferences.

Two latter tasks require that the classification should be rather simple and intuitively clear to common users of terminological products. Classification based on linguistic functions of the data seems to be one of the best candidates in this respect. Search based on the linguistic functions of the data effectively supplements the search by the names of data categories and has the same function as subject directory in a library.

In this article we have focused on the classification of data related to the description of LSP expressions as it is the most important type of data presented in terminological databases. Classifications of other types of data, such as information about sources, users and terminology management transactions, will be available in the specifications of the *ContentFactory* project which is aimed at designing an ontology-based platform for distributed collaborative terminology work. The project ends in 2010 and a large part of its internal documentation will be made public.

BIBLIOGRAPHY

Hartmann R. R. K. 2001: *Teaching and Researching Lexicography*. New York: Longman.

ISO 1087-2:2000 *Terminology Work – Vocabulary. Part 2: Computer Applications*.

ISO 12616:2002 *Translation-Oriented Terminography*.

ISO 12620:1999 *Computer Applications in Terminology – Data Categories*.

ISO 704:2000 *Terminology Work – Principles and Methods*.

Kudashev I. 2006: Additional Meaning Components of Terms and their Treatment in LSP Dictionaries. – Lehtinen, Esa – Niemelä, Nina (toim.) *Erikoiskielet ja käännösteoria. VAKKI-symposiumi XXVI. Vaasa 11.–12.2006*, Vaasa: Vaasan yliopisto, 143–149.

Кудашев И. С. 2007: *Проектирование переводческих словарей специальной лексики*. Хельсинки: Yliopistopaino.

Кудашев И. С., Хаютин А. Д. 2003: К вопросу о формах терминоэлементов и терминов. – *Лексикология. Терминоведение. Стилистика*: Сборник научных трудов, Москва, Рязань, 101–107.
Sanastotyön käsikirja 1989: *Sanastotyön käsikirja. Soveltavan terminologian periaatteet ja työmenetelmät* / Toimittanut Tekniikan sanastokeskus. Helsinki: Suomen Standardoimisliitto.

APPENDIX 1. CLASSIFICATION OF TERMINOLOGICAL DATA BASED ON ITS LINGUISTIC FUNCTIONS

Values in angle brackets are candidates for the formal representation of the categories. These values are used in *Appendix 2* for the sake of brevity.

1. Data related to form. *<termForm>*
1.1. Type of LSP expression by its form. *<termForm: termType>*
1.2.1. Data related to written form. *<termForm: writtenForm>*
1.2.2. Data related to oral form. *<termForm: oralForm>*
1.3.1. Data related to canonical form. *<termForm: canonicalForm>*
1.3.2. Data related to term formation. *<termForm: termFormation>*
1.3.3. Data related to inflection. *<termForm: termInflection>*
2. Data related to meaning. *<termMeaning>*
2.1. Type of LSP expression by correspondence of its meaning to its form. *<termMeaning: termType>*
2.2. Data related to logical meaning. *<termMeaning: logicalMeaning>*
2.3. Data related to induced meaning. *<termMeaning: inducedMeaning>*
3. Encyclopedic description. *<termEncyclopedicDescription>*
4. Data related to usage. *<termUsage>*
4.1. Data related to restrictions of usage. *<termUsage: restrictions>*
4.1.1. Data related to national language restrictions. *<termUsage: restrictions: language>*
4.1.2. Data related to domain restrictions. *<termUsage: restrictions: domain>*
4.1.3. Data related to scientific school or theory restrictions. *<termUsage: restrictions: schoolOrTheory>*
4.1.4. Data related to chronological restrictions. *<termUsage: restrictions: chronological>*
4.1.5. Data related to geographical restrictions. *<termUsage: restrictions: geographical>*
4.1.6. Data related to organizational restrictions. *<termUsage: restrictions: organizational>*
4.1.7. Data related to proprietary restrictions. *<termUsage: restrictions: proprietary>*

4.1.8. Data related to register restrictions. <*termUsage: restrictions: register*>
4.1.9. Data related to professional group restrictions. <*termUsage: restrictions: professionalGroup*>
4.1.10. Data related to combinatory restrictions. <*termUsage: restrictions: combinatory*>
4.1.11. Data related to compliance restrictions. <*termUsage: restrictions: compliance*>
4.2. Data related to frequency of use. <*termUsage: frequency*>
5. Data related to systematic relations. <*termRelations*>
5.1. Data related to synonymous relations. <*termRelations: synonymousRelation*>
5.2. Data related to antonymous relations. <*termRelations: antonymousRelation*>
5.3. Data related to homonymous relations. <*termRelations: homonymousRelation*>
5.4. Data related to paronymous relations. <*termRelations: paronymousRelation*>
5.5. Data related to generic relations. <*termRelations: genericRelation*>
5.6. Data related to partitive relations. <*termRelations: partitiveRelation*>
5.7. Data related to non-hierarchical ontological relations. <*termRelations: non-hierarchicalOntologicalRelation*>
5.8. Data related to equivalence relations. <*termRelations: equivalenceRelation*>
6. Data related to origin and development. <*termOriginAndDevelopment*>
7. Other types of data related to the description of an LSP expression <*termOtherRelatedData*>

APPENDIX 2. EXAMPLE OF MAPPING OF SOME ISO 12620:1999 DATA CATEGORIES INTO THE FUNCTIONAL CLASSIFICATION OF TERMINOLOGICAL DATA

See *Appendix 1* for the decryption of the formal representation of the categories which are used for the sake of brevity. Question mark after a data category means that the presence of the specified kind of data is subject to the interpretation of the data category. If subcategories can be mapped in the exact same way as their parent category, only the parent category is described.

Due to space limitations only the first subgroup of data categories from the ISO 12620:1999 standard is covered.

A.1 term

Note: term (headword) lies outside the scope of classification which covers data categories related to the *description* of LSP expressions. However, technically data category *term* is identical to the written form of the term.

Functional classification: not applicable or *<termForm: writtenForm>*

A.2.1 term type

Note: in the ISO 12620:1999 standard this category includes various kinds of LSP expressions and may contain different types of data. See subcategories for more information.

A.2.1.1 main entry term

Note: information about the structure of the entry is not data related to the description of LSP expressions. However, this category indirectly reflects preference.

Functional classification: not applicable or *<termUsage: restrictions: compliance>*

A.2.1.2 synonym

Note: when opposed to main entry term, this category reflects preference. Its primary function is, however, reflection of synonymous relations.

Functional classification: *<termRelations: synonymousRelation>*; *<termUsage: restrictions: compliance>*?

A.2.1.3 quasi-synonym

Note: same as previous.

Functional classification: *<termRelations: synonymousRelation>*; *<termUsage: restrictions: compliance>*?

A.2.1.4 international scientific term

Note: depending on the interpretation of this category and its contents, this data category may provide information about compliance, language restrictions, origination and development.

Functional classification: *<termUsage: restrictions: compliance>*; *<termUsage: restrictions: language>*?; *<termOriginAndDevelopment>*?

A.2.1.5 common name

Note: common name is a synonym of an international scientific term that is used in general discourse. This data category may provide information about restrictions related to register, compliance and professional group.

Functional classification: <*termUsage: restrictions: register*>; <*termUsage: restrictions: compliance*>; <*termUsage: restrictions: professionalGroup*>?

A.2.1.6 internationalism

Note: depending on the contents this data category may provide information related to language restrictions, term formation, origination and development.

Functional classification: <*termUsage: restrictions: language*>; <*termForm: termFormation*>?; <*termOriginAndDevelopment*>?

A.2.1.7 full form

Note: this data category provides information about the type of term by its form.

Functional classification: <*termForm: termType*>

A.2.1.8 abbreviated form of term

Note: depending on the contents, this data category and all its subcategories (**A.2.1.8.1 abbreviation**, **A.2.1.8.2 short form of term**, **A.2.1.8.3 initialism**, **A.2.1.8.4 acronym** and **A.2.1.8.5 clipped term**) may provide information related to type of term by its form, oral form; term formation, origin and development.

Functional classification: <*termForm: termType*>; <*termForm: oralForm*>?; <*termForm: termFormation*>?; <*termOriginAndDevelopment*>

A.2.1.9 variant

Note: depending on the contents, this data category may provide information about preference, term formation, origin and development.

Functional classification: <*termUsage: restrictions: compliance*>; <*termForm: termFormation*>?; <*termOriginAndDevelopment*>?

A.2.1.10–A.2.1.14 (transliterated form, transcribed form, romanized form, symbol and formula)

Note: these data categories provide information about the type of term by its form.

Functional classification: <*termForm: termType*>

A.2.1.15–A.2.1.17 (equation, logical expression, materials management categories)

Note: these units can not be headwords in terminological databases.

Functional classification: not applicable.

A.2.1.18 phraseological unit

Note: in ISO 12620:1999 this category is split into three subcategories: *collocation*, *set phrase* and *synonymous phrase*. Collocation can not be counted

as a phraseological unit as it does not correspond to the definition of a phraseological unit provided in the standard. *Synonymous phrase* seems to be an unnecessary category. The same information can be expressed with two other data categories: *set phrase* and *synonym*.
Functional classification: not applicable.

A.2.1.18.1 collocation

Note: this data category provides information about combinatory restrictions.
Functional classification: *<termUsage: restrictions: combinatory>*

A.2.1.18.2 set phrase

Note: this data category provides information about the type of term by correspondence of its meaning to its form.
Functional classification: *<termMeaning: termType>*

A.2.18.3 synonymous phrase

Note: as was stated above, this category is probably unnecessary. If used, it provides information about the type of term by correspondence of its meaning to its form and also about synonymous relations.
Functional classification: *<termMeaning: termType>*; *<termRelations: synonymousRelation>*

A.2.1.19 standard text

Note: standard texts can not be headwords in terminological databases.
Functional classification: not applicable.

TERMINIJOS RINKINIŲ SUDERINAMUMO GERINIMAS TAIKANT DUOMENŲ KATEGORIJŲ KLASIFIKACIJĄ

Viena iš informacijos ir ypač terminologijos išteklių plėtros tendencijų yra atskirų duomenų bazių jungimas į didelius portalus ir sąsajų tarp tokių portalų kūrimas. To pavyzdys gali būti EuroTermBankas (http://www.eurotermbank.com). Didėjant informacijos apimčiai ir vartotojų tokiems produktams keliamiems reikalavimams, auga poreikis užtikrinti išlpėstinę paiešką, apimančią visus žodynų straipsnių laukus, ir galimybę pateikti tik tas informacijos kategorijas, kurios tuo metu domina vartotoją. Šiems uždaviniams spręsti reikalinga duomenų kategorijų klasifikacija, kurią būtų galima taikyti kaip terminologinių bazių, turinčių skirtingą struktūrą, „bendrą vardiklį". Svarbus tokios klasifikacijos reikalavimas – jos paprastumas ir aiškumas paprastiems terminologijos produktų vartotojams. Straipsnyje aptariama lingvistinėmis duomenų funkcijomis paremta klasifikacija galėtų tapti tokiu „bendru vardikliu".

ИСПОЛЬЗОВАНИЕ МЕТАКЛАССОВ ДАННЫХ КАК СРЕДСТВО УЛУЧШЕНИЯ СОВМЕСТИМОСТИ ТЕРМИНОЛОГИЧЕСКИХ КОЛЛЕКЦИЙ

Одной из тенденций развития информационных ресурсов в целом и терминологических в частности стало объединение разрозненных баз данных в крупные порталы и создание общих интерфейсов к ним. Примером может служить ЕвроТермБанк (http://www.eurotermbank.com). По мере роста объемов информации и требований пользователей к подобным продуктам растет потребность в обеспечении расширенного поиска, охватывающего все поля словарных статей, а также возможности отображения лишь тех категорий информации, которые в данный момент интересуют пользователя. Эти задачи требуют наличия классификации информационных категорий, которая могла бы служить «общим знаменателем» для терминологических баз с различной структурой. Немаловажным требованием к подобной классификации является ее простота и понятность для простых пользователей терминологических продуктов. В статье описывается один из возможных кандидатов на роль «общего знаменателя» – классификация, основанная на лингвистических функциях данных.

Gauta 2009-10-14

Igor Kudashev
University of Helsinki
Palmenia Centre for Continuing Education
P.O. Box 239 (Paraatikenttä 6)
FIN-45100 Kouvola, Finland
E-mail: igor.kudashev@helsinki.fi