Domain Classification for Terminology Management: Principles of Compilation

Igor Kudashev
Palmenia Centre for Continuing Education
University of Helsinki

В больших терминологических коллекциях, содержащих лексику из разных предметных областей, отраслевые пометы являются одним из наиболее важных типов информации. Еще в 1980-х гг. многие исследователи пришли к выводу, что библиотечные классификации и информационно-поисковые тезаурусы не подходят для целей классифицирования терминологии. В связи с этим широко обсуждался вопрос о необходимости создания классификации предметных областей, которая была бы адаптирована для целей терминологической работы. Тем не менее нам не удалось найти подобной классификации в открытом доступе. В проекте TermFactory было принято решение разработать классификацию предметных областей, ориентированную на потребности коллоборативной терминологической работы. В статье обсуждаются принципы создания подобной классификации.

Keywords: domain classification, classification system, terminology management, terminology management system, term bank

1 Introduction

Users consider domain labels one of the most important data categories in LSP collections, especially multi-domain ones (Kudashev 2007: 294). In the 1980s, the need for a domain classification specifically designed for the purposes of terminology management has been discussed actively (see Nedobity 1988), but such a classification seems to have never been created or at least made public.

In the *TermFactory* project which is aimed at creating a platform and a workflow for collaborative ontology-based terminology work, it has been decided to create a dedicated domain classification that would be multilingual, generic (not too culture and language-specific), based on widely acknowledged divisions, simple, user-friendly and extensible. In my paper, I will discuss the reasons for using a domain classification in a term bank, consider some general problems in the compilation of domain classifications, and describe the principles of compilation of the *TermFactory* core domain classification.

2 Overview of the TermFactory Project

TermFactory (further on TF) is a part of a larger project *ContentFactory* (2008–2011) carried out at several departments of the University of Helsinki and Aalto University. The project is financed by the Finnish Funding Agency for Technologies and Innovation (Tekes) and a number of language industry companies.

TermFactory aims at creating an architecture and workflow for collaborative ontology-based terminology work. The architecture of the TF platform includes three components:

– A schema for representing terminologies as OWL ontologies.
– Software for distributed storage and retrieval of ontologized terminological resources.
– Web services that provide methods to query, discuss and edit terms on various collaborative platforms.

Another major outcome of the *TF* project is a *Quality Manual of Collaborative Terminology Work* that proposes a workflow and quality assurance instruments for collaborative terminology work. By *collaborative terminology work* we mean terminology work done by a community in a decentralized, Wikipedia-style manner.

3 Reasons for using a Domain Classification in a Term Bank

One of the specific characteristics of a large terminological collection is that it contains terminology from many different domains. There are a number of reasons to indicate the domain to which a particular LSP unit belongs.

Domain labels indicate the area of application of LSP units and also give a clue about their meaning, which is particularly important in situations when proper semantic description, such as definition or note, is not provided. In electronic collections, domain labels play a very important role in disambiguation – selection of the correct headword

from a list of homonyms. Domain classification also allows organizing terminological records thematically and managing them in a systematic way.

In a collaborative terminology management platform, domain classification can also be an important means of managing user rights and roles. As users provide information about their special field competence, it is possible to restrict their editing rights to those records which correspond to their domain(s) of expertise. Even if editing rights are not restricted, other users can see which experts have contributed to the entry and whether their competence corresponds to the domain of the term.

4 Overview of Domain Classifications used in Finland

Domain classifications are used in many areas of application, for example, in statistics, planning, accounting, and classification of publications according to their principal subject. Domain classifications are a part of library classifications, different thesauri and, lately, upper ontologies.

In Finland, the most wide-spread library classification is YSA (General Finnish Thesaurus), although some libraries use other classifications, such as UDC (Universal Decimal Classification) or HKLJ (Helsinki City Library Classification). YSA has been lately ontologized, and its revised and extended version has become YSO (Finnish General Upper Ontology).

Today, many classifications are either translations or localized versions of international classifications originating from different international organizations and consortia. The source language of these classifications is usually English. For example, economic classification used by Statistics Finland is based on the Eurostat's Statistical Classification, which in its turn is based on the United Nations' Classification. Field of Science and Technology Classification used in Finland is based on the Recommendations by UNESCO.

5 What Classifications are used in Existing Term Banks?

As the range of authoritative classifications is rather broad, a question arises which one of them is the most suitable for the purposes of terminology management. The fact that different term banks use different classifications although this complicates the exchange of terminological data (cf. Ubin 1992: 55) implies that there is no simple answer to this question.

Some term banks adopt external domain classifications as such. For example, *IATE* (EU inter-institutional terminology database) uses EuroVoc, which is a multilingual, multi-disciplinary thesaurus covering the activities of the EU and especially the European Parliament. In other term banks, tailored versions of existing classifications have been adopted or even dedicated in-house classifications have been created. For example, Canadian *Termium*, which is one of the biggest and oldest term banks in the world, uses a sophisticated domain classification developed at the University of Montreal (Hutcheson 2001: 670). In smaller term banks, such as Finnish *TEPA* or Swedish *Rikstermbanken*, no domain classification is used. As these term banks are mostly collections of glossaries, the name of the glossary usually plays the same role as the domain label.

6 Problems with Existing Classifications

Researchers who studied the applicability of library and documentary classifications to the needs of term banks in the 1980s (e.g. Nedobity 1988; Lingvističeskaâ koncepiâ 1989: 54), came to the conclusion that existing library classifications and thesauri might provide a good starting point but in most cases they could not be used as such for the purposes of terminology management. Below are provided a few examples of issues that complicate the use of existing classifications in terminology management systems.

Upper level classes of library classifications may refer to several domains, some of which are not even closely related. Such classes can not be used as domain labels as they are too broad (and often too long as well). For example:

– HKLJ, class 630: *Metal industry. Wood processing industry. Electrotechnology. Industry textile. Leather industry*.
– Library of Congress Classification, letter G: *Geography. Antropology. Recreation*.

Quite often classes in thesauri, ontologies and library classifications are themselves *keywords* (terms) rather than *names of domains*. For example, such classes in YSO as *heart*, *ECG*, *myocardial infarction* are *terms* that belong to the domain of *cardiology*.

Many classes in library classifications, thesauri and top ontologies are superfluous from the point of view of terminology management. Examples include:
– Proper names (e.g. YSO contains over 200 names of computer programs).
– Classes of publications by their genre or language in library classifications (e.g. HKJL, class 050: *General periodicals*; subclass 051.1 *Finnish-Swedish periodicals*).
– Abstract classes in ontologies (e.g. *abstract-concrete*, *endurant-perdurant* in YSO).

National classifications are rarely available in more than two or three languages, and they tend to be culture- and/or language-specific, at least partially. For example, YSA contains such culture-dependent keywords related to the Finnish educational system as *lukio* (≈ upper secondary school), *lyseo* (≈ secondary school) and *kansakoulu* (≈ elementary school). In Finnish classifications, one will find *valtiotiede*, which is a partial equivalent of *politology*, or *political science*.

Some thesauri and classifications (e.g. UDC) are aimed at professionals and are too difficult to be used by the general public. Quite often such classifications are not distributed free of charge, and it is difficult to obtain rights to them.

A heavy burden of many classifications and thesauri is version management. For example, a conversion table between two minor versions of EuroVoc is over 200 pages long.

7 Requirements to the Domain Classification

As domain classification is a very important instrument of quality assurance and role management in a collaborative system, in the *TermFactory* project it has been decided to create a dedicated domain classification for the purposes of multilingual, multi-user collaborative terminology work. To suit the needs of such work, domain classification should ideally meet the following requirements:

1. It should be free and available online 24/7, which in practice means that it has to be a part of the platform rather than a separate resource.
2. Classification should be multilingual.
3. The categories in the classification should be widely acknowledged.
4. Classification may not be too culture-specific.
5. Classification should be user-friendly and have simple organization and notation rules.
6. Classification should be extensible, i.e. there should be a possibility to add new classes if they are missing from the classification vocabulary.
7. Classification should have mechanisms of version management, so that older data could be made compatible with later versions.

8 General Problems of Compilation of Domain Classifications

While working on the principles of domain classification in TF, we have identified some common problems in the compilation of domain classifications. The first major problem is multiple alternative bases for classification. Domains can be classified in many different ways. For example, *Astronomy* can be classified according to the physical bodies that are the objects of study (solar / stellar / galactic astronomy, etc.) or according to the observed region of the electromagnetic spectrum (radio / infrared / optical astronomy). It is obvious that all the bases of division can not be included in the domain classification but sometimes the choice between them is not easy (cf. Hutchenson 2001: 671).

The second challenge is the choice of the appropriate depth of the classification. Granulation of domain classifications may vary from one to up to nine levels of hierarchical relations (ISO 12620:1999: 23). Too shallow classifications can be uninformative or even misleading (cf. Bergenholtz & Tarp 1995: 153), whereas too detailed classifications are hard to use and maintain (cf. Ubin 1992: 55; Grinev 1995: 85).

The third problem is related to the life cycle of the disciplines. At the end of the 20th century, the number of scientific disciplines used to double every 25 years (Grinev 1993: 8), and the pace has only grown faster. This poses several problems: how to guarantee that the classification is comprehensive at the time of its creation, how to keep the classification up-to-date in the future and how to know whether a new discipline will become established and generally acknowledged or is it only a short-living nonce word.

The fourth major problem has to do with the fact that classification schemes created in different countries and in different languages may differ both in terms of contents of classes and their location in the classification scheme. For example, relations between Russian *машиностроение* (a loan translation and a close relative of the German *Maschinenbau*), Finnish *metalliteollisuus* and English *mechanical engineering* are quite complex, although the final product of these industries is often the same.

Yet another problem is synonymy. Names of domains may have variants and synonyms, such as *animal geography* / *zoogeography*; *legal history* / *history of law*, etc. If synonymous names are not included in the domain classification, users may not be able to recognise the domain under a different name. If synonyms are included but not clustered properly, LSP units belonging to the same domain will end up in (formally) different domains.

9 Overview of TF Core Domain Classification

TF core domain classification has been created with the above mentioned challenges and requirements in mind. The classification is based on several existing classifications,

thesauri, ontologies and encyclopaedia. The list of primary sources of the classification can be found in the *Appendix 1*.

9.1 General Organization of the TF Domain Classification

The classification contains about 700 classes and aims at covering all domains of knowledge and activities. The benefit of having a relatively small classification of top-level classes is that these classes do not change so often and are less culture-, language- and theory-specific than the classes on deeper levels. Besides, such a classification is easy to browse and navigate.

TF classification is a two-level hierarchy. The top level consists of about 100 classes most of which are subdivided further. Below is an example of the domain *Physics*:

(1) **Physics**
acoustics; atomic physics (<- atom physics); biophysics; electrodynamics; geophysics; mechanics; molecular physics; nuclear physics; optics; particle physics; quantum physics; thermodynamics.

LSP units belonging to the second level classes are considered belonging to the corresponding upper class as well. For example, if a user has labelled an LSP unit as belonging to *Acoustics* which is a subclass of *Physics*, search for terms related to *Physics* will by default return units related to *Acoustics* as well.

Some second-level classes appear in the classification two or more times under different top-level classes. For example, *Legal history* can be found both under *History* and *Law and legislature*. This helps users locate domains of an interdisciplinary character in the domain selector. The fact that some domains can be found in the classification in several places does not affect the way in which they are documented in the TF platform. Each node in the classification has its own URI, i.e. it constitutes a complete classifier alone. In the example above, the domain label will always be *Legal history* and not *History: Legal history* or *Law and legislature: Legal history*.

9.2 Treatment of Disciplines of Broad Nature

The nature of some disciplines is so broad that they can be combined with almost any other domain. For example, such words as *philosophy*, *history*, *politics*, *sociology* or *psychology* can be added to almost anything. However, only major and the most important branches of the corresponding sciences could be included in the core domain classification.

9.3 Multiple Domain Labels

Users can label an LSP unit as belonging to several domains. For example, the term *fuel wood* may be labelled as belonging to both *Energy production* and *Logging*. It should be noted, however, that if the same object is considered from different points of view (for example, if *fuel wood* is defined as a *source of energy* in the energy sector and *timber assortment* in the forestry sector), it may be advisable to place different meanings in separate records and provide them with different domain labels.

9.4 Treatment of Disciplines of Complex Nature

Some long-established areas of research and activities are of complex nature. For example, *Marine research* is a complex conglomerate of different individual disciplines such as (marine) biology, hydrology, meteorology, geography, geology, etc. Such complex and somewhat loose conglomerates are split into individual disciplines in the TF core domain classification.

9.5 Treatment of Culture- and Language-specific Domains

TF core domain classification aims to be as generic and international as possible, so culture-specific divisions and classes have been avoided. When compromises had to be made (e.g. what system to follow in dividing *Law and legislature* – civil law, common law, religious law, etc.), priority has been given to the divisions adopted in continental Europe.

9.6 Combining of Domains

Combining of several domains in a single class has in general been avoided. However, sometimes two or three domains are closely related or just often mentioned together. Examples include *Administration and management*, *Ethnology and ethnography*, *Cosmetology and beauty services*. In such cases combining is allowed. The maximum number of domains making up a single class is restricted to three. An additional requirement is that the domains may not stand too far from each other.

9.7 Treatment of Interdisciplinary Words

Interdisciplinary words and word combinations like *analysis*, *report*, *evaluation*, *document* or *method* can be placed in the TF classification into the class *General terms*.

9.8 Treatment of Variants and Synonyms

Common variants, synonyms and near-synonyms of the primary names of classes are included in brackets. In the domain selector, cross-references are used to link variants and synonyms to the main form. Synonyms help users to easier locate domains in the domain selector.

9.9 Support of Multilingualism

In the *TF* project, the core domain classification has been created in four languages: English, Finnish, Russian, and German. Other language versions are to be produced collaboratively in subsequent projects. Finnish was the source language of the classification.

The language of domain labels should correspond to the language of the LSP expression being documented. If the classification is not yet available in a particular language, it is advisable to use the English language version.

9.10 User Extensions

The core domain labels are mandatory, i.e. users always have to specify to which domain in the core classification each LSP unit relates. Users can also supplement domain labels from the core classification with their extensions following the principles described above.

Users can pick extensions already made by other users or add extensions of their own. The language of extension shall correspond to the language of the LSP expression. If users feel that none of the top level classes is suitable for their extensions, they can link them to a special class *Unclassified domains* or one of its subclasses – *Unclassified field of special knowledge* or *Unclassified field of activities*.

When making extensions, users are advised to rely on major existing classifications, thesauri and ontologies in the first place. Sources of extensions can be documented in the same way as any other sources.

10 Conclusion

We have described the main principles of compilation of the core domain classification designed in the *TermFactory* project for the purposes of collaborative terminology work. More information, including detailed principles of extending the core classification and suggestions concerning the implementation of the domain selector, is available in the *Quality Manual of Collaborative Terminology Work* that we hope to get published soon, together with the domain classification.

References

Bergenholtz, Henning & Sven Tarp (Ed.) (1995). *Manual of Specialised Lexicography. The Preparation of Specialised Dictionaries*. Amsterdam: Benjamins.

Grinev, Sergej V. (1993). *Vvedenie v terminovedenie*. Moskva: Moskovskij Licej.

Grinev, Sergej V. (1995). *Vvedenie v terminografiû*. 2nd ed. Moskva: MPU.

Hutcheson, Helen (2001). Practical Considerations for a Term Bank: Termium. In: *Handbook of Terminology Management*. Compiled by S.E. Wright and G. Budin. Vol. 2. Application-Oriented Terminology Management. Amsterdam: Benjamins. 666–676.

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

Kudashev, Igor (2007) *Proektirovanie perevodčeskih slovarej special'noj leksiki*. Helsinki: Yliopistopaino.

Lingvističeskaâ koncepciâ (1989) = *Lingvističeskaâ koncepciâ terminologičeskogo banka dannyh mašinnogo fonda russkogo âzyka: (Proekt)*. Moskva: MGUUÂ.

Nedobity, Wolfgang (1988). Classification Systems for Terminological Databanks. [online]. In Catriona Picken (Ed.). *Translating and the Computer 9. Proceedings of a conference… 12–13 November 1987, CBI Conference Centre, Centre Point, London*. London: Aslib. Available at: http://www.mt-archive.info/A slib-1987-Nedobity.pdf

Ubin, Ivan I. (1992). *ÈVM i slovar': (Metodičeskoe posobie)*. Moskva: VCP.

Appendix 1. Primary References used in Compilation of TermFactory Core Domain Classification

Note: All sites accessed April–August 2010.

1. Eurostat's Statistical Classification of Economic Activities in the European Community (NACE Rev. 2): http://ec.europa.eu/eurostat/ramon
2. Finnish version of NACE Rev. 2 with comments by Statistic Finland (Tilastokeskus): http://www.stat.fi/meta/luokitukset/tieteenala/001-2007/kuvaus.html
3. United Nations' International Standard Industrial Classification of All Economic Activities (ISIC Rev. 4): http://unstats.un.org/unsd/cr/registry/isic-4.asp
4. General Finnish Thesaurus (YSA): http://www.yso.fi/onki/ysa
5. Finnish General Upper Ontology (YSO): http://www.yso.fi
6. Eurovoc, the EU's Multilingual Thesaurus: http://eurovoc.europa.eu
7. Universal Decimal Classification (UDC). Homepage: http://www.udcc.org. Abridged Finnish version: http://www.kansalliskirjasto.fi/kirjastoala/fennica/fennica_udkkaavio.html
8. Wikipedia, the Free Encyclopedia: www.wikipedia.org
9. Helsinki City Library Classification (HKLJ): http://hklj.kirjastot.fi
10. Recommendations concerning the International Standardization of Statistics on Science and Technology by UNESCO: http://unesdoc.unesco.org/images/0008/000829/082946eb.pdf
11. Field of Science and Technology Classification (FOS): http://www.stat.fi/meta/luokitukset/tietee nala/001-2007/kuvaus.html
12. Library of Congress Classification (LCC): http://id.loc.gov/search