[*Translating and the Computer 9*. Proceedings of a conference… 12-13 November 1987, ed. Catriona Picken (London: Aslib, 1988)]

Classification systems for terminological databanks*

Wolfgang Nedobity

Infoterm, Vienna, Austria

INTRODUCTION

The subject code is one of the most important data elements for entries in terminological databanks. A concept can only be defined if it is known which subject field, i.e. which system of concepts it belongs to.

The subject field arrangement differs from databank to databank. According to their particular requirements they choose a more general or a more specific scheme on the one hand, and a more rough or detailed scheme on the other. For practical work, the various domains and topics are being encoded. The codes should have both mnemotechnic qualities and — at the same time — reflect the structure of the subject field. In particular those databanks which are term-oriented cannot do without subject codes. It allows them to differentiate homonyms and to delimit the scope and sphere of validity of a concept. The users can activate the subject code in order to obtain all entries belonging to a particular subject field.

Usually there are files available which offer an overview of all the subject fields and the codes that have been used in preparing the entries. Only when such a listing is hierarchically structured, can one speak of a classification.

In systematic dictionaries each concept is shown within its neighbourhood of related concepts. Such a neighbourhood, i.e. a conceptual field, can be characterised by a descriptor (or its code) from the classification. The same is done for each bibliographic reference in literature databases. Most of these bibliographic databases use a thesaurus in order to be able to control the descriptors used and in order to obtain more accurate results from the queries. Thesauri have a number of advantages over mere alphabetical listings of topics, but only a classification scheme can show in what detail a subject field has been structured.

*Revised and enlarged version of document Infoterm 7-84: Krommer-Benz, M. and Nedobity, W. Klassifikationssysteme und die terminologischen Datenbanken.

If the scope of a descriptor is not made clear, e.g. if it is wider than the user expected, a certain concept indexed by such a descriptor might be applied in the wrong context. Furthermore, a subject code comes in handy when terminological data is exchanged between various termbanks. In such a case, it is of great advantage if there is a common roof classification, otherwise concordances have to be produced. Finally, the determination of equivalences is facilitated if the required concept can be searched for in the set of concepts belonging to the same subject field.

SHORT DESCRIPTION OF SELECTED CLASSIFICATION SYSTEMS

In the following a few terminological databanks from various regions of the world have been selected on the basis of universality, length of operation and user availability.

La Banque de terminologie du Gouvernement canadien (Termium)/ Canadian Government Terminology Bank (Termium), Ottawa, Canada

The classification of the entries is carried out on the basis of a system devised by the terminology bank of the University of Montreal (BTUM). It is a hierarchical system consisting of three levels. This allows the user to widen or to narrow down the search strategy. The notation is composed of an alphabetical code. The twenty-six main classes of the first level stand for the main activities of the Canadian government and are subdivided into further fields which altogether offer 17,500 possibilities of qualifying an entry concept. The system can be characterised by high flexibility but little mnemotechnic support.

Since extensive research in a variety of projects has demonstrated to the Translation Bureau the inadequacy of documentary classification schemes when applied to terminology, it has retained the Termium classification scheme for the new generation of Termium III [1].

It has been updated, however, but the basic structure remains the same because this classification tool is designed to facilitate the retrieval of specialised information from the bank, either by entire fields or by very limited groupings of data.

Commission of the European Communities (CEC) - Eurodicautom

The terminological databank of the CEC is truly universal, because it covers virtually all subject fields. Its classification system was devised by Hans Lenoch who also created a special notation for it. The descriptors are, however, based on the Universal Decimal Classification which was adapted to the requirements of the CEC administration.

The system comprises forty-six main groups, which are represented by a two letter code. This code is based on the French version of the descriptors. Each main group can be divided into thirty-four subgroups which are represented both by the digits 1 to 9 and the letters A to Z ('I' was omitted in order to avoid confusion). The result is a three character code of high mnemonic value, especially since the third character is inheritable in meaning. An entry in the terminology file of Eurodicautom is usually classified by numerous notations. Nevertheless, the classification of concepts would be more accurate if a more detailed scheme was available. The system is actually a mixture between a classification and a thesaurus and was also taken over for the termbanks of the World Bank in Washington and the World Health Organisation (WHO) in Geneva [2]. A special aspect is the possibility of truncation by means of a dash. Theoretically the system would allow 1,564 subject notations, but in practice less than half are used. There is also available an alphabetical index of descriptors which is indispensable for practical work.

Siemens AG, Language Service - TEAM, Munich, FRG

Generally, this databank covers all subject fields with a strong emphasis on engineering, in particular electrical engineering and computer science. The classification scheme was taken from a catalogue prepared by the electrical industry for documentation purposes. This scheme has been enlarged in a very pragmatic way, following the originally established hierarchical order. The code consists of a single letter which is supplemented by one to four digits. Since this notation has no mnemonic value at all, the user has also the possibility to denote the subject field in the abbreviated form of the descriptor. There is a conversion programme available from one form of 'labelling' to the other and retrieval can be carried out by the full form, the abbreviation and the code [3].

Spravochnyi bank terminov 'Automatizirovannoi sistemy informatsionno - terminologicheskogo obsluzhivaniya' (SBT ASITO) [Terminological databank of the 'Automated System of the terminological information service']

This is a databank primarily for standardised terminologies. Thus the classification system applied is identical with the 'All-Union classifier of the state standards of the USSR'. It is a system which builds up notations consisting of ten characters. The first two characters represent the section of the aforementioned 'classifier', which means that the letter code is converted into numbers.

The third and fourth position in the string refer to the class and group in the classifier. The following six digits correspond with the standard identification number in the catalogue of standards combined with the divisionary number of the concept concerned. It is therefore possible to limit the query to the entries pertaining to a particular subject field, a particular standard-type document or

some other category. The design of the classification system is based on the document RD 50-379-83 'Contents and order of providing information for terminology standardization'. ASITO also facilitates the automation of the verification of All-Union classifiers for technical and scientific information and of thesauri for information retrieval [4].

The Danish Terminology Bank (Danterm)

By means of the classification mark the term is referred to one subject field whereby the user will be able to distinguish one term from homonyms classified in other subject fields in the bank. The classification system used is the common Scandinavian 'Nordterm' classification [5]. A proposal for a common classification for all Nordic termbanks was prepared by Ejvind Andersen of the Technical Library of Denmark. This proposal was discussed and expanded by Working Group 2 of Nordterm at their meeting in Helsinki in January 1985.

The notation consists of five characters starting with a single letter denoting the macro-units (A to Q) followed by four digits representing four levels in the hierarchy. The macro-unit should not encompass more than 2,000 terminological units.

It was decided at the above-mentioned meeting to translate the classification into all Nordic languages and to test it at various termbanks [6].

ANALYSIS AND COMPARISON OF THE SYSTEMS

Notations

The terminological databanks under review use for their notations either an alphabetical, a numerical or an alpha-numerical code, which varies in length from three to ten characters. None of the codes follows the widespread decimal systems of libraries and thus they are incapable of expressing intricate hierarchies or multidimensional facets. The codes are relatively simple and require no special program for machine-processing. The mnemotechnic qualities of the codes also vary and – as is the case with the Lenoch system – function only in one particular language.

Structure

The macro-units of the classifications under examination have been compiled with a view to the special purpose of the terminological databanks concerned. The micro-units, however, are determined by the number of entries, i.e. the size of the terminological databanks. The more concepts are stored under a particular subject heading, the more detailed the structure has to be. Some of the structures are more or less hierarchical (e.g. Termium, TEAM) while others are purely enumerative (e.g. Eurodicautom).

Compatibility

Compatibility depends on the selection of subject fields that are to be covered by the various databanks.

In the case of TEAM and ASITO the emphasis lies in the field of technology while Eurodicautom has a strong tendency towards mining and foundry work, due to historical reasons. Furthermore, the availability of certain languages is also to be considered before any exchange of data can be envisaged. The Canadian termbank for instance has very few entries in languages other than English and French. Therefore it is sometimes necessary to allow several classification schemes in one termbank.

CONSIDERATIONS ON A POSSIBLE ROOF CLASSIFICATION

The preparation of a subject classification which is obligatory for all terminological databanks was already the aim of the 'Werkgroep Terminologiebank' which met seven times between January 1978 and May 1979 and which also published a final report on its work [7]. In line with the recommendations of this working group, the following advantages of a roof classification can be stated:

—a unified classification would allow the direct exchange of data collections between databanks and would thus prevent any loss of information;
—joint dictionary projects in particular subject fields could be carried out without the need for conversion routines;
—a more detailed structure of terminological holdings could be accomplished in the form of a common project with the sharing of costs.

On the other hand a warning against the disadvantages of a rigid 'universal classification' has to be given as well. The difficulties which can occur in the course of the preparation of such a system have been reported by A. Bothe in his article 'La classification systématique des stocks terminologiques' [8].

New developments in the area of knowledge engineering and intelligent user interfaces could also be of interest to the producers of new terminological databanks. In any case there should be several ways of access to a terminology file because the users might search for information starting from a single term or from the vague idea of a conceptual field. The system has to cater for both approaches: the non-hierarchical classification and the descriptors of a thesaurus. H. Samulowitz describes the situation like this: 'It is like the movement of a wave: after the hey-day of classifications when primarily physical units had to be ordered, the peak of the natural language systems followed when it was deemed necessary to order logical units and thus thesaurus systems evolved. This phase is now superseded by a mixture of both systems.' [9]

An example of this new movement is the *Root Thesaurus* of BSI [10] which combines the advantages of both systems, i.e. the structure of the Bliss

classification with the natural language access of 5,500 non-descriptors.

As the *Root Thesaurus* is available in machine-readable form in various languages, it would be a suitable foundation, i.e. subject classification for a terminological databank. BSI and several other partners within ISONET classify standards (inclusive of terminology standards) with this tool. The next step would then be the integration of the concepts included in those standards into the system, where a number of them have already been utilised as descriptors. The system is adaptable to individual requirements and extendable in any direction. No wonder that a terminological databank is mentioned among the possible products of the *Root Thesaurus* in the promotional material.

All the information necessary for the production *of Root* is held on the main computer file as a sequence of records in the same order as the subject display. Attached to each descriptor are some additional data elements, not printed out in the present edition. These include management information (reference numbers for descriptors and synthesised terms, codes for hierarchical levels and categories of data), a code showing the source of the term, and a code relating to the availability of a definition.

It is planned to set up a secondary file holding definitions of all the descriptors. The main form of output envisaged is a set of cards, one for each descriptor, with definitions and related terms. Probably the most ambitious exploitation of the work done for the development of the *Root Thesaurus* and systems, a terminological databank could include English terms, details of context, French (and possibly other language) equivalents and definitions where available.

REFERENCES

1. Ashead, M. A new look for Termium. *TermNet News* (1985) 9, pp. 67-68.
2. Lenoch, H. Die Klassifikation in EURODICAUTOM und anderen Anwendungsbereichen [The classification in EURODICAUTOM and other areas of application]. *Terminology Bulletin (1981)* 38, pp. 159-178.
3. Hohnhold, I. Die Terminologie-Datenbank TEAM im Sprachendienst Siemens München [The terminological databank TEAM within the language service of Siemens-Munich]. *BDÜ-Mitteilungsblatt für Dolmetscher und Übersetzer 30* (1984) 3, p. 7.
4. Perstnev, I.P. and Volkova, I.N. The automated system of the terminological information service (ASITO). *TermNet News* (1985) 13, pp. 30-31.
5. Engel, G. and Nistrup-Madsen, B. Danterm. *TermNet News* (1985) 12, p. 8.
6. Picht, H. Klassifikation for termbanker. In: Picht, H. (Ed), *Nordisk Terminologikursus II 'Rolighed'*, Skodsborg, DK, 5-16 August 1985. Copenhagen Handelshojskø1en, 1985, pp. 464-477.
7. Werkgroep Terminologiebank. Schlußbericht der Arbeitsgruppe Klassifikation (15-6-1979) [Final report of the classification working group (15-6-1979)]. *Philips terminology bulletin* (1980) (9) 1/2, pp. 9-11.

8. Bothe, A.A.P. La classification systématique des stocks terminologiques. In: *INFOTERM. Terminological databanks.* Proceedings of the first international conference convened in Vienna, 2-3 April 1979, by Infoterm. Munich/New York/ London/Paris: K. G. Saur, 1980 (Infoterm Series 5), pp. 56-63.
9. Samulowitz, H. 'Renaissance' der Klassifikationssysteme? ['Renaissance' of classification systems?]. *Nachrichten für Dokumentation* (1982) (33) 4/5, p. 188.
10. BSI. *Root Thesaurus.* Part 1: Subject display. Part 2: Alphabetical list. Hemel Hempstead: British Standards Institution, 1981, Part 1: 620 pp.; Part 2: 667 pp.

APPENDIX

Example 1: Termium

Care		Chassis	JDF
(General)	NKA	Checkers	LID
Cargo	VEB	Chemical Compounds	CAC
Cast Iron	OCB	Chemical Engineering	
Causeways	DLF	(General)	CIA
Cellars	DEJ	Chemical Industries	CHB
Cellulosic Plastics	CFA	Chemistry	C
Ceramics		Chess	LID
(General)	JGN	Chimneys	DEK
(Trades and Occupations)	JGI	Chronology	SBG
Ceramics (Applied Arts)	LCB	Circuit Breakers	IFG
Cerebral Surgery	NCK	Circulatory System	NFD
Characterology	REI	Circus	LAF
Charts	SKD	City and Regional Planning	GDB

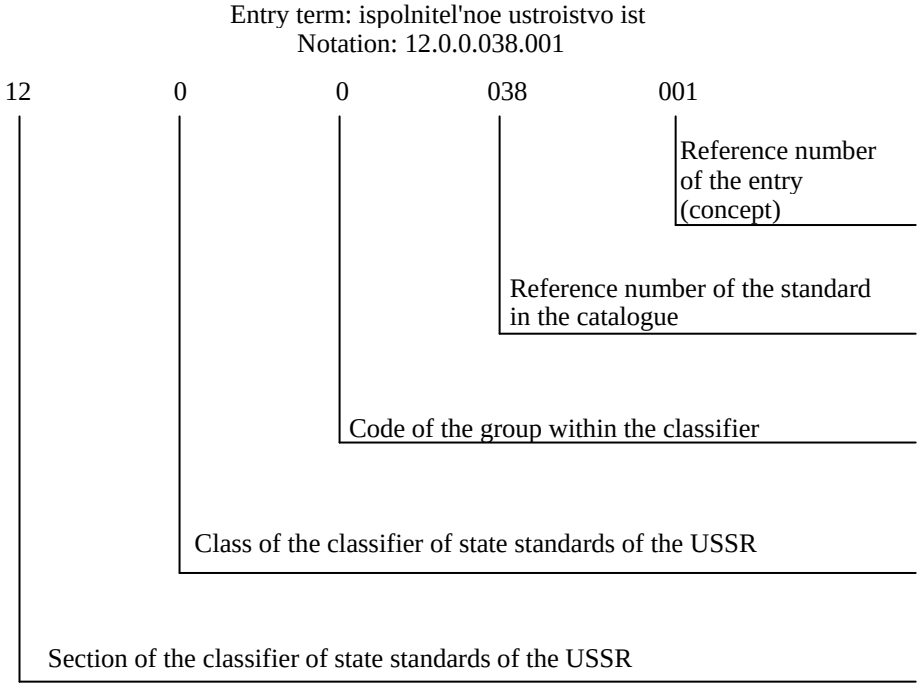
Example 2: Eurodicautom

AG3	Organisations agricoles
AS5	Assurances de personnes
AS6	Assurances de choses
AS 7	Assurances-transport
EN1	Environnement, généralités
EN4	Protection de l'environnement
FIA	Crédits et paiements
JUG	DROIT pénal
SP1	Sports, généralités
TV1	Travail, généralités
TV4	Marché du travail

Example 3: TEAM

Fachgebiet		*Code*
Automatic Control		E3311
Automatische Schreibmaschinen	WP.	E4013
Automatische Sprachübersetzung	MAT.	E4283
Automatische Textverarbeitung	Text	E4119
Automatisieren (allgemein)		E5000
Automatisierung, Ämter der -		E5002
Automatisierung, Betriebe der -		E5002
Automatisierung, Büro	OffAut.	E4016
Automatisierung, Institut der -		E5002
Automatisierung, Organisationen der -		E5001
Automobilclubs	KfzIns.	E9419
Automobilsport	MotRac.	S0200
Automotoren	KfzMot.	E9331
Autorennsport	MotRac.	S0200
Autozubehor	KfzExtra.	E9417
BAB-Bau	RdCon.	E9140
Baderbau		E9131
Badminton	Bad.	S0300

Example 4: ASITO



Example 5: Danterm

Makroenhed Q
Landbrug, Fiskeri

Q 0000	*Generelle anliggender*
Q 0001	geografi, historie
Q 0010	landmåling
Q 0015	mål, vægt
Q 0030	uddannelse, konsulentvirksomhed
Q 0100	miljøspørgsmal (i f.m. denne makroenhed)
Q 0150	organisationer, institutioner, marter, udstillinger
Q 0200	lovgivning, administration
Q 0300	*Økonomiske anliggender*
Q 0400	*Bygninger og anlaeg i landbruget*
Q1500	*Landbrugsmaskiner* Traktorer, motorer og andre kraftmaskiner i f.m. landbrug (+ summen af de under de følgende delområder anførte maskiner; notationen er --1-, fx Q 2110)
Q 2000	*Jordbund, jordbundsundersøgelser*