Medical Informatics Europe '96
J. Brender et al. (Eds.)
IOS Press, 1996

Towards a new generation of terminologies and coding systems

Angelo ROSSI MORI
Reparto Informatica Medica, Istituto Tecnologie Biomediche, CNR
viale Marx 15, I-00156 Roma e-mail: angelo@color.irmkant.rm.cnr.it

Abstract The features of a new generation of terminological systems are introduced: they are computer-based, multihierarchical, extensible, mappable. This performance is achieved by a compositional approach: each phrase is systematically represented by predefined descriptors, according to a "categorial structure" (a model of semantic categories and their relations). A terminological system is therefore made of four interrelate components: categorial structure, system of descriptors, system of phrases and their systematic representations. The role of the European standardization body (CEN) to support the development of such terminological systems is outlined.

1. Introduction

Health care management is based on transmission and processing of a large amount of heterogeneous data. Telemedicine applications and computer-assisted rationalisation of health care demand for multiple views on federated databases on patients, activities and resources.
Terminological systems in health care are the main support enabling this performance, and thus the need for their extension, rationalisation and integration is growing [1].
In Health Care Information Systems too many coding systems are in use, not compatible among them [e.g. 2-6]; many of them are very large and range between 10 000 and 100 000 concepts, to cover all the aspects of the medical disciplines and organisational knowledge.

1.1. Terminological phrases as triggers of knowledge within applications

Ongoing efforts are extending collections of sensible phrases in all health-care domains (e.g. diseases, procedures, drugs, laboratory quantities, medical devices); they consider differentiating details to assign actual "individual" items to relevant classes for defined tasks (starting from statistics up to routine use). Coding systems and terminologies are developed by independent organisations: nevertheless, they should progressively converge and increase their interoperability, in the shortest time as possible and avoiding inconsistencies.
Terminological phrases in health care are not the usual "terms" from terminology theories [7]: in fact, they are expressions pragmatically created by users to specify an adequate amount of details, able to trigger administrative, scientific and clinical applications (see the tables used to place orders for services, to schedule the use of resources, to ask for reimboursement).
To allow software applications to exploit the knowledge behind each phrase, all the involved potential details should be made explicit. The set of details evoked by a "motivated" phrase goes beyond the juxtaposition of the meanings of each word in that phrase; extensive computer-based exploitation of nomenclatures requires to make explicit at least a part of this knowledge using predefined descriptors, to complement the list of the sensible phrases.

1.2. Need for a representation based on predefined descriptors

Each phrase from a nomenclature could be represented by descriptors taken from a "system of descriptors" (i.e. from an intermediary thesaurus — *inter-thesaurus* — independent from the source corpora). Descriptors may be used for intensional definitions and to build systematic names for synonymous phrases, according to a suitable set of generative patterns [16].

This representation will support maintenance of individual terminological systems, allowing dynamic arrangement of phrases according to different criteria, to satisfy different purposes. It may provide a background to translation of phrases into various languages, indexing, and semi-automatic coding (including search for classes which apply for a particular individual). Moreover, it allows to compare nomenclatures, and to cross-map among their entries.

2. Three generations of terminological systems

Computer-based systems increase life-time and accessibility of data; a major consequence is the need to convert coded data from one environment to another; extreme precision of meanings is required for multiple use of represented data. Paper-based terminological systems are no more able to satisfy the increasing needs of computer-based processing.

2.1. First generation: paper-based terminological systems

Paper support does not support multiple organization of the rubrics, according to different criteria. Maintenance is expensive, and extension to meet local needs is difficult. Until now few attempts were made to merge concept systems on a subject field into multipurpose tools. Paper-based Information Systems were not requested to support a large re-use of the same data; the contacts among professionals in the Health Care System were either partially structured (by pre-defined forms, eg. insurance claims), either informal (being based on human-to-human communication, eg. ad hoc paper-based or oral communication of results).

2.2. Second generation: compositional systems (with categorial structures and descriptors)

Terminological systems aimed at routine use (eg. [3, 6]) are migrating towards more complex structures and performance, envisaging a second generation: new systems, conceived for computer use, process explicit details and are mappable one to the other; they are based on a compositional approach using predefined descriptors. Complexity and evolution rate preclude paper presentation. Guidelines and vocabulary provided by available domain-independent standards from various committees of International Standard Organisation (ISO) — on terminologies (ISO TC37), classifications (ISO TC69-SC1), codes (ISO-IEC JTC1-SC1), and thesauri (ISO TC46-SC3) [8-13] — result mostly inadequate and sometimes inappropriate against the specific requirements in Medical Informatics, outlined in this paper.
In particular, the approach proposed by ISO TC37 on Terminology [8, see also 14] is insufficient; in fact, that approach was the starting point of the SESAME project (1990-1991) in the Program 'Advanced Informatics in Medicine' of European Union [15], which tried to clarify it and make it more effective, in order to be applicable to sets of very large concept systems.
The project introduced the idea of "structure" of a system of concepts, that was then reworked and resulted in the "categorial structure" as described by CEN ENV 12264 [16].
The structure of a concept system consists of a list of involved categories (with reference to the available authoritative sources for detailed values) and their typical relations. Any large concept system is organised, by describing relations among concepts, and results in the systematic description of a subject field (see the organisation implied by SNOMED [6] or MeSH [4]). In a large concept system, the categorial structure may be therefore identified, and could be the explicit framework for the organisation of detailed concepts [15, 16].

2.3. Third generation: formal systems (universal, parsimonious, precise)

Experiments on advanced compositional systems seem promising [18-20]. A 3rd generation is appearing: *formal systems*, based on "universal" models, allowing advanced computational performance. They represent all and only the medically-sensible statements, provide validation capability, and are parsimonious (by a generative approach). But development of formal models is more expensive and resource consuming than building structured representations

with descriptors; in particular, many problems of global coherence and "normalization" in a large model are still unsolved; appropriate software tools have to be developed, and familiarity should be reached by an adequate number of experts, before a practical wide application of this approach can be put in place. The *GALEN project* [17] (1992-1995) in the Program 'Advanced Informatics in Medicine' of European Union (followed by a demonstration project, GALEN-IN-USE, 1996-1998), is defining in a computable way, by *formal intensional definitions* or conventional subsumption, a large number of expressions.

3. Responsibilities and roles in developing a second-generation system

A terminological system of second generation is developed by an iterative process, up to an adequate level of complexity and robustness. The process is suitable for decentralization and for progressive involvement of an increasing number of developers, and defines responsibilities and roles for standardization bodies, experts and national coding centers [29].

3.1. The four components of a second-generation system

The development process produces four complementary results:
- a *categorial structure* [16] describes semantic categories, semantic links and structural patterns for a system of concepts on the given topic; it requires a modest amount of resources in a short time; it is suitable for standardization activity and may result in a standard;
- an *inter-thesaurus* provides descriptors suitable for a given categorial structure; it organises them by multiple hierarchies and cross-relations; it requires initially a modest amount of resources to work out most descriptors [27]; development is complex for a standardization initiative, but if results are produced elsewhere, they may be included in a standard;
- a family of structured *sub-systems of phrases*, with systematic names; it requires continuous maintenance and local adaptations; it is not suitable for standardization (except for ancillary sub-systems, as a "reference classification" and a "reference nomenclature");
- a *knowledge base* with systematic representation of each phrase made of descriptors from the inter-thesaurus, according to the structural patterns, ie. a unique combination of descriptors able to identify each phrase within a source and to compare similar phrases from different sources. It is supplementary to the previous results, able to validate and refine them, and it has to rely on independent initiatives from external sponsors. This knowledge base is not another coding system, nor a new terminology; although systematic names can be built from it as reference "lingua franca", they are not suitable for use by end-users in routine applications.

Categorial structure, inter-thesaurus, and knowledge base are not only results per se, but they are also suitable for further formalisation, eg. they are used in the GALEN-IN-USE Project [17], to prepare an intermediate representation for a semi-automatic translation into the GRAIL language, in a demonstrator about multicentric cooperative production of an integrated model of surgical procedures, for advanced computer-based exploitation [29].

3.2. Role of standards in the development process

Standards have a precise role in the process of developing new terminological systems.
In the short term, standards can only provide for registration of coding schemes (eg. ENV1068 [23], similarly to ASTM and HL7, provides a prefix to identify each coding scheme in healthcare messages).
In the medium term, standard categorial structures on individual topics can support spontaneous convergence and systematic development of terminological systems about that topic; of course, those standards are not aimed at end-users, but their main target group is made of developers of terminological systems. More in general, terminological modelling can clarify the expectations about the content of a coding system in the design of information systems: categorial structures allow also the integration of concept systems within the patient

record and messages for data interchange, by matching the items of the first with the items in the information model of the others.
In the long term, the categorial structures on the different (overlapping) topics should be harmonized on a deep ontological basis: individual categorial structures will be made coherent by referring to a unique metastructure, facilitating also the development of "universal" formal models of third generation. The ultimate goal is the cooperative development of this "*shared ontology*" for medicine, and thus of an integrated system of concepts; intermediate results are effective in the short and medium term in narrower subject fields and with limited resources.

3.3. Categorial structure of concept systems as a mean for convergence

The above approach was adopted by the European Committee for Standardisation (CEN) to facilitate a gradual and *spontaneous* convergence of large coding systems. In particular *Working Group 2* (WG2) on 'Terminology, semantics and knowledge bases' of the Technical Committee on *'Medical informatics'* (TC 251) is developing a series of standards according to a Work Plan [24] (up to now, on surgical procedures [25], properties (including quantities) in laboratory medicine [26], medical devices [27]). The *methodological background* was set up by Project Team CEN/TC251/PT003 'Model for representation of Semantics in medicine' (MoSe) [28], which produced ENV 12264 [16] on description of categorial structures.
Categorial structures are a powerful tool to synthetically describe the content of large concept systems, to allow their comparison and to facilitate future convergence by a more systematic design, even independently from computer-based applications. From the same categorial structure various compatible terminological systems may be built, suitable for different tasks (eg., in the field of laboratory: collection of a specimen, reimbursement of a service, request of a service, communication of a result, etc.). The same categorial structure may be used to prepare principled classifications and nomenclatures (with systematic names), multi-hierarchical systems, combinatorial systems, and so on.

4. Conclusions

The management of medical semantics is a key issue of future clinical information systems.
A unique environment for information processing and communication in Health Care is being established; different speciality-related information systems and purposive coding systems are facing and conflicting in this environment. Computer allows for coexistence and integration of multiple coherent coding systems in the same information system, for different tasks.

New terminological systems can represent the needed level of details on clinical cases (as opposed to *classify* them) to cluster them dynamically, according to varying user's needs; multiple uses of the same data, with appropriate conversion, are theoretically possible. End-users can benefit of advanced interfaces to classify more faithfully individuals by a particular nomenclature and to browse one or more nomenclatures according to multiple viewpoints.

Advanced terminological tools should be able to couple different cultural and clinical environments to serve the increasing communication needs. Availability of new terminological tools will influence many scenarios on exploitation of knowledge:
- to link knowledge bases to medical records in Intelligent Information Systems,
- to transform (using previously formalised knowledge) information available in books into a knowledge-based module (and to add it to an existing knowledge base),
- to process medical records to extract knowledge (or to browse a clinical database, based on natural language) and to store, compare and merge the extracted knowledge.

Techniques able of supporting the required increase of performance are not yet fully worked out, but there is a desperate need for immediate solutions. It is important that as much as possible of the effort required to achieve quick solutions contribute to the long term solution as

well and vice versa. Integrated terminological systems need a joint effort of co-operative development in order to enhance their effectiveness.

The approach outlined in this paper can be applied to every topic were there is a need to facilitate maintenance, rationalisation and spontaneous convergence of existing nomenclatures into an integrated terminological system.

Acknowledgements. Work partially supported by contract HC1018 "GALEN-IN-USE" from European Union.

References

1 Rossi Mori A. Coding systems and controlled vocabularies for hospital information systems. *Int J Biom Comp* 39 (1995) 93-98
2 World Health Organisation. *International Classification of Diseases,* 9th revis. Geneva: WHO, 1977
3 Read JD. The Read codes. CEN/TC251/WG2/N62, Copenhagen: CEN/TC251/WG2 1992
4 National Library of Medicine. *MeSH Medical Subject Headings*. Bethesda, MD: NLM (yearly)
5 Humphreys BL, Lindberg DA. The unified medical language system project: a distributed experiment in improving access to biomedical information. In: KC Lun et al., eds. MEDINFO 92. Amsterdam: Elsevier Science Publishers, 1992, pp 1496-500
6 Rothwell DJ, Coté RA, Brochu L (eds), *SNOMED International*, Northfield, IL: College of American Pathologists, 1993, 3rd ed.
7 Galeazzi E, Agnello P, Gangemi A (et al). What is a medical term ? Terms and phrases in controlled vocabularies and continuous discourses. In: Barahona P, Veloso M, Bryant J (eds):*Proceedings of the 12th Congress of the European Federation of Medical Informatics,* Lisbon, 22-26 May, 1994, 234-239
8 ISO 1087 Terminology -Vocabulary
9 ISO/IEC TR 9789 Coding methods and principles
10 ISO/DIS 3534-1, -2, Statistics — Vocabulary and symbols
11 ISO 704 Principles and methods of terminology
12 ISO 2788 Documentation — Guidelines for the establishment and development of monolingual thesauri
13 ISO 5127/1 Documentation and information — vocabulary — part 1: basic concepts
14 Nordic Terminology Centers. Practical guide for terminology work: Nordic proposal for CEN/TC251, included in CEN/TC251/WG2/N209) Copenhagen: CEN/TC251/WG2, 1994
15 SESAME Compilation of Deliverables c/o P de Vries Robbé, POBox 9101, NL-6500 HB Nijmegen
16 CEN ENV 12264:1995. Medical Informatics — Categorial structure of systems of concepts — Model for representation of semantics. Brussels: CEN, 1995
17 GALEN and GALEN-IN-USE documentation, available from the main contractor AL Rector, Medical Informatics Group, Dept. Computer Science, Univ. Manchester, Manchester M13 9 PL, UK (e-mail galen@cs.ac.man.uk; home page http://www.cs.man.ac.uk/mig/galen)
18 Campbell K, Musen MA. Representation of clinical data using Snomed III and conceptual graphs. In: *Proceedings of the 16th Symposium Computer Applications in Medical Care*. November 1992 pp 354-8
19 Friedman C, Johnson S, Cimino J, Conceptual Graph Representation of Radiology Findings — Merged Model (Version 3, 29 Nov. 1993 for the Canon Group)
20 Rector A. Compositional models of medical concepts: towards re-usable application-independent medical terminologies. In: Barahona P, Christensen JP eds. *Knowledge and Decision in Health Telematics*. Amsterdam: IOS Press, 1994: 109-14
21 Gangemi A, Galanti M, Galeazzi E, Rossi Mori A. Beyond UMLS: computational semantics for medical records. In: KC Lun et al., eds. MEDINFO 92. Amsterdam: Elsevier Science Publishers, 1992, pp 703-8
22 Rossi Mori A, Gangemi A, Galanti M. The coding cage. In: Reichert A, Sadan BA, Bengtsson S, Bryant J, Piccolo U eds. MIE 93. London: Freund Publishing House, 1993, pp 466-72
23 CEN ENV 1068:1993. Medical Informatics — Health care information interchange — Registration of coding schemes. Brussels: CEN, 1993
24 CEN/TC251, Directory of the European Standardisation Requirements for Health Care Informatics and Programme for the Development of Standards (Version 2.0). Gent: CEN/TC251, 1995
25 CEN ENV 1828:1995. Health care informatics — Structure for classification and coding of surgical procedures. Brussels: CEN, 1995
26 CEN ENV 1614:1994. Health care informatics — System of concepts for systematic names, classification, and coding for properties, including quantities, in laboratory medicine. Brussels: CEN, 1994
27 CEN/TC251/PT2-015. Medical informatics — Categorial structure of systems of concepts — Medical devices (First Working Document). Brussels: CEN, 1995
28 Rossi Mori A, Bernauer J, Pakarinen V, et al. Models for representation of terminologies and coding systems in medicine. In: De Moor GJE, McDonald C, Noothoven van Goor J, eds. *Progress in Standardisation in Health Care Informatics*, Amsterdam: IOS Press, 1992: 92-104
29 Rossi Mori A. The Cooperation Document. (internal report for GALEN-IN-USE) Rome: ITBM-CNR, 1996 (available through [17])