ISO / TC215 / WG3

Supporting document for the WI

"Vocabulary on terminological systems"

Angelo Rossi Mori, CNR, Rome, Italy angelo@itbm.rm.cnr.it

Table of contents

1. Terminological challenges in healthcare information systems 2
2. Evolution of coding systems towards controlled representations 2
Difficulties in merging multiple "visions of the world" 3
Free text, coding systems and modifiers 4
Kinds of knowledge that define a concept 4
The delimitation of a concept by its "context" 5
a) Context of a rubric inside a terminological system 5
b) Context of a phrase inside the original composition or transaction 5
c) Context of a code, when used inside an information system 5
3. New approaches: LOINC, SDM, Clinical Terms 6
LOINC 6
DICOM-SNOMED Microglossary 6
Clinical Terms in UK 6
4. Three generations of terminological systems 6
First generation: traditional terminological systems 7
Characterization of different kinds of terminological systems 7
Classifications, nomenclatures and thesauri 9
Usages and tasks for first-generation systems 10
Features of first-generation systems 11
Second generation: compositional systems 11
The four components of a second-generation system 11
Usages and tasks for second-generation systems 12
Features of the second-generation systems 12
Third generation: formal systems 13
Synergy between second and third generation 13
Features of the third-generation systems 14
5. Role of standards in the development of terminology systems 15
Co-existence and convergence among systems on the same subject field 15
Maintenance and upgrade of traditional systems 16
Production of systematic names and reference lists 16
Increasing coherence among systems on overlapping subject fields 17
Development of second-generation systems 18
A scenario on standard-driven developments 19

1. Terminological challenges in healthcare information systems

Present coding systems are hardly able to support the evolution of the next generation of Healthcare Information Systems.

In fact, computer-based coding systems and terminologies are requested to evolve too, in order to satisfy the incoming challenges:

1) many applications, including hospital sub-systems and GP's record systems, each using its own *sub-language*, are increasingly facing each other within a "federated environment" and are communicating with systems in other institutions.

   A Babel Tower is becoming evident: each professional belongs to various overlapping clinical sub-communities (characterised by profession, task, purpose, tools) and exchanges information within them, e.g. inside hospitals or across the hospital barriers. Computer-based solutions should be able to map among the concepts represented by these sub-communities, allowing to process, present and transmit a variable quantity and quality of details.

2) in an ideal Electronic Healthcare Record System, data should be entered only once and stored near their source; afterwards these data should be selected, converted and transmitted — when, where and how requested to comply with the receiver's needs — in the same or in a different *context.*

The integrated management of the same data for diverse purposes could assure: reliability, efficiency, timeliness, friendliness (data should be transformed ad hoc for various actors, tasks, time scales).

A selection of routine clinical data, validated by an effective use, might be forwarded also for administrative, planning, auditing purposes.

In this chapter we discuss the present trends in terminological systems, and their relation with the present standards. More broad applications of terminological know-how to health informatics — of relevance to the future standardisation activities of WG II — are described in chapter 3.

2. Evolution of coding systems towards controlled representations

Clinical data may be entered into information systems according to various approaches. Each approach is very effective within its scope:

- free text (redundant, slow and ambiguous, but complete and flexible);
- coded data (rigid and artificial, but in principle based on consensus);
- locally controlled vocabulary (friendly, effective, but not standard).

The most frequent uses of this information was so far consisting in retrieving the suitable item, transmitting and presenting it to the users without a particular semantic processing.

The scenario is changing; priorities are going towards the representation and conversion / adaptation of standardised information; whatever is the input solution, the internal representation of the *relevant* information should be suitable for further processing of its semantic content.

The focus is in the organisation of information, appropriately preserving detail for further re-use.

A pertinent reduction of the proportion of unprocessed free text and a suitable representation of the clinical concepts in the context of the structure of the patient record has to be achieved.

The full exploitation of free text information contained in clinical documents is feasible, but laborious: it can be achieved in full only after appropriate pre-processing, i.e., if the information is adequately interpreted and represented; since the issues on free text and language processing are out of the scope of this report, let us focus on coding systems, controlled vocabularies and on internal representation of concepts, considering also the potential trends on this topic.

The main goal of most coding systems is to reduce information *(classification)*, disregarding details which are considered not significant for a given purpose; a posteriori coding by trained professionals is used mainly for economic, statistical and epidemiological tasks.

The possibility for computers to manage a larger amount of data is producing a shift to more precise and detailed labels *(nomenclatures)*, and is inducing the use of multiple coding, with effective "structured input" performed directly by users.

A controlled vocabulary can be used to represent concepts and their relations in complex structures, building an (artificial) language with all the functions needed to communicate information: it is similar to natural language, but ambiguities and redundancy are reduced.

If the vocabulary is developed and maintained locally, it may result precise and effective, but is difficult to transmit information externally, reaCHING the same degree of precision and faithfulness ALSO IN THE RECEIVING APPLICATIONS.

Difficulties in merging multiple "visions of the world"

Collections of clinical terms may reach very large numbers, up to a million entries, but the amount of concepts that each individual user manages routinely is very limited.

The amount of details encapsulated in each item (granularity) and the width of the involved domain are inversely proportional, so that the total amount of concepts managed routinely by an individual user tends to remain relatively constant:

- a highly specialised physician deals with a narrow domain and uses the maximum number of details;
- a statistician deals with international comparisons in the whole healthcare system and uses the minimum of details.

The units of information that a user feels worthwhile to distinguish in a single application seem to be of the order of $10^{3\pm1}$ concepts: see

- DRG classes for administrators,
- ICD categories and tabulation lists for statisticians,
- ICPC combinations for general practitioners,
- the set of speciality lists of pre-defined diagnoses and procedures in DIOGENE I (Geneva) and
- the structured forms in a patient record (either on paper or on computer) for physicians, etc.

In the same team, different professions — e.g., a nurse vs. a physician — use different details to build their own controlled vocabulary.

Previous observations suggest to assume an operational way to define the appropriate granularity for an entry, as *the distinguishable unit of information that is able to influence the behaviour of someone* (in a given context) in a significantly specific way.

A coding system collects these units of information, to harmonise them within a given sub-community of users and usually provides ad hoc information to assure homogeneity in the use of each item. As a consequence, it imposes a pre-defined "vision of the word" in that sub-community.

The number and the variety of items faced by a class of users (plus the related tables and cross-references) may then involve coding difficulties, being at the origin of frequent errors in the coding process; it may require experience with handbooks and manuals.

The collection of concepts from different sub-communities (in order to satisfy a variety of users, purposes, contexts) produces of course a very large list, as mentioned above, and entails problems of equivalence, cross-references and multiple organisation of items referring to the multiple "visions of the word" carried by each of the various concept systems.

Free text, coding systems and modifiers

Free text does not allow immediate re-use of data within computers; it supports human-to-human interactions and some limited information retrieval but *codes* (or at least registered phrases) are needed for transmission of clinical data in shared care and to access tables for management of resources, reimbursement, audit.

Traditional coding systems provide codes and *huge lists of phrases* (classifications, nomenclatures), perhaps organized into *one or more hierarchies*. They satisfy therefore most needs from claims for health insurances and epidemiological-statistical uses, but fail to satisfy routine clinical needs about flexibility and expressiveness, characterizing the sub-languages of the different specialties and allied healthcare professions.

The usage of *thesauri* made of elementary *descriptors* is appropriate for indexing and information retrieval (eg. MeSH), but the lack of pre-coordinate phrases and of codes for combined entries makes them not suitable for clinical purposes.

The use of *modifiers* (ie. the combination of a main entry from a nomenclature with a variable number of elementary descriptors from a thesaurus) seems promising. It can reach an appropriate level of flexibility, but introduces a variable length in the codes and the need to have generic combination rules (which subset of modifiers is appropriate in which context ?) and to verify appropriateness of specific combinations. Moreover, it makes more difficult the production of reference tables (how a modifier of a procedure affects its cost ? how a modifier of a device affects its functions ?).

Kinds of knowledge that define a concept

In a given coding system or in a reference database, the knowledge attached to each concept may consist of the following items:

- a **key** for its unique identification, e.g. a code
- **a rubric**, ie. a precise string that identifies the concept, e.g. in ICD the title of a class
- a set of **clinical phrases**, including preferred term and synonyms, often used (with permutations) to build the indexes
- a **description** to explain, but also to fix, the meaning of the item, as:
    - textual and iconic definition
    - explicit inclusions and exclusions
    - multimedia representations
- **cross-references** to more elementary concepts
- a **formal description** acording to a pre-defined description logic
- **constraints** to check consistency with other data , e.g. age and sex; compatibility between diseases and procedures
- **inclusion in clusters**, made to group homogeneous data for given purposes (e.g. tabulation lists for statistics)
- **knowledge** to provide a systematic presentation of concepts, e.g. hierarchical and causal relationships

- **pragmatic information** (perhaps stored in look-up tables) needed by the information systems, e.g. cost-related information

All these items should be considered in standards about computer-based terminological systems.

The delimitation of a concept by its "context"

A concept is never isolated, but it is always part of a context.
Here we analyze three kinds of context:

a) context of a rubric inside a terminological system
b) context of a phrase inside the original document
c) context of a code, when used inside an information system

a) Context of a rubric inside a terminological system

Precise sense of a concept is determined by the terminological system in which it is embedded, i.e. by:

names of classes; related terms (preferred term, synonyms);"Presentation Trees" (systematic list, perhaps hierarchical); descriptions in free text (dictionary, glossary); explicit relations with other concepts (cross-relations, conditions for inclusion and exclusion in a class, ...).

The sense of each conceptual unit delimits and is delimited by the others.

Relevance of the implicit mutual definitions in a structured set of concepts is underestimated.

Quality and quantity of information provided to the user change from system to system; they reflect the expected use of the system and the degree of standardisation of models and concepts already present in the professionals of the domain.

b) Context of a phrase inside the original composition or transaction

Context of each phrase in the application contributes to define the concept to be coded.

For example, in a healthcare record from a surgical department, the simple string "excision" may fill in a <procedure field>, with an obvious, implicit reference to a previous <disease field> in the same document (where the involved body part or lesion is fully specified).

The "producer" is aware of the organisational and cultural context and preserves a part of the context with or within the phrase (e.g. "in cardiopathic")

c) Context of a code, when used inside an information system

In a specific information system, in a given moment, a code belongs to a complex of data that is managed together by an application.

It may consist of the set of concepts that is carried by the same message or in the values that are stored in selected database fields of the same healthcare record.

Which amount of this context must be transmitted by a message (or stored for further use), in order to allow the receiver (or the next user) to process a concept correctly ?

A preliminary answer is given by the OCC (original complex component) in ENV 13606, but further work on this issue is still needed.

3. New approaches: LOINC, SDM, Clinical Terms

New approaches are being introduced into routine applications:

LOINC

It is a recent system on laboratory observations.

Items are systematically organized by segments and sub-segments.

Items are selected from lists provided by vendors and medical societies; most items are intended to be directly usable in laboratory messages with a unique code. If more details are really needed, the system provides also a way to specify the kind of detail (e.g. to express and transmit additional details about site of measurements, devices and methods) and some mechanism for combination.

It provides also examples (a non-exhaustive short list) of allowed modifiers in each specific context.

It abandoned "positional" codes in favor of pure sequential ones; hierarchies and cross-relations should be managed through additional tables (e.g provided by SNOMED-RT).

Note that LOINC is a collection of names, each with an associate data type for the “answer”, and with an associate value domain (range or answer-list). In this respect, it is different from the usual coding systems that deal with value domains and it is similar to a metadata registry.

DICOM-SNOMED Microglossary

It is a system for the description of imaging procedures and the content of images (at present, under revision).

It provides a database where each exhaustive set of phrases or elementary descriptors is assigned to a "context group".

Context groups can be structured into a "template", ie. a model to represent phrases on a given topic.

Different medical societies are directly involved in the development of the system.

Clinical Terms in UK

Clinical Terms Version 3 includes terminology for:

- Clinical Findings (which includes diseases – known as Disorders – as well as Clinical History and Observations, Social History and Observations, Lab test observations, for example)
- Operations and Procedures (which includes surgical procedures as well as non surgical procedures, regimes and therapies particularly relevant to non-medical professions, investigations)
- Administration pertaining to the UK healthcare system
- Drug and appliance of all NHS prescribable items in the UK
- Occupations

Diseases and procedures are being developed by a compositional approach, i.e. each phrase corresponds to a set of predefined elementary descriptors that can fill in a number of slots according to a predefined frame.

The hierarchy of elementary descriptors is coherent with the hiearchy of the compound items.

4. Three generations of terminological systems

Different kinds of languages can be processed by humans and computer, with different degrees of flexibility, expressiveness, easiness of input.

clinical language:
for human communication
interpreted and re-adapted by humans according to context

formal systems:
for advanced computer-based processing
symbols are used by computers, final results are interpreted by humans

coding systems:
for multiple computer usages
counting (statistics; table construction)
retrieval (epidemiology, audit; presentation of cases)
management (access to additional knowledge in look-up tables)
they are a compromise (they link "frozen" clinical language to standard labels, usually codes)

Here we focus on different kinds of coding systems and formal languages.

Traditional coding systems are considered here as "first generation".

Recently introduced systems, as Clinical Terms Version 3 (The Read Codes), LOINC, DICOM-SNOMED Microglossary are considered as precursors of a second generation.

Advanced systems under development, such as formal systems (eg. the GRAIL-based models in GALEN), are instead considered as third generation.

First generation: traditional terminological systems

Various approaches to terminological systems (controlled vocabularies, nomenclatures, taxonomies, coding systems) were developed in the past to perform a range of functionalities, depending mainly on the cultural environment where they were conceived. In correspondence, a set of domain-independent standards provide guidelines and vocabularies to deal with each particular kind of terminological system.

These standards were independently produced by various committees of the International Standard Organisation (ISO), e.g. on terminologies (ISO TC37), classifications (ISO TC69-SC1), codes (ISO-IEC JTC1-SC1), and thesauri (ISO TC46-SC3).

Standards produced by the above different communities are not harmonized among them and thus provide incoherent vocabulary and definitions; they result mostly inadequate and sometimes inappropriate against the multiple requirements in Healthcare Informatics.

Characterization of different kinds of terminological systems

An effort of systematization was made in the European Prestandard CEN ENV 12264, where the different kinds of terminological systems were described *within the same framework*, as far as possible in agreement with the various standards produced by the above mentioned ISO Technical Committees.

Note that only nomenclatures and controlled vocabularies are made of *actual expressions* currently used by professionals of the field (e.g. in patient records and in human-to-human communication).

Classifications contain instead artificial expressions using metalanguage (eg. "other location", "not otherwise specified") not used during care provision but only to build statistical classes.

Descriptors of a thesaurus are instead "standardized" atomic concepts used to index documents or sentences.

The following table summarizes the standardized usage as required by CEN ENV 12264.

vocabulary	In a vocabulary every term of a particular *subject field* should be the result of terminological work and should be accompanied by a definition. A vocabulary should be adopted by a particular clinical linguistic community only in force of some agreement.
terminology	The set of terms representing a *system of concepts* is referred by ISO 1087 as terminology. In a broader sense, 'terminology' may include also the system of concepts and thus the definitions, becoming a synonym of 'vocabulary'. A set of codes may be provided to identify terms (may be also to distinguish preferred terms from synonyms) or concepts.
nomenclature	A nomenclature should involve a system of *terminological phrases* elaborated according to pre-established naming rules, ie. should consist of *systematic names*. A coding scheme (i.e. the labelling of each name by a code) may be optionally provided with a nomenclature. If there are codes, they may be multiaxial — reflecting the rules for constructing names —, or hierarchical, or just sequential. In computer systems, relations should be represented by mechanisms that are more explicit and flexible than multiaxial or hierarchical codes, so that the use of those partially meaningful codes can become gradually less relevant.
controlled vocabulary	A controlled vocabulary should be a limited set of *terminological phrases* used within an organization for a given purpose in a specific subject field. Designers need not to provide all the definitions, due to the homogeneity of users and uses; they can put a terminological phrase into such a list considering only its local use (and perhaps a specific application software) and are not required to search for agreement within a larger clinical community. Within a computer-based application, the controlled vocabulary can be organized in tables, one for each data base field or entry considered by the application. A set of codes — to be managed by the application software — may be provided, to identify each terminological phrase within each table or in the whole application. A unique comprehensive controlled vocabulary should be built and maintained by each organisation, merging all the over-specialized vocabularies of the teams or departments. The controlled vocabulary should provide unambiguous *transformation* tables towards relevant coding systems.
thesaurus	A thesaurus should be a collection of atomic concepts of a controlled indexing language, formally organized so that a priori relationships between concepts (e.g. 'broader' and 'narrower') are made explicit (ISO 2788).
classification (taxonomy)	A classification should be a terminological system whose system of concepts is structured by generic relations only. Classes may consist of objects or concepts; in consequence, two kinds of activities may be performed on classes: - in the preparation of a hierarchy of classes (**taxonomy**), developers identify the essential characteristics of each concept in a class, and use them to <u>organize</u> *classes of concepts*; - in the routine application, only the **subset of the most specific classes** is used; users <u>assign</u> each object to a *class of objects*, according to the predefined characteristics.

coding scheme	A coding scheme is a collection of rules to represent items of a 'coded set' (e.g. rubrics) into the elements of a 'code element set', or 'code set' (e.g. codes); it does not include the elements of the two sets (see ISO 2382-4). The first set may be: a) a set of objects; b) a set of concept designations; in particular it may be: b1) a set of codes.
coding system	A coding system shall be a combination of: a system of concepts; a terminology (rubrics); a set of code values; at least one coding scheme to relate the codes to the concepts or the terms. In a coding system, the system of concepts or the terminology are the coded set, and the set of code values is the set of code elements (ISO/IEC TR 9789).

Classifications, nomenclatures and thesauri

We briefly describe here the main features of three relevant kinds of terminologic al systems: classifications, nomenclatures and thesauri.

Classification

set of classes to arrange individuals; taxonomy among classes is optional

examples: ICD9/10, ICD9-CM, ICD9-Procedures, DRG; specific systems developed by national and international clinical societies (Read Codes can be considered both as a classification and as a nomenclature)

- Uniqueness of event-to-code relationship; (an event or patient corresponds to one and exactly one code). Minor exceptions have been verified in some cases where the same classification is required for different purposes, and an additional code may be requested in order to specify details.
- Classes are mutually exclusive. Rare exceptions have been verified when different "clinical schools" conflict
- Propensity for comprehensive expressions including several clinical concepts (designers, not users, decide on the amount of details to be preserved and represented)
- Limited coverage of clinical field. Coverage usually comprehends only diseases and procedures
- Existence of a wide variety of metalanguage expressions. e.g. “NOS”, “NEC” , “other”, “with mention of”
- Conceived for off-line coding by professional coders; information is considered already available in a document previously written — without constraints — by physicians

Nomenclature

list of motivated expressions

example: SNOMED system (Systematized Nomenclature in Medicine) it combines elementary terms and complex expressions

- Propensity to consider any expression as "unitary" concept, even very long expressions
- Uniqueness of concept-to-code relationship; each preferred term corresponds to a code
- Complex situations expressed by multiple codes; normally a complex concept is represented by a user-defined combination of codes (the user selects the amount of details he wants to preserve)
- Diverse solutions are available for complex expressions ; most Diseases and Procedures have explicit pointers to “elementary axes” (unique code vs multiple codes); different ad hoc arrangements are normally possible for multiple coding
- Extensive coverage of clinical terminology; SNOMED III has 12 modules; one module is about "modifiers" or "qualifiers"
- Conceived for direct coding by users and case-based processing

Thesaurus

system of predefined atomic descriptors
example: EMTree by Elsevier Publishing, MeSH (Medical Subject Headings), included in UMLS/META-1

- Typically organized by “facets”; large chapters containing tree structures, often using also associative relations; occurence of pragmatical exceptions of different nature and origin
- Descriptors represent elementary concepts to be used in combination among them
- Extensive coverage of clinical field and outside clinical field (e.g. Countries)
- Presence of "position codes". In most cases a single concept can be found under different positions, with specific codes (called “contexts” in UMLS/META-1)
- Conceived for direct usage by users; normally a complex concept is represented by a user-defined combination of descriptors (the user selects the amount of details he wants to preserve)

Usages and tasks for first-generation systems

The performance of first generation systems is poor from a semantic point of view; nevertheless they are able to satisfy most of the needs in paper-based information systems and the basic needs of computer-based applications. Typical components are a list of phrases, a list of codes, a coding scheme to map between items from the previous lists, perhaps hierarchies on phrases or codes. The respective potential usages and tasks are listed in the following table.

Usages and tasks allowed by various components of a first-generation system

component	**usages and tasks**
list of phrases	storing, transmitting, retrieving by string
list of codes	storing, transmitting, retrieving by code;providing key to look-up tables
coding scheme	converting string <--> code;converting code --> group or tabular entry
hierarchy	coding assistance by refinement;clustering of codes

Features of first-generation systems

Paper-based presentation does not support *dynamic re-organization* of rubrics according to different criteria.

Moreover in the past there have been relatively few attempts to *merge*, or make compatible, various systems on the same subject field into multipurpose tools. In fact, paper-based information systems do not demand for large re-use of data. The underlying process is either partially structured (by pre-defined forms, eg. insurance claims), either informal (based on human-to-human communication, eg. ad hoc paper-based or oral communication). Re-use of data for secondary goals goes through human interpretation of phrases in the context of the original documents.

Maintenance is expensive; extensions and adaptations to meet local needs cannot be communicated in a structured way between computers.

Relevant features of first-generation systems are summarized in the following table.

Summary of features of first-generation systems

FIRST GENERATION

• presentation:
  systematic list (typically with codes),
  alphabetical index with permutations
• organisation:
  fixed (typically one, hierarchical)
• purposes:
  devoted to a single application
• flexibility and extension:
  NO; pre-defined list of allowed expressions
• processing on semantics:
  NO; only storage, transmission
  and retrieval of strings and codes

Second generation: compositional systems

The four components of a second-generation system

A terminological system of second generation can be developed only by an iterative process, up to an adequate level of complexity and robustness. The development process produces and refines four complementary results:

1. a *categorial structure* describes semantic categories, semantic links and most relevant structural patterns; an interim release can be produced by a modest amount of resources in a short time; it can be the topic for a standard;

2. a *cross-thesaurus* provides a set of descriptors organized according to the categories in 1.; it requires initially a modest amount of resources to work out most descriptors; complete development could be complex as a standardization initiative, but results refined elsewhere may be included in a standard;

3. a family of structured *lists of classes or phrases*; lists requires continuous maintenance, and phrases require local adaptations; they are not suitable for standardization bodies (except for basic collections, as a "reference taxonomy", a few "reference classifications" and a "reference nomenclature");

4. a *knowledge base of dissection*, where each phrase of 3. is represented by descriptors from the cross-thesaurus 2., according to the structural patterns from 1. Production of the knowledge base has to rely on independent initiatives, from sponsors external to the standardization environment.

The categorial structure is just a synthetic model of the semantic field; it is not imposing how segments of systematic names or fields of databases should be arranged; designers can implement the structure in various ways.

Usages and tasks for second-generation systems

Second-generation systems have some capabilities of semantic processing.

Each component is able to support new potentialities of specialized software (see table below).

Usages of component 3. (ie. the family of lists of classes or phrases) are similar of the ones described in table 3 for components of first-generation systems.

Usages and tasks allowed by various components of a second-generation system

component	usages and tasks
1. categorial structure	• providing metadata to build queries; • providing criteria to organize phrases
2. cross-thesaurus	• providing values for queries; • mirroring hierarchies on dissections
4. knowledge base of dissections	• dynamic cross-referencing of concepts; • suggesting details for structured input; • clustering, indexing phrases by detail; • clustering phrases by pattern

Features of the second-generation systems

New terminological systems, conceived for computer use, process explicit details and are mappable one to the other by comparing dissections (if based on a common cross-thesaurus). Their typical complexity and evolution rate preclude presentation on paper. By appropriate design, second-generation systems can accommodate various first-generation systems, each providing particular functionalities (previously typical of isolated classifications or nomenclatures).

Relevant features of second-generation systems are summarized in the following table.

Summary of features of second-generation systems

SECOND GENERATION

• presentation:
categorial structure + cross-thesaurus + lists (classifications, nomenclatures, ...)
+ knowledge base of dissections
• organisation:
dynamic (multiple, hierarchical)
• purposes:
multiple
• flexibility and extension:
new atoms can be added by users, new combinations can be made by users
• processing on semantics:
(using categorial structure and dissections)
clustering of phrases according to criteria
structured extension of lists
extract and rearrange details
structured input interfaces (suggest details)

Third generation: formal systems

An emerging kind of system for concepts representation is based on **formal models**. It provides a set of symbols (eg. lexical marks on paper or their equivalent in a computer system) and a set of formal rules to manipulate them.

They can be interpreted as denoting a set of concepts and some set of relations between them. Each concept shall be represented in a formal model by a unique canonical form, made up of symbols; non-canonical formal expressions shall be convertible to the canonical one. Symbols can be given names that human users find meaningful, but behaviour of the model shall be independent of names. Models are designed to allow an engine to manipulate these symbols in a formal way, so that symbols behave in ways that correspond to human use of concepts represented.

Formal models are based on a formalism, eg. conceptual graphs and GRAIL. They include all the constructs defined for second-generation systems, but in a continuum. For example, there is no distinction between categories, descriptors, composite concepts: they are all "concepts" that can be manipulated by the engine. Similarly there is no distinction between structural patterns and dissections: they are all manipulated as "canonical forms".

The expressive power of a formal model largely depends on the formalism. In significant systems for practical applications, the complexity of the formalism and of the manipulations goes beyond the power of human processing, and suitable computer software for manipulation and browsing is needed. A formal model is too complex and specialized to be presented directly to users: suitable interfaces are needed to access the model. An appropriate delivery procedure is via a "terminology server" (a software module that communicates with user's applications).

Synergy between second and third generation

The presence of an active engine determines high performances for third-generation systems. By exploiting hierarchies and inheritance, formal systems perform more efficiently all the tasks of second-generation

systems; in addition, they are able in principle to represent all and only the clinically-sensible statements of a subject field and provide for "automatic" validation capability.

But building structured representations with descriptors (convenient for humans assisted by computers) is simpler than developing formal models (convenient for computer processing); modelling of formal systems — of adequate quality for routine use in real settings, at the expected level of performance — requires growth of a cultural context, ie. familiarity of experts with the compositional approach. In addition, large formal models present problems of global coherence and normalization.

For these reasons, diffusion of second-generation systems could be considered as a concrete advantage for formal models. Important issues — regarding any kind of semantic representation — can be discussed and solved by simplified tools of second generation on a less constrained model, and developers of formal systems can concentrate the efforts on issues regarding the formalism and the more robust structure that is required by formal models. Moreover, categorial structure, cross-thesaurus, and knowledge base of dissections are not only results per se, but they can also be used as an **intermediate representation** for a semi-automatic translation into a formal model. This approach is being used for multicenter cooperative production of an integrated model of surgical procedures in GALEN-IN-USE, a demonstration project (1996-1998) following the GALEN project (1992-1995). On the reverse, modelling the knowledge base by formal tools (eg. the ones of GALEN-IN-USE) can provide more coherence and robustness to the knowledge base itself.

Features of the third-generation systems

In summary, features as universality, parsimony, precision present as series of advantages and disadvantages. Advanced performance is balanced by more demanding modelling efforts.

Relevant features of third-generation systems are summarized in the table below.

Summary of features of third-generation systems

THIRD GENERATION

- presentation (internal):
  universal model + engine by combinatorial rules (parsimonious, precise)
- organisation:
  dynamic (multiple, hierarchical)
- purposes:
  multiple
- flexibility and extension:
  new combinations validated by computer (by predefined combinatorial rules)
- processing on semantics:
  complete (formal processing)

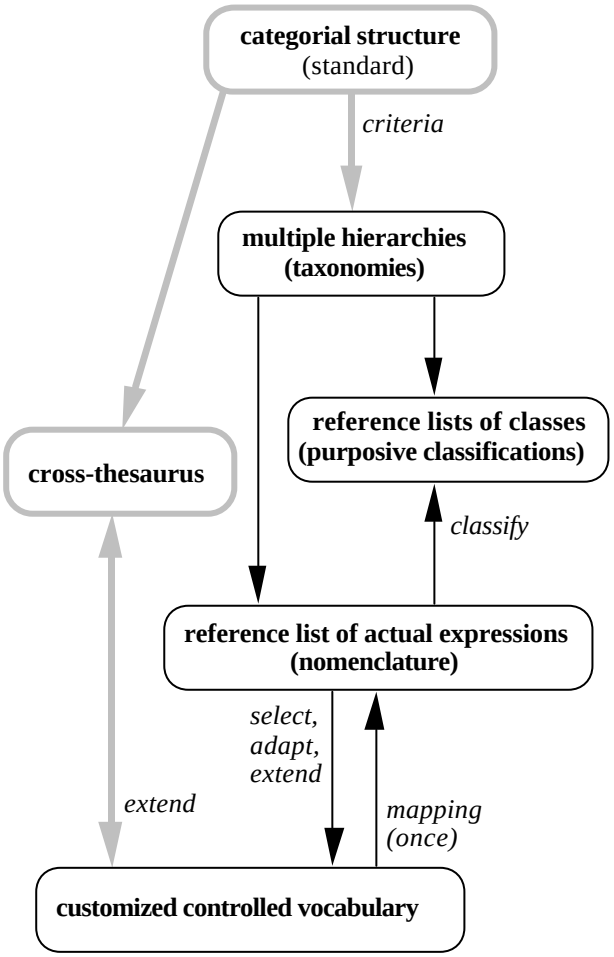
5. Role of standards in the development of terminology systems

Co-existence and convergence among systems on the same subject field

Synergy among generations of terminological systems is not limited to translation of a knowledge base of dissections into a formal model (to refine the former or to feed the latter).

A system of second generation should include a family of lists of first generation, such as classifications and nomenclatures, and can be customized by users. In turn, one or more first-generation systems will benefit from an underlying categorial structure and a cross-thesaurus.

The scenario shown in the figure below could therefore refer either to the distribution of a second-generation system, or to a step in the evolution towards it.



A scenario for a coherent family of first-generation lists developed according to a standard categorial structure and unique cross-thesaurus (only most relevant influences are made explicit)

Maintenance and upgrade of traditional systems

Production and appropriate use of an advanced compositional system (of second generation) can facilitate the work of developers of traditional terminological systems in two ways:

- easier re-organization and maintenance of existing individual paper-based systems;
- harmonization and cross-referencing of diverse existing paper-based systems, and the simplified development of new specialized systems.

In particular, description of the categorial structure of large systems (e.g. ICD, SNOMED, Read codes) allows systematic comparison of such systems — at least to a first approximation — so that steps can be taken towards their future convergence and towards the more systematic design of future systems.

In fact, a relevant application of categorial structures and cross-thesauri is to support spontaneous convergence among first-generation systems on the same topic.

Each terminological system was built for a single specific task. Even where terminological systems addressed the same topic and used similar differentiating criteria, there were usually differences in the order in which differentiating criteria were applied and the degree of detail captured.

Comparing different terminological systems requires therefore making these differences explicit, by providing a description of the categorial structure of each corresponding system of concepts (if possible, against the same standard categorial structure).

This is a significant but not conclusive step, since no description of a categorial structure, however detailed, can cover all the subtle differences in the exact rubrics and in the definitions attached to concepts that reflect different intended usages by different target groups (see the kinds of knowledge that define a concept, described in §1.1).

Developers of traditional systems should become active; they should perceive the motivations to prepare dissections for all the terminological phrases of their system, by recognizing the added value to their own system. Effort should be gradual, always aligned with the perceived benefits. This benefit was already perceived, in our opininon, by developers of a few systems (LOINC, IUPAC, SNOMED-RT, Clinical Terms in UK, ICNP, SDM) and is driving the evolution of their systems.

This activity will result in upgrading each first-generation system, by increasing its regularity and making this regularity more explicit also in more self-explaining phrases. The criterion to judge on adequacy of a dissection (and to limit the effort) should be the uniqueness of each dissection in the context of the whole terminological system, i.e. its ability to identify the related phrase; phrases with identical dissection will be considered as synonyms for that context.

Another result is the possibility to compare similar dissections from different systems. Unnecessary diversity can be removed, residual ambiguities (same dissection but different intended meaning) can be discovered and clarified. This process is actually taking place between LOINC and IUPAC.

As already stressed, the existence of second-generation systems does not mean that applications based on first-generation systems are no more useful for their precisely stated purpose.

Distribution and usage of first-generation systems is easier and cheaper. More sophisticated software — which is needed if one wants to use second generation systems — could be effective in limited contexts where appropriate, in particular to adapt data to variable contexts in telematic applications, e.g. by a “mechanical” transformation of data from a more detailed system to a less detailed one, according to predefined rules.

Production of systematic names and reference lists

Systematic names are terminological phrases created according to pre-established rules (as in chemistry); they could be a reference "lingua franca" for standard communication, but in most cases they are not conceived to be used by healthcare professionals in routine applications, where local adaptations or working

names are preferred. Systematic names can be generated using a categorial structure and a cross-thesaurus according to well-defined (standard) generation rules.

Most of the communication tasks require an explicit reference list of allowed (ie. meaningful) terminological phrases. A reference list may be required also as a "basement" to build look-up tables and knowledge bases (i.e. reference knowledge).

Users (or, better, user groups) can be allowed to customize the reference lists — i.e. to select, adapt and expand a list to obtain a controlled vocabulary for specific purposes —, using descriptors from the cross-thesaurus according to predefined strategies and constraints. The cross-thesaurus can be used to extend the local vocabulary in a controlled manner, and new descriptors can be defined by users and be incorporated into the cross-thesaurus at the next revisions of the related standard.

Note that *standards* should intervene only if this process requires coherence in the development of multiple systems or in data transmission. In principle systematic names and reference lists could be applied to individual systems — even without considering an official standardization process — for example to LOINC and to the Read Thesaurus, i.e. where the categorial structure is explicit and validated by a large amount of disssections.

Note also that a reference nomenclature should collect phrases from real settings, and could be refined using the continuous input from customized controlled vocabularies. The goal of the reference nomenclature should be to provide guidance to avoid unnecessary diversity in local vocabularies.

Increasing coherence among systems on overlapping subject fields

CEN/TC251/WG II- is producing a series of European prestandards on categorial structures in different overlapping subject fields. These prestandards could be used to facilitate a "bottom-up" convergence towards a common framework.

A first step is to use the categorial structure on a field to expand a category in another structure, eg. by using the SNOMED module on <topography> to extend any complex list involving the category of <anatomy>.

In fact, integrating different standards on overlapping topics, independent terminological systems can be in turn systematically integrated and expanded: for example, a nomenclature on <medical devices> could be used as a supplement to:

- a nomenclature on <surgical procedures>, extending its list of <interventional equipments>;
- a nomenclature on <nursing interventions>, extending its list of <objects> or <means>.

Another useful achievement could be the comparison of ontologies in similar "atomic" domains, eg. about <actions> in <nursing interventions> and <deeds> in <surgical procedures>.

More in general, categorial structures on individual subject fields can be compared and eventually integrated into a comprehensive categorial structure. Analogously, cross-thesauri can be merged into a multi-domain thesaurus.

Experts could produce more coherent prestandards, by considering various "top-down" results, namely: the Semantic Network of Meta-1 in the UMLS initiative of the US National Library of Medicine, the studies on the integrated ontology for healthcare and the future European prestandard on the most relevant classes of semantic links, planned with highest priority by CEN/TC251.

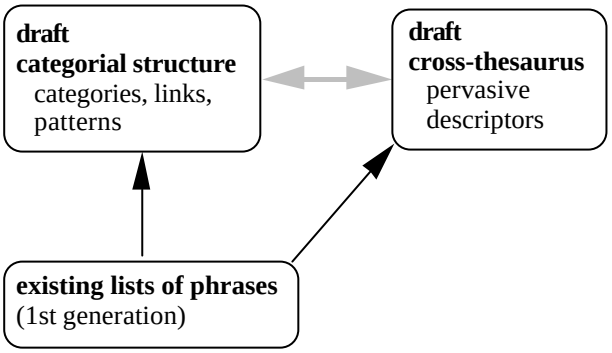
This effort will eventually allow the establishment of a comprehensive "universal" computer-based formal model for representation of healthcare concepts.

Development of second-generation systems

The process to develop a second-generation terminological system is suitable for decentralization and for progressive involvement of an increasing number of developers, and defines responsibilities and roles for standardization bodies, national coding centers, domain experts and terminology experts. The ideal process should consist — in our perspective — of three main phases:

1. extract an initial categorial structure and an interim cross-thesaurus;
2. populate the knowledge base and refine the categorial structure and the cross-thesaurus;
3. process feedback from additional systematic developments.

In phase 1 the experts on the subject field examine one or more existing terminologies (classifications, nomenclatures, ...) and analyze the phrases in order to extract an initial categorial structure.



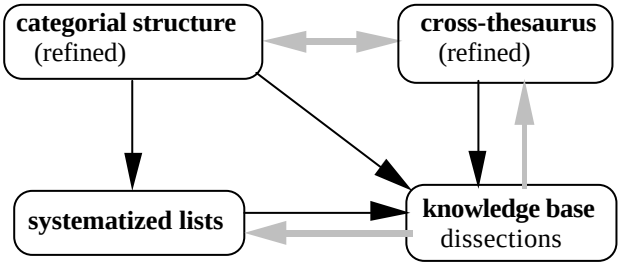
At the same time they extract the candidate descriptors — the most frequent ones, that appear to be useful to organize the system of concepts — and align them with the provisional categories.

Note that this is the kind of work actually performed by Project Teams in CEN, for the production of ENV on categorial structures.

In phase 2 more resources are needed, to populate the knowledge base of dissections according to provisional categorial structure and cross-thesaurus. Dissections should be able to identify each distinct terminological phrase in the context of a source list and to allow comparison of similar phrases from different lists.

A focussed harmonization should take place, by comparing similar dissections, eliminating unwanted differences and making wanted differences explicit.

The systematic work on the knowledge base may be used to validate the previous components: it implies a refinement of categorial structure and cross-thesaurus; also the original lists can be made more coherent, because inconsistencies and irregularities will be discovered and fixed.



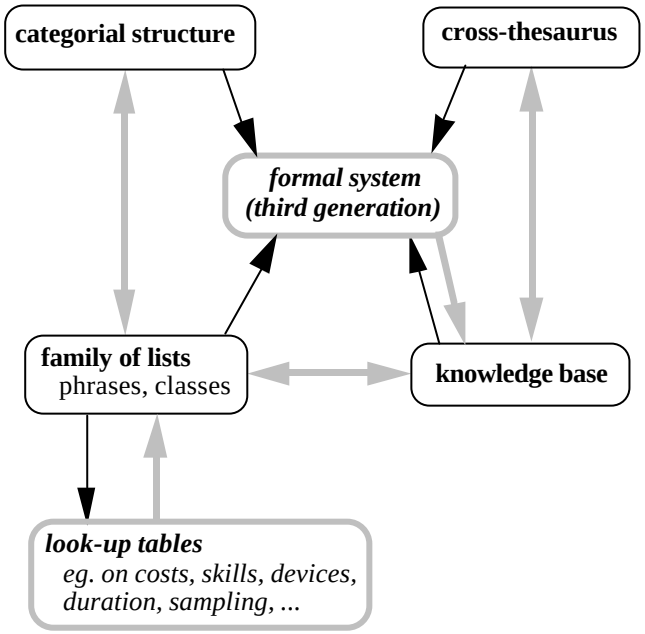
Development can be considered complete and systems are usable; maintenance and field tests by users in their current environments will refine the content of all the components.

As a possible complement to the previous activities, a phase 3 with formal and/or operational exploitation might be performed, respectively by:

- semi-automatic translation into a formal system,
- synchronization between each list of entries on a topic and related look-up tables (i.e. the master files that contain information about each entry relevant for a defined administrative or organizational purpose).

The latter could be a systematic criterion to pragmatically evaluate the completeness of a list and of the related knowledge base of descriptors.

As we will see in the next paragraph, this is not in the remit of standard development and would require a substantial amount of resources. Nevertheless, it could be a crucial proof of the quality of the previous steps.



A scenario on standard-driven developments

In this paragraph we present a synthesis of the activities that are compatible the standardization process in CEN, considering the amount of resources required for design and developments, the years necessary to perform consensual revisions and re-approval of new releases, the organization appropriate for collecting feedbacks and users' reactions.

We consider three levels of involvement for CEN:

1. *production* of a complete and stable standard according to the CEN regular process;
2. *adoption* of a standard developed by external resources, e.g. a by-product of a Research Project of the European Union;
3. *maintenance* of a standard (with a cycle of revision of the order of 3-5 years).

Suitability for CEN to produce, adopt or maintain standards on various components of terminological systems

component	PRODUCTION	ADOPTION	MAINTENANCE
categorial structure	YES	YES	YES
rules for systematic names	YES	YES	YES

detailed list of systematic names	NO	??	NO
cross-thesaurus	interim	YES	??
reference nomenclature	NO	YES	??
reference classification	??	YES	??
knowledge base of dissections	NO	NO	NO
formal models	NO	NO	NO

In principle, CEN standards on terminological systems cannot be directly aimed at end-users (ie. healthcare professionals), but their main target group should be made of developers of terminological systems.

It is proved that usual CEN mechanisms can produce prestandards on categorial structures and cross-thesauri. Rules for systematic names can be also produced. In particular domains, CEN could also produce reference classifications, when starting material is close to a solution.

The relevance of making available a categorial structure as a CEN standard should be considered with relation to two complementary processes:

- the iterative process to develop a terminological system of second generation
- the harmonization of the categorial structure and of the cross-thesaurus across subject fields, to obtain a coherent framework on the whole healthcare sector.

The first process is in particular conceived for decentralization and for progressive involvement of an increasing number of developers, with defined responsibilities and roles for standardization bodies, experts and national coding centers. A standard on categorial structure should be the framework for subsequent efforts (each iteration can be performed with more stable components and thus by a more distributed initiative).

Building a complete and robust system of descriptors (ie. the cross-thesaurus) is resource consuming — therefore initiatives should be external to CEN —, but also powerful to assure coherence across participating initiatives. An interim cross-thesaurus with most frequent descriptors can be developed and approved as standard, to guide further steps. Refinement should be driven by experimental work on existing systems, to produce the knowledge base of dissections.

After a few iterations, the system of descriptors becomes stable and has a slow evolution, consisting mainly in extensions, by adding a limited number of new "building blocks"; it can be revised and approved as CEN standard, because maintenance, after initial debugging, is compatible with the CEN revision cycle.

The adoption of standards on reference lists is possible but the risk of freezing their evolution by lack of timely maintenance is high. In fact, actual classifications and — more — nomenclatures are characterized by a rapid evolution (the extension consists in a lot of new combinations of the building blocks); therefore they require a permanent institution outside CEN, for continuous updating and for validation of extensions.

In our opinion, knowledge bases of dissections, as well as formal models, are not suitable for activities in the standardization bodies.