# Medical Terminology Systems

**Katharina Kaiser**

**Vienna University of Technology, Vienna, Austria**
Institute of Software Technology & Interactive Systems – Information Engineering Group

Asgaard-TR-2007-1 May 2007

Authors: **Katharina Kaiser**

kaiser@ifs.tuwien.ac.at
http://ieg.ifs.tuwien.ac.at

Contact: **Vienna University of Technology, Vienna, Austria**
Institute of Software Technology & Interactive Systems
Information Engineering Group

Favoritenstraße 9-11/188
A-1040 Vienna
Austria, Europe

Telephone: +43 1 58801 18839
Telefax: +43 1 58801 18899
Web http://ieg.ifs.tuwien.ac.at

# Contents

## Abstract

Medical terminologies provide a powerful instrument to both abstract medical information or to enrich it in order to make it better understandable. The various application areas of terminology systems have not only led to a variety of such systems, but also to attempts developing and maintaining systems that support a multitude of application purposes. We provide an overview over commonly used terminology systems for various purposes. Furthermore, we present desirable features such systems should cover in order to satisfy today's requirements.

# Chapter 1

# Introduction

Medical knowledge is very complex and comprehensive. Thus, in the 18th century there were already attempts to classify diseases systematically (e.g., ”London Bills of Mortality” by John Graunt, ”Nosologia methodica” by Francois Bossier de Lacroix, ”Genera morborum” by Linnaeus, ”Synopsis nosologiae” by William Cullen). These first attempts were further developed to clinical vocabularies, terminologies or coding systems. The systems are structured list of terms which together with their definitions are designed to describe unambiguously the care and treatment of patients. Terms cover diseases, diagnoses, findings, operations, treatments, drugs, administrative items and so on, and can be used to support recording and reporting a patient’s care at varying levels of detail, whether on paper or, increasingly, via an electronic medical record.

## 1.1 Definitions

First, we will give a short description of some concepts:

- A **nomenclature** is a relatively simple system of names
- A **vocabulary** is a system of names with explanations of their meanings
- A **classification** is a systematic organisation of things into classes
- A **thesaurus** (such as MeSH) is designed to index medical literature and support search over bibliogaphic databases

But many of the terms used in this field can prove difficult to define accurately, and their use in practice can be inconsistent.

## 1.2 Desirable Features of Computerized Medical Terminology Systems

Application of vocabulary systems has turned out to be a difficult issue due to differences of the controlled vocabularies. In order to ease their application standards have been developed. But their adoption has been slow as system developers indicated that they would not meet their needs. Thus, medical informatics researchers have examined the structure and content of existing vocabularies to determine why they seem unsuitable for particular needs.

Cimino carries some solutions forward by keeping in mind that the desired vocabularies must be multipurpose (i.e., they have to be used for capturing clinical findings, natural language processing, indexing medical records and literature, or representing medical knowledge) [1]:

1. Ensure domain completeness
2. Maintain systematics approaches to updating vocabularies
3. Apply concept-oriented vocabularies where terms must be non-vague, non-ambiguous, and non-redundant
4. Concept permanence, i.e., the concept and the meaning of a concept must remain even if it is inactive or archaic
5. Only use non-semantic concept identifiers
6. Allowing multiple hierarchies (polyhierarchy) to better express concepts in their contexts
7. Maintain formal definitions to be able to support automated vocabulary management, collaborative vocabulary development, and methods for converging distributed development efforts
8. Reject ”not elsewhere classified” terms (that may be established to deal with incompleteness in the vocabulary)
9. Provide multiple granularities to meet the needs of different users
10. Maintain multiple consistent views: ensure consistency over different views of a hierarchy
11. Represent context specific information to maintain the relationship between a concept and the context in which it is used
12. In vocabulary evolution give detailed descriptions of what changed and why (e.g., simple addition, refinement, precoordination, disambiguation, obsolescence, discovered redundancy, minor name changes)
13. Be able to recognise redundancy where the same information is expressed in different ways

## 1.3 Aims of these Systems

Not only the type of knowledge, but also the type how knowledge is stored depends on the goal of the developer. For instance, the clinical focus of ICD is to provide the basis for the compilation of national mortality and morbidity statistics by WHO Member States, SNOMED CT’s clinical focus is to advance excellence in patient care, and MeSH’s clinical focus is indexing documents containing information about healthcare and biomedicine.

But medical coding and classification systems form part of current moves towards implementing a standardized ”language for health”: a common (computerized) medical language for global use. The use of standardized clinical terminologies facilitates collecting, retrieval, and reuse of electronic data for multiple purposes (e.g., disease surveillance, clinical decision support, patient safety reporting).

Following, we will describe some of the most common terminology systems in detail beginning with systems with the purpose of abstraction followed by multipurpose systems.

# Chapter 2

# Coding Systems for Medical Record Abstraction

Many coding systems have been developed with the aim of simplifying the data, converting it to a general form which is easier to manipulate. For instance, patient records can be classified by such codes in various granularities. Such classified records can then be retrieved when cases of certain types are needed (e.g., for statistical measures). The coding represents only a simplified abstract of information extracted from the record and thus this kind of coding is referred to as abstraction.

## 2.1 ICD — International Statistical Classification of Diseases and Related Health Problems

In 1898, the American Public Health Association (APHA) recommended that the registrars of Canada, Mexico, and the United States adopt Jacques Bertillon's Classification of Causes of Death, which was introduced in 1893 and already adopted by a number of countries. Furthermore, the APHA agreed in revising the system every ten years to keep it remained current with medical practice advances. The first international conference to revise the International Classification of Causes of Death convened in 1900 with revisions occurring every ten years thereafter.

In 1948, the World Health Organization (WHO)[1] assumed responsibility for preparing and publishing the revisions to the ICD every ten years. WHO sponsored the seventh and eighth revisions in 1957 and 1968, respectively.

### 2.1.1 ICD-9 and ICD-9-CM

In 1977 the WHO published the ICD-9 containing 6,969 codes [2]. The coding system consisted of three-digit core codes (see Table 2.1 which shows ICD-9-CM, but is completely compatible with ICD-9 at this level). A fourth digit (after the decimal point) provided an additional level of detail. Usually .0 to .7 are used for more specific forms of the core term, .8 is usually used for an ”other” category, and .9 for ”unspecified” category. The arrangement of the terms is based on a strict hierarchical order (see Table 2.2).

Due to immediately following criticism regarding ICD-9's inadequacy for general coding and specific specialty coverage [3], the National Center for Health Statistics (NCHS)[2]

[1] http://www.who.int/

[2] http://www.cdc.gov/nchs/

**Table 2.1:** ICD-9-CM Volume 1 Diagnosis codes

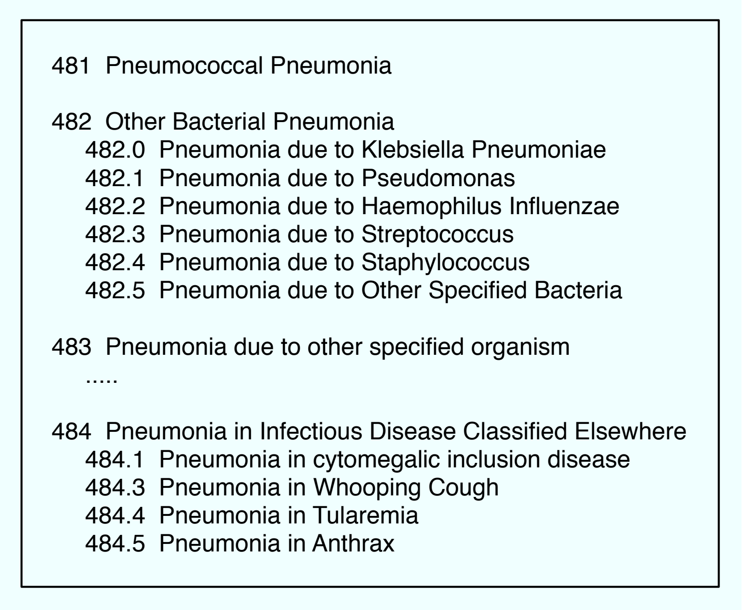
| | |
|---|---|
| 001-139 | Infectious And Parasitic Diseases |
| 140-239 | Neoplasms |
| 240-279 | Endocrine, Nutritional And Metabolic Diseases, And Immunity Disorders |
| 280-289 | Diseases Of Blood And Blood-Forming Organs |
| 290-319 | Mental Disorders |
| 320-389 | Diseases Of The Nervous System And Sense Organs |
| 390-459 | Diseases Of The Circulatory System |
| 460-519 | Diseases Of The Respiratory System |
| 520-579 | Diseases Of The Digestive System |
| 580-629 | Diseases Of The Genitourinary System |
| 630-677 | Complications Of Pregnancy, Childbirth, And The Puerperium |
| 680-709 | Diseases Of The Skin And Subcutaneous Tissue |
| 710-739 | Diseases Of The Musculoskeletal System And Connective Tissue |
| 740-759 | Congenital Anomalies |
| 760-779 | Certain Conditions Originating In The Perinatal Period |
| 780-799 | Symptoms, Signs, And Ill-Defined Conditions |
| 800-999 | Injury And Poisoning |
| E800-E999 | Supplementary Classification Of External Causes Of Injury And Poisoning |
| V01-V86 | Supplementary Classification Of Factors Influencing Health Status And Contact With Health Services |

created an extension of it so the system could be used to capture more morbidity data and a section of procedure codes was added. This extension was called ”ICD-9-CM” [4], with the CM standing for ”Clinical Modification”. The extensions provided an additional level of detail in many places by adding a fifth digit to the code, corresponding to another level in the hierarchy (see Table 2.3) [3]. Referring to the mentioned table we can see that due to the fact that ICD-9-CM is a strict hierarchy every concept is only coded once and its position in the hierarchical tree is not always unambiguously reproducible. A way to overcome this limitation offers MeSH by polyhierarchies.

ICD-9-CM consists of three parts: (1) a tabular listing of diagnosis codes (see Table 2.1 for codes extended for CM), (2) an index for diagnosis codes, and (3) hospital inpatient procedure codes (see Table 2.4).

Several other classification schemes arose due to the deficiencies of ICD-9, such as Yale University’s *Diagnosis Related Groups (DRGs)* for use in prospective payment in the Medicare program [5], or the *International Classification of Primary Care (ICPC)* from the World Organization of National Colleges, Academies and Academic Associations of General Practitioners/Family Physicians (WONCA) [6].

**Table 2.2:** Bacterial Pneumonia coded in ICD-9 (adapted from [3]).



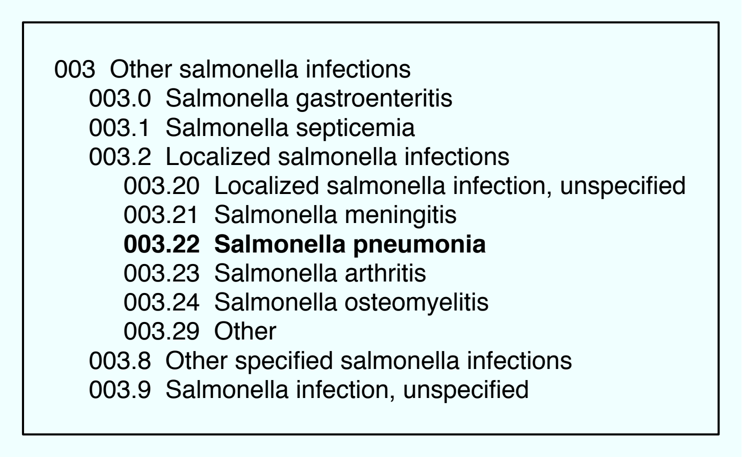
481 Pneumococcal Pneumonia

482 Other Bacterial Pneumonia
- 482.0 Pneumonia due to Klebsiella Pneumoniae
- 482.1 Pneumonia due to Pseudomonas
- 482.2 Pneumonia due to Haemophilus Influenzae
- 482.3 Pneumonia due to Streptococcus
- 482.4 Pneumonia due to Staphylococcus
- 482.5 Pneumonia due to Other Specified Bacteria

483 Pneumonia due to other specified organism
- .....

484 Pneumonia in Infectious Disease Classified Elsewhere
- 484.1 Pneumonia in cytomegalic inclusion disease
- 484.3 Pneumonia in Whooping Cough
- 484.4 Pneumonia in Tularemia
- 484.5 Pneumonia in Anthrax

**Table 2.3:** Example of ICD-9-CM. The four-digit codes are identical to those of ICD-9; the five-digit codes were introduced in ICD-9-CM. Note that Salmonella Pneumonia has been added as a child in 003 rather than in 482 (Other Bacterial Pneumonia) or 484 (Pneumonia in Infectious Disease Classified Elsewhere) (see Table 2.2). Adapted from [3].



003 Other salmonella infections
- 003.0 Salmonella gastroenteritis
- 003.1 Salmonella septicemia
- 003.2 Localized salmonella infections
  - 003.20 Localized salmonella infection, unspecified
  - 003.21 Salmonella meningitis
  - **003.22 Salmonella pneumonia**
  - 003.23 Salmonella arthritis
  - 003.24 Salmonella osteomyelitis
  - 003.29 Other
- 003.8 Other specified salmonella infections
- 003.9 Salmonella infection, unspecified

**Table 2.4:** ICD-9 Procedure codes

| | |
|---|---|
| 0. | Procedures and Interventions, not Elsewhere Classified (00) |
| 1. | Operations on the Nervous System (01-05) |
| 2. | Operations on the Endocrine System (06-07) |
| 3. | Operations on the Eye (08-16) |
| 4. | Operations on the Ear (18-20) |
| 5. | Operations on the Nose, Mouth, and Pharynx (21-29) |
| 6. | Operations on the Respiratory System (30-34) |
| 7. | Operations on the Cardiovascular System (35-39) |
| 8. | Operations on the Hemic and Lymphatic System (40-41) |
| 9. | Operations on the Digestive System (42-54) |
| 10. | Operations on the Urinary System (55-59) |
| 11. | Operations on the Male Genital Organs (60-64) |
| 12. | Operations on the Female Genital Organs (65-71) |
| 13. | Obstetrical Procedures (72-75) |
| 14. | Operations on the Musculoskeletal System (76-84) |
| 15. | Operations on the Integumentary System (85-86) |
| 16. | Miscellaneous Diagnostic and Therapeutic Procedures (87-99) |

### 2.1.2 ICD-10

In 1992 ICD-10 [7] was published by the WHO with an enormous increase to 12,420 codes. It is now used to report mortality; for morbidity ICD-9-CM is still in use. It uses alphanumeric categories instead of numeric (see Table 2.5), it changed chapters, categories, and titles and regrouped conditions.

**Table 2.5:** ICD-10 categories

| Chapter | Blocks | Title |
|---|---|---|
| I | A00-B99 | Certain infectious and parasitic diseases |
| II | C00-D48 | Neoplasms |
| III | D50-D89 | Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism |
| IV | E00-E90 | Endocrine, nutritional and metabolic diseases |
| V | F00-F99 | Mental and behavioural disorders |
| VI | G00-G99 | Diseases of the nervous system |
| VII | H00-H59 | Diseases of the eye and adnexa |
| VIII | H60-H95 | Diseases of the ear and mastoid process |
| IX | I00-I99 | Diseases of the circulatory system |
| X | J00-J99 | Diseases of the respiratory system |
| XI | K00-K93 | Diseases of the digestive system |
| XII | L00-L99 | Diseases of the skin and subcutaneous tissue |
| XIII | M00-M99 | Diseases of the musculoskeletal system and connective tissue |
| XIV | N00-N99 | Diseases of the genitourinary system |
| XV | O00-O99 | Pregnancy, childbirth and the puerperium |
| XVI | P00-P96 | Certain conditions originating in the perinatal period |
| XVII | Q00-Q99 | Congenital malformations, deformations and chromosomal abnormalities |
| XVIII | R00-R99 | Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified |
| XIX | S00-T98 | Injury, poisoning and certain other consequences of external causes |
| XX | V01-Y98 | External causes of morbidity and mortality |
| XXI | Z00-Z99 | Factors influencing health status and contact with health services |
| XXII | U00-U99 | Codes for special purposes |

Several countries have already created their own extensions to ICD-10 (e.g., ICD-10-CA in Canada or ICD-10-AM in Australia). NHCS has received permission to create a clinical modification of ICD-10 (ICD-10-CM for diagnosis codes and ICD-10-PCS for procedure codes), but only draft versions are currently available and there is not yet an implementation date announced.

## 2.2 CPT — Current Procedural Terminology

CPT is a nomenclature used to report medical procedures and services performed by physicians. It is developed by the American Medical Association (AMA), who introduced it in 1966 [8]. The 2006 version of CPT contains 8,568 codes and descriptors. In the US it is used for billing and reimbursement. CPT codes specify information about the codes which differentiates them based on their cost (see Table 2.6 for codes regarding tonsillectomy). CPT codes can also describe information about the reasons for a procedure.

**Table 2.6:** Different CPT codes for tonsillectomy.

| Code | Description |
|---|---|
| 42820 | Tonsillectomy and adenoidectomy; younger than age 12 |
| 42821 | Tonsillectomy and adenoidectomy; age 12 or over |
| 42825 | Tonsillectomy, primary or secondary; younger than age 12 |
| 42826 | Tonsillectomy, primary or secondary; age 12 or over |
| 42960 | Control oropharyngeal hemorrhage, primary or secondary (eg, post-tonsillectomy); simple |
| 42961 | Control oropharyngeal hemorrhage, primary or secondary (eg, post-tonsillectomy); complicated, requiring hospitalization |
| 42962 | Control oropharyngeal hemorrhage, primary or secondary (eg, post-tonsillectomy); with secondary surgical intervention |

The current version is CPT-4. There are three categories of CPT codes:

1. **Category I CPT codes** are designated for services (or procedures) common in contemporary medical practice and being performed by many physicians in clinical practice in multiple locations. For each, there is a five digit code and a text descriptor.

2. **Category II CPT codes** are focused on performance measurement. They are invented to facilitate data collection by coding certain services and/or test results that are agreed upon as contributing to positive health outcomes and quality patient care. This category of codes is a set of optional tracking codes for performance measurement. These codes may be services that are typically included in an *Evaluation and Management (E/M)* service or other component part of a service and are not appropriate for Category I CPT codes. The use of tracking codes for performance measures will decrease the need for record abstraction and chart review, thus minimizing administrative burdens on physicians and survey costs for health plans.

3. **Category III CPT codes** deal with emerging technology. The purpose of this category is to facilitate data collection on and assessment of new services and procedures. These codes are intended to be used for data collection purposes to substantiate widespread usage or in the FDA approval process.

## 2.3 MeSH — Medical Subject Headings

The Medical Subject Headings (MeSH)[3] thesaurus is a controlled vocabulary produced by the National Library of Medicine (NLM)[4] and used for indexing, cataloging, and searching for biomedical and health-related information and documents.

In 1954 the first official list of subject headings published by the NLM appeared under the title *Subject Heading Authority List.* In 1960 the *Index Medicus* was initiated and a new and thoroughly revised Medical Subject Headings appeared.

[3] http://www.nlm.nih.gov/mesh/

[4] http://www.nlm.nih.gov

The MeSH structure can be described by three major components: (1) the Headings themselves, (2) the Subheadings (also known as Qualifiers), and (3) the Supplementary Concept Records. Furthermore, the concepts are connected by various relationships, such as equivalence relationships (these include entry terms and synonyms), hierarchical relationships, and associative relationships.

The 2005 version of MeSH contained a total of 22,568 subject headings, also known as *descriptors*. Most of these are accompanied by a short definition, links to related descriptors, and a list of synonyms or very similar terms (known as *entry terms*). Because of these synonym lists, MeSH can also be viewed as a thesaurus.

**Table 2.7:** Top level of MeSH tree structures

| | **Category** | **Title** |
|---|---|---|
| 1. | [A] | Anatomy |
| 2. | [B] | Organisms |
| 3. | [C] | Diseases |
| 4. | [D] | Chemicals and Drugs |
| 5. | [E] | Analytical, Diagnostic and Therapeutic Techniques and Equipment |
| 6. | [F] | Psychiatry and Psychology |
| 7. | [G] | Biological Sciences |
| 8. | [H] | Natural Sciences |
| 9. | [I] | Anthropology, Education, Sociology and Social Phenomena |
| 10. | [J] | Technology, Industry, Agriculture |
| 11. | [K] | Humanities |
| 12. | [L] | Information Science |
| 13. | [M] | Named Groups |
| 14. | [N] | Health Care |
| 15. | [V] | Publication Characteristics |
| 16. | [Z] | Geographicals |

The descriptors are arranged in a hierarchy and may appear at several places in the tree hierarchy (see Table 2.7 for the top level of MeSH's tree structure). The tree locations carry systematic labels (i.e., tree numbers), and consequently one descriptor can carry several tree numbers. For example, following Fig. 2.1, D stands for Chemicals and Drugs, D02 for Organic Chemicals, D02.755 for Phenols, and D02.755 for Hydroxybenzoic Acids. A second tree number for Hydroxybenzoic Acids is D02.241.511.390. As seen from the graphic, Hydroxybenzoic Acids has three locations in the hierarchy and thus carries three different tree numbers. The tree numbers of a given descriptor are subject to change as MeSH is updated. Every descriptor also carries a unique alphanumerical ID that will not change. The ability of multiple occurrence within a hierarchy is referred to as *polyhierarchy*. In contrast to strict hierarchies (e.g., ICD) concepts can be linked to a multitude of their parents allowing multiple contexts; however, allowing a concept to appear in multiple contexts may lead to some ambiguity about its meaning.

### 2.3.1 Kinds of Concepts

Besides descriptors, which characterize the subject matter or content, MeSH contains also Qualifiers and Supplementary Concepts.

**Headings** (i.e., descriptors) characterize the subject of matter of content. They contain sev-

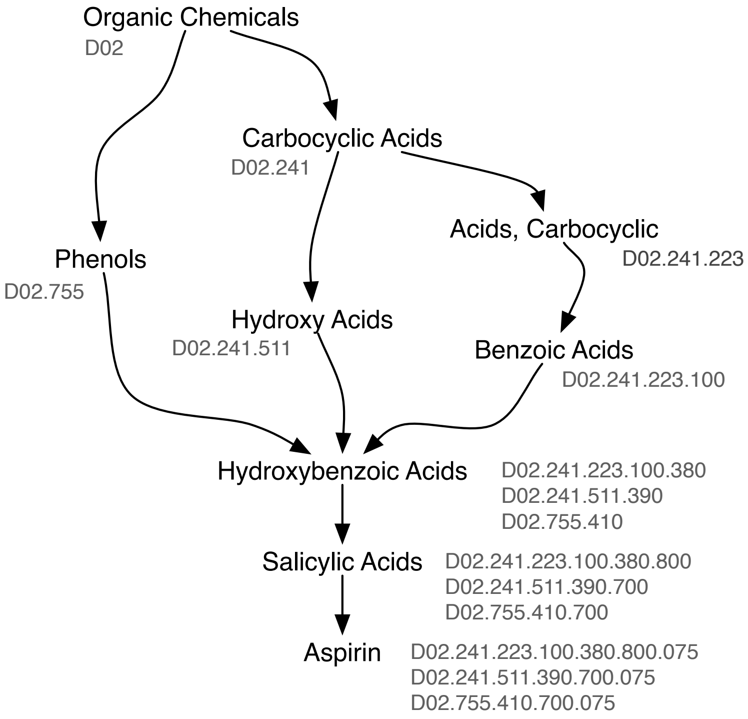


**Figure 2.1:** Hierarchical relationships in MeSH 2007: *Aspirin* and all of its broader terms

eral elements, for instance, the MeSH heading (i.e., the term used in the MEDLINE database), which reflects a meaning, entry terms (i.e., synonyms, alternate forms, and other closely related terms), tree numbers, semantic types (from the UMLS Semantic Network), forward cross references, or pharmacological actions (for chemicals).

**Qualifiers** are used with descriptors and afford a means of grouping together those documents concerned with a particular aspect of a subject. There are 83 topical qualifiers used for indexing and cataloging in conjunction with descriptors. Not every qualifier is suitable for use with every descriptor. They are also grouped by a hierarchy.

**Supplementary Concepts** (SCR) are used to index chemicals, drugs, and other concepts for MEDLINE. They do not have tree numbers, but each SCR is linked to one or more descriptors. Currently, there are over 150,000 SCRs with more than 400,000 terms.

### 2.3.2 Cross References

We can differentiate between ***See* cross-references** and other cross-references. The former are synonyms, alternate forms, and other closely related terms in a MeSH record (e.g., *Abscess, Abdominal* **see** *Abdominal Abscess*). In printed MeSH there exists also a ”backwards cross-reference” (e.g. *Abdominal Abscess* **X** *Abscess, Abdominal*). Additionally, there are three kinds of informative references:

1. **See related** references, also known as ”associative relationships” are used for a variety of relationships between *descriptor* records, where an other descriptor may be more appropriate for a particular purpose.

   *Bone and Bones* **see related** *Osteogenesis*
   for a relationship between an organ and a physiological process

In printed MeSH this reference is labeled ”**XR**” (e.g., *Bone and Bones* **XR** *Osteogenesis*).

2. **Consider also** references to another *descriptor* having related linguistic roots.

   *Brain* **consider also** *terms at CEREBR- and ENCEPHAL-*.

3. **Entry combination** is used if certain *Descriptor/Qualifier* combinations are prohibited. For instance, descriptor *Accidents* cannot be used with qualifier *prevention & control*, but the descriptor *Accident Prevention* should be used instead:

   *Accidents/prevention & control* **see** *Accident Prevention*

MeSH tree structures are hierarchical relationships that can be seen as cross references to broader and narrower terms. This ”relation” is depicted by the *TreeNumber* element (see also Fig. 2.1).

### 2.3.3 Describing Complex Concepts

Complex concepts can be described in three different ways.

1. Coordination – the combined use of two or more separate descriptors. For example, *jejunal enteritis* may be expressed by the use of *Jejunum* and *Enteritis*.

2. Qualifiers can be used in conjunction with appropriate descriptors. A *deficiency of monoamine oxidase* may be indexed as *Monoamine Oxidase/deficiency*. The direct linkage of the qualifier to the descriptor to which it relates avoids the possibility of false coordination that may occur if two descriptors are used to represent a single concept.

3. Many pre-coordinated descriptors are contained in MeSH for frequently encountered subjects. If MeSH has a pre-coordinated descriptor such as *Heart Surgery*, the indexer or cataloger uses it rather than a descriptor-qualifier combination. If a descriptor-qualifier combination is available, it will be used in preference to coordinating two descriptors.

Various online systems provide access to MeSH and the vocabulary is available in several online systems (e.g., the MeSH Browser and the UMLS Metathesaurus with links to many other controlled vocabularies).

# Chapter 3

# Multipurpose Coding Systems

Despite for reimbursement coding schemes are also used with medical records. Since the advent of computers electronic medical records (EMRs) are evolving and their application can be improved by augmenting the information contained with standardized codes. But most of the systems described in the previous chapter failed to be implemented within EMR systems. Amongst other things the schemes were too coarse grained to be applicable to an EMR system's task. For instance, a code "Pneumonia Due to Other Specified Bacteria" is too less detailed to select an appropriate antibiotic [3]. Thus, early EMR systems have developed their own coding system (e.g., PTXT for the HELP system [9], COSTAR Directory [10]).

However, developing and maintaining terminologies demands for great efforts and thus it seems to be more appropriate investing in shareable, reusable, multipurpose coding systems that can be adapted for their applications' needs. These terminology systems should then be enabled to support:

- Classification and coding systems
- Electronic medical records (EMRs)
- Decision support systems
- Knowledge management systems
- Natural language processing

## 3.1 SNOMED CT — Systematized Nomenclature of Medicine – Clinical Terms

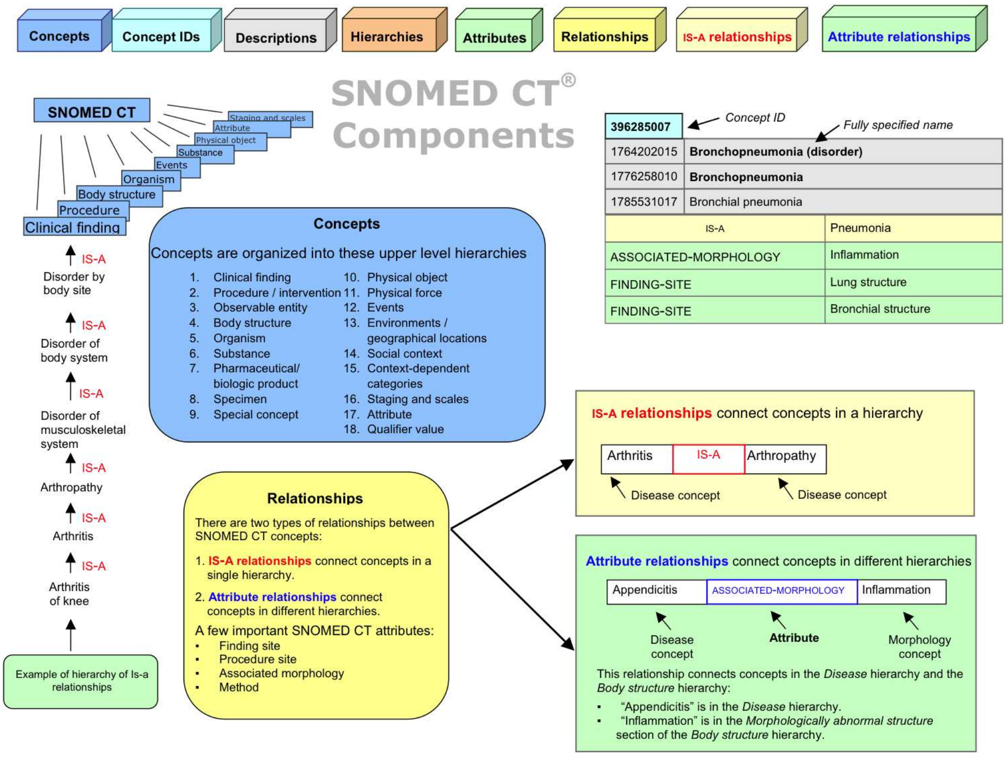
SNOMED [11] is a coding system, controlled vocabulary, classification system and thesaurus. It is a comprehensive clinical terminology; designed to capture information about a patient's history, illnesses, treatment and outcomes.

SNOMED is developed by the College of American Pathologists (CAP) and was introduced in 1977. SNOMED derived from SNOP (Systemized Nomenclature of Pathology) which was introduced in 1965 and expanded this lexicon to include all of general medicine. Further significant revisions included SNOMED-II in 1979 and SNOMED-III (International) in 1993. SNOMED-RT succeeded SNOMED version 3.5 (1998), and included over 340,000 explicit relationships. SNOMED-RT represented a significant update whilst remaining compatible with SNOMED International, presenting data in a completely machine-readable format. What had previously been a relatively flat, multi-axial system became a true semantic network.

SNOMED CT[1] resulted from a merger between SNOMED-RT (Reference Terminology), which is strong in in specialty medicine, and the England and Wales National Health Service's Clinical Terms, a UK-based terminology for primary care previously known as the Read Codes [12]. SNOMED CT is considered to be the first international terminology.

SNOMED CT is probably the most comprehensive medical terminology developed to date and can be used to support patient data capture, transfer, querying and storage via an electronic patient record. It includes a Semantic Net of over 300,000 medical concepts and their relationships. It has multiple axes and hierarchies; at the top level are three main hierarchies (finding, disease, procedure) and 15 supporting hierarchies (see Fig. 3.1). Additionally, over 7 million relationships are defined that can be hierarchical relationships (i.e., IS-A relationship) and attribute relationships that connects concepts in different hierarchies (see also Fig. 3.1).

As the basis for its concept representation, it uses a description logic (KRSS). It includes concepts covering multiple use scenarios: diagnosis, drug definitions, findings, procedures, anatomy, and so on. Many other terminologies can map to SNOMED CT including LOINC and ICD9.



**Concepts**

Concepts are organized into these upper level hierarchies

1. Clinical finding
2. Procedure / intervention
3. Observable entity
4. Body structure
5. Organism
6. Substance
7. Pharmaceutical/ biologic product
8. Specimen
9. Special concept
10. Physical object
11. Physical force
12. Events
13. Environments / geographical locations
14. Social context
15. Context-dependent categories
16. Staging and scales
17. Attribute
18. Qualifier value

| 396285007 | |
|---|---|
| 1764202015 | Bronchopneumonia (disorder) |
| 1776258010 | Bronchopneumonia |
| 1785531017 | Bronchial pneumonia |
| IS-A | Pneumonia |
| ASSOCIATED-MORPHOLOGY | Inflammation |
| FINDING-SITE | Lung structure |
| FINDING-SITE | Bronchial structure |

**Relationships**

There are two types of relationships between SNOMED CT concepts:

1. IS-A relationships connect concepts in a single hierarchy.
2. Attribute relationships connect concepts in different hierarchies.

A few important SNOMED CT attributes:

- Finding site
- Procedure site
- Associated morphology
- Method

IS-A relationships connect concepts in a hierarchy

| Arthritis | IS-A | Arthropathy |
|---|---|---|

Attribute relationships connect concepts in different hierarchies

| Appendicitis | ASSOCIATED-MORPHOLOGY | Inflammation |
|---|---|---|

This relationship connects concepts in the *Disease* hierarchy and the *Body structure* hierarchy:

- "Appendicitis" is in the *Disease* hierarchy.
- "Inflammation" is in the *Morphologically abnormal structure* section of the *Body structure* hierarchy.

**Figure 3.1:** SNOMED-CT components [13]

SNOMED CT is able to apply both pre- and post-coordination. That means that the user can choose a pre-coordinated concept that fits best and can endorse it by attributes (i.e., post-coordination) if a more detailed description is necessary and an adequate pre-coordinated concept is not available.

[1] http://www.snomed.org/snomedct/index.html

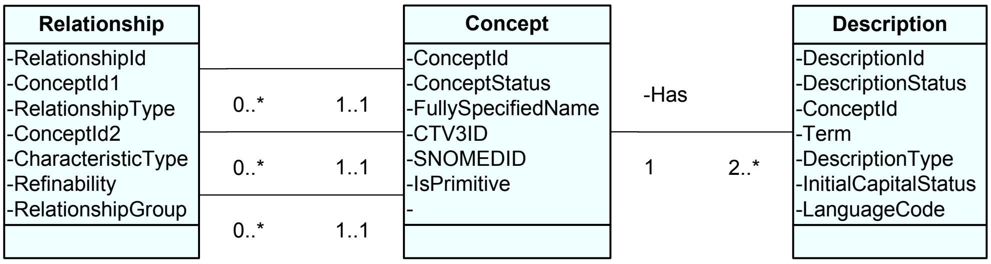


**Figure 3.2:** SNOMED CT table structure [14]

## 3.2 LOINC — Logical Observation Identifiers, Names and Codes

LOINC [15] codes are universal identifiers for laboratory test results and other clinical observations. It has been developed to facilitate particularly the transmission and storing of clinical laboratory results to support clinical care, outcomes management, and clinical research. Nearly 31,000 of around 41,000 observation terms contained within LOINC relate to laboratory testing. LOINC is designed to be compatible with HL7 messages. Each record in the LOINC database identifies a clinical observation and contains a formal 6-part name, a unique name for tests identifying code with check digit, synonyms, and other useful information.

LOINC is developed by the Regenstrief Insitute, an internationally respected non-profit medical research organization associated with the Indiana University, and the LOINC committee.

LOINC codes are not intended to transmit all possible information about a test or observation. They are only intended to identify the test result or clinical observation. Other fields in the message can transmit the identity of the source laboratory and special details about the sample. (e.g., the result code may identify a blood culture, but the message source code can be more specific and identify the sample as pump blood.) The level of detail in the LOINC definitions was intended to distinguish tests that are usually distinguished as separate test results within the master file of existing laboratory systems [16].

Each LOINC record corresponds to a single test result or panel. The record includes fields for specifying:

1. Component (analyte) — e.g., potassium, hemoglobin, hepatitis C antigen.
2. Property measured — e.g., a mass concentration, enzyme activity (catalytic rate).
3. Timing — i.e. whether the measurement is an observation at a moment of time, or an observation integrated over an extended duration of time — e.g., 24-hour urine.
4. The type of sample — e.g., urine; blood.
5. The type of scale — e.g., whether the measurement is quantitative (a true measurement) ordinal (a ranked set of options), nominal (e.g., E. coli; Staphylococcus aureus), or narrative (e.g. dictation results from x- rays).
6. Where relevant, the method used to produce the result or other observation.

### Major ”Parts” of a Test/Observation Name

The fully specified name of a test result or clinical observation has five or six main parts including: the name of the component or analyte measured (e.g. glucose, propranolol), the property observed (e.g. substance concentration, mass, volume), the timing of the measurement (e.g. is it over time or momentary), the type of sample (e.g. urine, serum), the scale of measurement (e.g. qualitative vs. quantitative), and where relevant, the method of the measurement (e.g. radioimmunoassay, immune blot). These can be described formally with the following syntax.

<Analyte/component>:<kind of property of observation or measurement>:<time
aspect>:<system (sample)>:<scale>:<method>

The colon character, ”:”, is part of the name and is used to separate the main parts of the name.

The first part of the name can be further divided up into three subparts, separated by carats (^). The first subpart can contain multiple levels of increasing taxonomic specification, separated by dots (.). The third and fourth parts of the name (time aspect and system/sample) can also be modified by a second subpart, separated from the first by a carat. In the case of time aspect, the modifier can indicate that the observation is one selected on the basis of the named criterion (maximum, minimum, mean, etc.); in the case of system, the modifier identifies the origin of the specimen if not the patient (e.g. blood donor, fetus, blood product unit).

Glucose^2H post 100 g glucose PO:MCnc:PT:Ser/Plas:Qn

mGentamicin^rough:MCnc:PT:Ser/Plas:Qn

mABO group:Type:PT:Bld^donor:Nom

mBody temperature:Temp:8H^max:XXX:Qn

LOINC was the first of a proposed set of five uniform standards for the electronic exchange of clinical health information to be adopted across all US federal agencies (including Departments of Defense, Health and Human Services and Veterans Affairs health care facilities). LOINC is also the standard used by the Centers for Disease Control and Prevention and state health departments for reporting communicable diseases. The US National Committee for Quality Assurance (NCQA) 2005 edition of its Health Plan Employer Data and Information Set (HEDIS) supports the use of LOINC codes for some measures.

## 3.3 GALEN — General Architecture for Languages, Encyclopedias and Nomenclatures in Medicine

The GALEN project [17] evolved from Rector’s PEN&PAD electronic medical record system [18]. Within the project a reference model for representing medical concepts independent of the language being recorded and of the data model used by an EMR system was developed.

The reference model consists of four parts:

1. The high level ontology [19], also referred to as high level schemata, which describes the sorts of concepts and the broad patterns by which they can be composed to produce more detailed concepts.

2. The Common Reference Model (CRM) itself, which consists of the reusable parts of anatomy, surgical deeds, diseases, clinical signs, and so on, their definitions, descriptions, and the constraints on fitting them together. The CRM itself is broad and shallow and expected to be shared by most applications. Its internal structure is highly modular and can be subdivided if needed.

3. Extensions required for specific applications or specific subdomains

4. The model of surgical procedures and other similar models which define composite concepts made up of the parts from the CRM and its extensions.

GALEN provides a model from which concepts can be composed. The classification of composite concepts is automatic and based on formal logical criteria. The concept model is separated from the model of use. Thereby, extensions can be added separately in order to make the conceptual model fit the requirements if the concept model does not correspond to the model of use.

The representation scheme that is used to build the CRM is known as GRAIL – the GALEN Representation And Integration Language [20].

## 3.4 UMLS — Unified Medical Language System

The UMLS is a controlled compendium of a large number of national and international vocabularies and classifications (over 100) and provides a mapping structure between them. Thereby, an important issue of UMLS is to assume continuing diversity in the formats and vocabularies of different information sources and in the language employed by different elements of the biomedical community. It is not an attempt to build a single standard biomedical vocabulary [21].

The UMLS is developed by the National Library of Medicine, USA. The UMLS R&D project was initiated in 1986.

The UMLS is made up of three main knowledge components: (1) the Metathesaurus, (2) the Semantic Network, and (3) the SPECIALIST Lexicon.

### 3.4.1 Metathesaurus

The Metathesaurus contains medical concepts from more than 100 vocabularies. The sources were developed by varying purposes, they have different structures and properties. The sources are thesauri (e.g., MeSH, CRISP, NCI), statistical classifications (e.g., ICD-9-CM), billing codes (e.g., CPT), and clinical coding systems (e.g., SNOMED CT). The large set of sources was incorporated due to different types of information, levels of specificity, or organizational and political boundaries. The Metathesaurus enables the continuing diversity of the formats and vocabularies of different information sources instead of building a single standard vocabulary [21].

Besides medical concepts it also contains inter-concept relationships whose primary purpose is to map between coding systems (see Fig. 3.3) and provide information exchange between different clinical databases and systems (see Table 3.1). It is organized by concept or meaning, and each concept has specific attributes that define the meaning. Identical or almost identical concepts are linked together with hierarchical context from the different vocabularies and relationships between the concepts are explained and represented. If a new vocabulary is added to the Metathesaurus its constituent terms are linked whenever possible to existing concepts. The Metathesaurus contains 5 million concept names identifying more than 1 million biomedical concepts and can be used to overcome problems caused by discrepancies in different terminologies [3, 22].

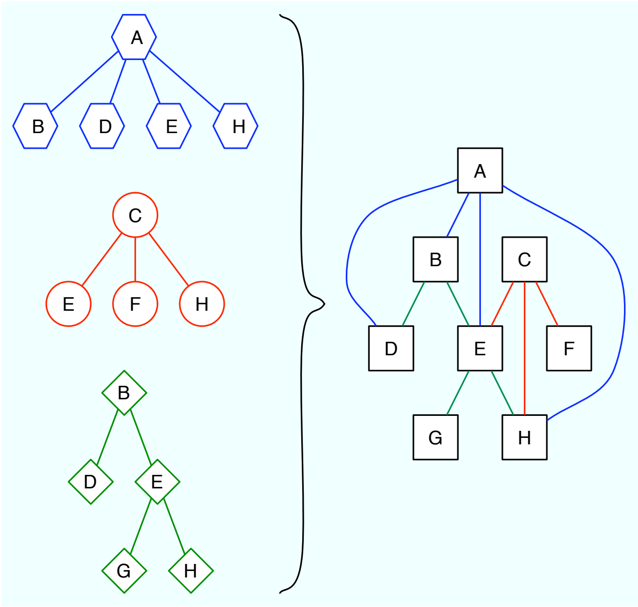


**Figure 3.3:** Metathesaurus: Combining structures from different source vocabularies together to a (directed) graph.

**Table 3.1:** Metathesaurus: Synonymous terms are clustered into a concept. Preferred term is chosen and unique identifier (CUI) is assigned.

| | | | |
|---|---|---|---|
| Addison’s disease | Metathesaurus | PN | |
| Addison’s disease | SNOMED CT | PT | 363732003 |
| Addison’s Disease | MedlinePlus | PT | T1233 |
| Addison Disease | MeSH | PT | D000224 |
| Bronzed disease | SNOMED Intl 1998 | SY | DB-70620 |
| Deficiency; cortocorenal, primary | ICPC2-ICD10 Thesaurus | PT | MTHU021575 |
| Primary Adrenal Insufficiency | MeSH | EN | D000224 |
| Primary hypoadreanlism syndrome, Addison | MedDRA | LT | 10036696 |

**C0001403 Addison’s disease**

### 3.4.2 Semantic Network

Within the UMLS project an upper-level ontology was developed, the so called Semantic Network. It provides an overarching conceptual framework for all UMLS concepts [23]. The Semantic Network [24, 25] specifies categories to which medical concepts defined in the Metathesaurus can belong and the semantic relationships that can be assigned between these concepts and their attributes. Thus, it helps to orient users to the vast knowledge content of the Metathesaurus [26]. 135 semantic types and 54 semantic network relationships exist.

The Semantic Types serve as *high level categories* assigned to each Metathesaurus concept, independent of its position in its original source hierarchy.

There are major groupings of semantic types including organisms, anatomical structures, and the like. The links among semantic types provide the structure for the network and show important relationships between the groupings and concepts (see Figurer 3.4). The primary link between semantic types is the *isa* link. This primary link establishes a hierarchy to decide the most specific semantic type to assign to a Metathesaurus concept.

The network also has five major non-hierarchical relationship categories – categories of what are called ”associational” relationships – (1) *physically related to*, (2) *spatially related to*, (3) *temporally related to*, (4) *functionally related to*, and (5) *conceptually related to*. The semantic type information includes an identifier, hierarchy, definition and its associated relationships.

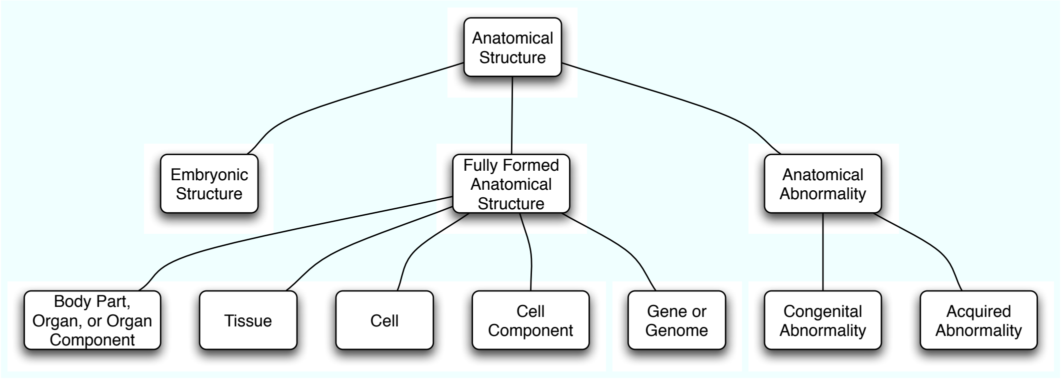


**Figure 3.4:** Semantic Network: Anatomical Structure

### 3.4.3 SPECIALIST Lexicon

The SPECIALIST Lexicon contains syntactic (i.e., how words are put together to create meaning), morphological (form and structure), and orthographic (spelling) information for biomedical and common words in the English language. ”The Lexicon and its associated lexical resources are used to generate the indexes to the Metathesaurus and also have wide applicability in natural language processing applications in the biomedical domain.” [27]. The SPECIALIST Lexicon includes over 200,000 lexical items.

In order to support application developers with lexical variation and text analysis when using the UMLS the *Lexical Systems Group* of the *Lister Hill National Centre for Biomedical Communications*[2] has developed the SPECIALIST NLP tools. There are three tool packages available: (1) the Lexical tools, (2) the Text tools, and (3) the Spelling tools.

The Lexical tools are designed to manage lexical variations. They use the SPECIALIST lexicon for their purposes. The Text tools are used to analyze free text. They can tokenize strings into words, terms, phrases, sentences and sections. The Spelling tools are used to find close or related terms from an index.

Based on the SPECIALIST NLP tools NLM has developed the *MetaMap* program [28]. MetaMap is designed for biomedical researchers and maps Metathesaurus concepts to free text. It consists of five modules: (1) Parsing, (2) Variant Generation, (3) Candidate Retrieval, (4) Candidate Evaluation, and (5) Mapping Construction. Within these modules the input text is parsed and tokenized, lexical variants are generated from the resulting phrases, Metathesaurus candidates are retrieved who match the variants, and after evaluation by a final mapping algorithm the candidates highest ranked are returned.

[2] http://lhncbc.nlm.nih.gov/

### 3.4.4 MetamorphoSYS

The Metathesaurus merges a multitude of vocabulary sources. A lot of information and many concepts occur many times in the Metathesaurus due to their multiple appearance in various sources. Furthermore, the database containing all sources is very large. But it is not feasible using the Metathesaurus with all sources, as this would lead to an enormous processing time on the one hand and delivering too many results on the other hand. Thus, one can limit the sources contained in her own UMLS version by using the MetamorphoSYS. This application can be used to configure the Metathesaurus regarding license restrictions, language restrictions, input and output options, source vocabularies, precedence or suppressibility of a source, attribute and relationship filters, and semantic type filters.

### 3.4.5 Conclusion

The UMLS is not a single medical terminology, but provides access to multiple terminology systems and mappings among them. Furthermore, it serves as a superior description of the terminology by connecting the concepts with its Semantic Network.

The UMLS can be understand as an ontology — a system describing instances, concepts, attributes, and relations. But other than common ontologies it has a two-level structure:

1. The Metathesaurus, a unified collection of many different medical terminologies, a compilation of terms, concepts, relationships, and associated information

2. The Semantic Network containing semantic types (one may think of semantic types as high-level concepts, i.e., broad categories), organized in a hierarchy of IS-A links.

Lee and Geller call such a structure a *Terminological Knowledge Base (TKB)*: ”... any structure that consists of (1) a semantic network of semantic types; (2) a thesaurus of concepts; and (3) assignments of every concept to at least one semantic type” [29].

# Chapter 4

# Conclusions

A very large number of coding and classification systems have been developed for healthcare. Their development has been driven by different and specific goals.

### Completeness, Comprehensiveness

On the one hand, many systems have been designed mainly to support administation (e.g., billing). So they have typically included, for example, only a limited number of diagnosis codes for each encounter and can lose clinical information. Widely-used but essentially administration-oriented system, such as ICD, have been mandated by government agencies and/or payor organizations but capture clinical data at an insufficient level of detail to support clinical needs that lie outside the limited range of activities they were designed to support.

On the other hand, systems designed to cover clinical information have tended to cover a relatively narrow subset of healthcare, such as nursing procedures or problem lists. Some systems that concentrate on coding fine-grained primary clinical data have been proprietary, custom-built, limited, difficult for clinicians to use and have resulted, in some cases, in low user acceptance.

Thus, we can say that existing medical vocabularies vary in their coverage and completeness, but many classifications overlap. Nevertheless, it can be difficult to compare clinical coding systems. Content, structure, completeness, detail, cross-mapping, taxonomy, definitions, clarity vary between existing vocabularies. Interoperability is a significant problem.

### Semantics

Although recent emergence of description logic encoded medical terminologies – particularly SNOMED CT – exists, many of the established systems lack a precise semantic substantiation. It is expected that medical terminologies encoded in such a way have the potential to facilitate transition to the Semantic Web.

### Structure

Medical terminologies are evolving from relatively ”simple code-name-hierarchy structures, into rich, knowledge-based ontologies of medical concepts” [30].

### System Integration

Comprehensive clinical terminology systems are needed to help integrate patient data with health information technologies such as electronic medical records.  SNOMED CT aims to help structure and computerize the medical record but needs to be used correctly and consistently to preserve data quality and maximize shareability.

# Acknowledgements

This work is supported by ”Fonds zur Förderung der wissenschaftlichen Forschung FWF” (Austrian Science Fund), grant L290-N04.

# Bibliography

[1] James J. Cimino. Desiderata for controlled medical vocabularies in the twenty-first century. *Methods of Information in Medicine*, 37(4-5):394–403, Nov. 1998.

[2] World Health Organization. *International Classification of Diseases, 9th revision*. World Health Organization, Geneva, 1977.

[3] James J. Cimino. Coding systems in health care. *Methods of Information in Medicine*, 35(4/5):273–284, 1996.

[4] Commission on Professional and Hospital Activities. *International Classification of Diseases*. Ninth Revision, with Clinical Modifications (ICD-9-CM). United States National Center for Health Statistics, Ann Arbor, 1978.

[5] 3M Health Care. AP-DRGs: all patient diagnosis related groups. 3M Health Care, updated annually.

[6] H. Lambert and M. Wood, editors. *International Classification of Primary Care*. University Press, Oxford, 1987.

[7] World Health Organization. *International statistical classification of diseases and related health problems, 10th Revision*. World Health Organization, Geneva, 1992.

[8] American Medical Association. Current Procedural Terminology. The Association, updated annually.

[9] Gilad J. Kuperman, Reed M. Gardner, T. Allan Pryor, and Bruce I. Blum. *Help: A Dynamic Hospital Information System*. Computers and Medicine. Springer Verlag, New York, 1991.

[10] G. Octo Barnett, N.S. Justice, M.E. Somand, J.B. Adams, B.D. Waxmand, P.D. Beaman, M.S. Parent, F.R. Van Deusen, and J.K. Greenlie. COSTAR — a computer-based medical information system for ambulatory care. *Proc. of the IEEE*, 67(9):1226–1237, Sept. 1979.

[11] R. Coté, D. Rothwell, J. Palotay, R. Beckett, and L. Brochu, editors. *The Systematized Nomenclature of Medicine*. SNOMED International, Northfield, IL: College of American Pathologists, 1993.

[12] NHS Centre for Coding and Classification. *Read Codes, Version 3*. NHS Management Executive, Department of Health, London, 1994.

[13] SNOMED International. SNOMED CT. http://www.snomed.org/snomedct/index.html (last assessed: March 6th, 2007), 2006.

[14] SNOMED International. SNOMED CT user guide. Technical report, College of American Pathologists, Jan. 2007.

[15] S. M. Huff, R. A. Rocha, C. J. Mcdonald, G. J. De Moor, T. Fiers, W. D. Bidgood, A. W. Forrey, W. G. Francis, W. R. Tracy, D. Leavelle, F. Stalling, B. Griffin, P. Maloney, D. Leland, L. Charles, K. Hutchins, and J. Baenziger. Development of the logical observation identifier names and codes (loinc) vocabulary. *J Am Med Inform Assoc*, 5(3):276–292, 1998.

[16] Regenstrief Institute, Indianapolis, IN. *Logical Observation Identifiers Names and Codes (LOINC®) Users' Guide*, December 2006.

[17] Jeremy Rogers, Angus Roberts, Danny Solomon, Egbert Van der Haring, Christopher Wroe, Pieter E. Zanstra, and Alan L. Rector. GALEN ten years on: Tasks and supporting tools. In Vimla L. Patel, Ray Rogers, and Reinhold Haux, editors, *Proceedings from the Medinfo 2001 World Congress on Medical Informatics*, volume 84 Studies in Health Technology and Informatics, pages 256–260. IMIA, IOS Press, 2001.

[18] W.A. Nowlan, Alan L. Rector, S. Kay, B. Horan, and A. Wilson. A patient care workstation based on user centred design and a formal theory of medical terminology: PEN & PAD and the SMK formalism. In *Proc. of the 15th Annual Symposium on Computer Applications in Medical Care (SCAMC'91)*, 1991.

[19] Alan L. Rector, Jeremy E. Rogers, and P. Pole. The GALEN high level ontology. In *Proc. of the 14th International Congress of the European Federation for Medical Informatics (MIE-96)*, Copenhagen, Denmark, 1996.

[20] Alan L. Rector, S. Bechhofer, Carol Goble, Ian Horrocks, W. Nowlan, and W. Daniel Solomon. The GRAIL concept modelling language for medical terminology. *Artificial Intelligence in Medicine*, 9(2):139–171, Feb. 1997.

[21] Betsy L. Humphreys and P.L. Schuyler. The Unified Medical Language System: Moving beyond the vocabulary of bibliographic retrieval. In N.C. Broering, editor, *High-performance medical libraries: advances in information management for the virtual era*, pages 31–44. Meckler, Westport, CT, 1993.

[22] Betsy L. Humphreys and D.A. Lindberg. The Unified Medical Language System project: A distributed experiment in improving access to biomedical information. *Methods of Information in Medicine*, 7(2):1496–1500, 1992.

[23] Alexa T. McCray. An upper level ontology for the biomedical domain. *Comparative and Functional Genomics*, 4:80–84, 2003.

[24] Alexa T. McCray. UMLS Semantic Network. In *Proc. of the 13th Annual Symposium on Computer Applications in Medical Care (SCAMC'89)*, pages 503–507, 1989.

[25] Alexa T. McCray and W. Hole. The scope and structure of the first version of the umls semantic network. In *Proc. of the 14th Annual Symposium on Computer Applications in Medical Care (SCAMC'90)*, pages 126—130, 1990.

[26] Alexa T. McCray and Stuart J. Nelson. The representation of meaning in the UMLS. *Methods of Information in Medicine*, 34:193–201, 1995.

[27] Stuart J. Nelson, Tammy Powell, and Betsy L. Humphreys. The Unified Medical Language System (UMLS) project. In Allen Kent and Carolyn M. Hall, editors, *Encyclopedia of Library and Information Science*, pages 369–378. Marcel Dekker, 2002.

[28] Alan R. Aronson. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proc. of the AMIA Symposium*, pages 17–21, 2001.

[29] Yugyung Lee and James Geller. Semantic enrichment for medical ontologies. *Journal of Biomedical Informatics*, 39(2):209–226, April 2006.

[30] James J. Cimino. Terminology tools: State of the art and practical lessons. *Methods of Information in Medicine*, 40(4):298–306, 2001.