From Terminologies to Classifications – the Challenge of Information Reduction

Hans Rudolf STRAUB[a], Maurus DUELLI[a], Norbert FREI[b], Hugo MOSIMANN[a] and Annette ULRICH[a]
[a] *Semfinder AG, Kreuzlingen, Switzerland*
[b] *Interstate University of Applied Sciences of Technolog NTB, Buchs, Switzerland*

Abstract.
A description of a medical cases can – as any statement about reality – contain more or less information. The aim of a classification is to express as much as possible with a minimum of words (classes). For this purpose the information contained in a terminology must be reduced. Is such a reduction an obvious process? In this paper we examine this question by considering practical aspects arising from the task of "teaching" computers automated ICD-10 coding of diagnoses in text form.
We first assess the extent of information reduction and then discuss the path along which this reduction takes place. The role and conditions of a true hierarchical structure are discussed, as well as the questions that stem from reduction of the many semantic dimensions to the single dimension of a formal hierarchy. Special attention is given to the *sum/summands problem*, a major challenge for automated classification in practice.
Are medical classifications necessary at all? Just because extracting class information from terminological data is not self-evident, the classification holds information which is not otherwise available.

1. Introduction

The information available about a medical case, a patient, is always less than the information that could theoretically be found at the moment of observation of the real case. The language that we use to describe the patient can be differentiated according to several characteristics: the divide between ontological and epistemological viewpoints has recently been discussed [1,2] and the discussion looks set to continue. In this paper we do not emphasize this distinction, but we do look more closely at the question of granularity, e.g. of the information content of a language.

"Language" is used in a broad sense in this context and includes "free" natural languages, standardized and structured terminologies like SNOMED CT (with a fine granularity and a large information content) and classifications like ICD-10 (with a coarse granularity and a poor information content). Of course, nobody believes that it would be possible to extract information in a fine granular language (terminology) from the information found in the terms or codes of a classification. But is it – on the other hand – possible to go in the other direction and assign a case to a classification with the aid of the terms in the terminology alone? At first glance this seems self-evident. A closer look, however, reveals general problems.

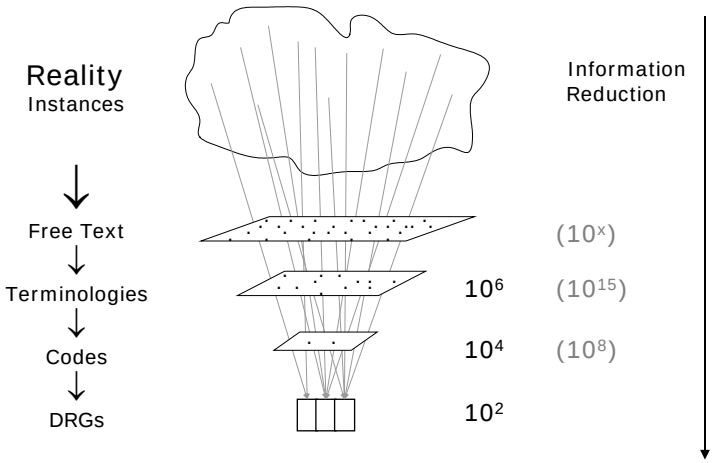

Figure 1: From Reality to Terminologies to Classifications

2. Information reduction

2.1. From reality to observation

In reality, every hair of a patient can be counted. But this information is not what the physician wants to know. Nor does he want to know the condition of every single red blood cell. It is sufficient for him to know that most appear to be normal and that they occur in numbers within a certain range of normality. If anaemia is present, it is not necessary to know all the details of the single cells; it is sufficient to know their condition and numbers in general terms. Obviously only a very small part of the information relating to the real case is observed by the physicians, nurses and laboratories, yet this is not a shortcoming, but a desirable outcome, since we do not need every single piece of information to cure the patient. Too much information would confuse the observer and he wouldn't be able to see the wood for the trees.

The fact that the look is of limited closeness implies a reduction of information, but closeness is not the only aspect. Also the *direction* of the look means a selection of what is possibly observed. This selection is intended, too. The complaints of the patient direct the views of the medical professionals. When he complains about acute abdominal pain, the doctor will most probably not perform a CT scan of the head.

All in all, the reduction of information content from reality to observation is obviously huge.

2.2. From observation to medical records

Not every observation is worth recording and of course only a small part of the information in doctors' and nurses' heads finds its way into medical records. Information in the records can be in pictures, in numbers (quantitative) or in words (qualitative). For purposes of this paper we confine ourselves to the words, they carry the qualitative information, which is the main scope of terminologies and medical classifications.

2.3. From medical records to diagnoses

The diagnoses are usually a small part of the information in the medical record.

2.4. From diagnoses to codes and DRGs

Again there is a reduction of the amount of information and again this reduction is intended. The fewer codes or DRGs (diagnosis related groups) there are, the easier it is to compare cases statistically in groups.

2.5. Estimation of the information content on each layer of granularity

In Figure 1 the information content of the layers of granularity is estimated roughly. The number of permitted instances in the layers provides an estimate of the information content of a selected instance (selective information content according to Shannon [10] and MacKay[6]).

DRGs usually amount to several hundred groups and usually include less than a thousand groups. The ICD-10 has roughly 15,000 codes, depending on the version in question. SNOMED CT contains more than 1 million terms. Compared to these still small numbers, the information content of a medical case is impossible to quantify in reality. In Figure 1 it is shown as a cloud, which represents the huge amount of information as well as its lack of form at this stage of interpretation.

The number in brackets (and the points in the three quadrilaterals representing the interpretation layers) in Figure 1 reflect the fact that, although there are several ICD-10 codes for one case, there is by definition just one DRG for the same case. The information content of the single ICD-10 codes is multiplied and the information content of the whole is the product of the contents of the single codes. In Figure 1 we assume that each DRG has two codes. This is of course a rough estimate. Not every combination of codes is possible, but usually there are more than two diagnostic and therapeutic codes per case.

What is true for the codes is true for the terms. Many terms combine to give one code. Not every term is used for ICD-10 coding. Therefore not only is the information content of one SNOMED clinical term reduced to one ICD-10 code, but several terms in the medical record lead to just one code.

2.6. Amount of the information reduction

As can be seen from Figure 1, the amount of information explodes when we go from the bottom (DRGs) to the top (free text in the medical record). The information in the real case (cloud) is again much richer than the information in the medical record (we shan't offer a quantitative estimate at this point). In the other direction, from the real case to the codes and the DRGs, the information content of the medical case is radically reduced.

2.7. The coding process

Our group creates programs for automated ICD-coding with computers. The installations are designed around an inference machine, which reads the free text (noun

phrases) and produces ICD-10 codes. If the input is not precise enough, the program requests the missing information in the form of a context specific multiple-choice question. As an internal representation language we use concept molecules [12,14], which permit precise and structured modelling of the descriptive [6] information content of the words in the physicians' natural language as well as the information contained in the ICD-10 codes.

3. Is the result of the coding process naturally deducible?

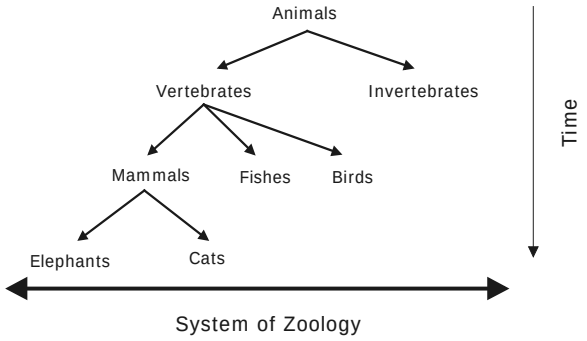

Figure 2: A True Hierarchy

3.1. Deduction in a hierarchical tree

A hierarchy (Figure 2) has two conditions: *disjunctivity* and *unidirectionality*.
Disjunctivity means that the siblings on each level are mutually exclusive. If a mammal is a dog, it cannot be a cat at the same time.

Unidirectionality in a hierarchy means that the branchings go in only one direction: mammals can be differentiated as dogs, cats, cows, elephants, etc. However, this differentiation cannot apply in the other direction: elephants are mammals and can never be fish. If a hierarchy were not strictly unidirectional, it would contain ring structures and would not be a hierarchy, but a *net*.

If the two conditions apply, we have a true hierarchy and this means that we can easily make conclusions based on the leaves of the hierarchical tree back to the branches: if we know that the subject is an elephant, we can conclude that it is a mammal and that it is a vertebrate. Furthermore we can pass the properties of the elements in the upper layer to those in the lower layers. The elephant inherits all the properties of mammals as well of those of vertebrates.

This is a stroke of luck for knowledge representation: we don't need to show all the information about elephants, dogs, cats etc. again for each species, as it is sufficient to show the common information just once at the upper level. This saves space in the representation and makes maintenance easier and more transparent.

A hierarchical tree is therefore ideal for knowledge representation purposes.

Properties are passed from the root to the leaves, from coarse granular to fine granular levels. Class information, however, is deduced in the opposite direction, from fine granular to coarse granular levels (elephant → mammal). This deduction is self-evident in a hierarchical tree, *but is dependent on the two conditions explained above*.

A natural deduction of this kind from fine to coarse granular levels would be exactly what we are striving for in the coding process described in Section 2.7. If the information reduction "funnel" in Figure 1 could be designed as a hierarchy, we could easily deduce the identity of a medical case on the coarse level from its description on the fine granular level. In other words, we could safely deduce the ICD-10 code from the description of the case in medical terms without external assistance.

Is this possible?

3.2. Difference between the zoological system and the system of diseases

Unfortunately the system of diseases cannot be arranged naturally in a hierarchical tree. The reason for this is linked to the two conditions required for a hierarchy: disjunctivity and unidirectionality both apply naturally in the case of animals and plants but are absent in the case of diseases.

In zoology the disjunctivity condition is naturally guaranteed by the fact that two species cannot mix (*species barrier*). Because cats and dogs cannot have offspring together, the two species are definitively disjunct.

The unidirectionality condition is based on the history of the evolution of species. Since species have evolved along the unidirectional time line, this evolution cannot be reversed. Elephants cannot evolve into fish in the future.

The evolution of zoological species is, however, a special case in nature. The fact that this system is in the form of a perfect hierarchical tree is due to the history behind its evolution.

Such a history is absent in the development of diagnoses. Diseases do not evolve from other diseases as zoological species evolve over time from more ancient species.

Certainly diseases are related to each other. One disease can lead to another. But these relationships are much more complicated than the ones in zoology.

Because the two conditions, disjunctivity and unidirectionality, are not present in the system of diseases, this system does not occur naturally in the form of a hierarchy. If we want to make it into a hierarchy for practical reasons – and there are good reasons for this! – we have to create it artificially. As soon as a structure is artificial, however, its shape can be altered and becomes arbitrary.

Statistical methods (variance reduction with regard to a target variable) can be used to perfect a system as Fetter has done in creating the first DRG systems [5]. Such statistically created systems are designed to serve specific needs (economic ones in the case of DRGs), they are artificial and do not have natural and unpassable boundaries like the above described species barriers in zoology.

The ICD-10 classification is designed as a hierarchy. This does offer many advantages, but we have to remember that its structure is arbitrary – however well designed it may be. If we want to assign ICD-10 codes to diagnoses, we must reduce the complex information of the real case diagnoses until it fits into the artificial hierarchical tree. What information gets lost? Additional rules – inclusiva et exclusiva – are necessary for this task.

If we try to obtain ICD-10 codes automatically from natural language diagnoses, we can see how more complex structured information is arranged in a hierarchical tree. In the next section I intend to show how this is done.

4. ICD-10 coding of arterial hypertension

4.1. Semantic dimensions (degrees of freedom)

If we want to code a diagnosis, we first have to analyse the characteristics used in the target coding system. The terms used to describe the codes are best arranged in groups of the same semantic "flavour".

Terms of the same "flavour" represent tokens of the same type. Usually, for each "flavour", just one token can be assigned independently to a diagnosis, so that the diagnosis has as many tokens assigned to it as there are "flavours". The "flavours" can be seen as semantic dimensions, as axes in a semantic space or as degrees of freedom, the latter in order to express the independence of each dimension. They are related, but not completely identical, to the partitions and features (qualities) of the semantic web [9]. The exact differences between the methods of partitioning in the semantic web and the here depicted semantic dimensions as well as the consequences of these differences must be the subject of an additional paper yet to appear.

The semantic dimensions or axes form a multifocal [13] grid, which we use in the concept molecules method [12,14] as a basis for representing term meaning and semantic reasoning. The semantic dimensions do not give rise to ontological statements. They are just a way of representing the meaning of the physicians' words and the classes in the coding systems and serve as a technical medium for translating from language to code.

If we look at the actual ICD-10 German Version 2006 [4] we can identify the following five semantic classes or dimensions used for the partitioning of the diagnosis hypertension:

- Cause: primary, renovascular, renoparenchymatous, endocrinous
- Malignancy: malignant, benign
- Involvement of organs: with cardiopathy, with nephropathy, without such involvement
- Symptoms of organic involvement: with cardiac insufficiency, with renal insufficiency, without such symptoms
- Hypertensive crisis: with crisis, without crisis

The five dimensions are now inserted into the one dimension, which is available in ICD-10 (see Figure 3).

After merging the five dimensions into one, we notice that some combinations which are theoretically possible are represented, but not all:

- With/without crisis can be added as a fifth digit to every four-digit hypertension code, even though this is not shown throughout Figure 3 in order to save space.
- Organic insufficiencies can only be assigned to hypertensions with organ involvement. This is obvious, since there must be a basis for the insufficiency. The two involved axes are not in fact independent. (Nevertheless it is of

practical use to represent them as individual, but dependent axes. The underlying architecture of the concept molecules allows them to be modelled in this way.)

- It is not possible to represent the malignancy of secondary hypertension. This fact is excusable considering that malignant hypertension in Germany is very rare, since blood pressure controls are frequent and hypertension is usually detected before malignancy develops. As secondary hypertension is rare, the combination of secondary and malignant hypertension will be very exceptional and the inability to represent it is no great loss.
- Organic involvement and cause of hypertension cannot be shown at the same time. Again, it can be argued that, due to the symptoms of the primary disease, secondary hypertension will be detected early, usually before secondary organic diseases develop. Therefore it is not essential to represent the combination of secondary hypertension with organic complications.

The way the five dimensions are distributed in one hierarchy is typical of the way one-dimensional coding systems represent multi-dimensional semantics.

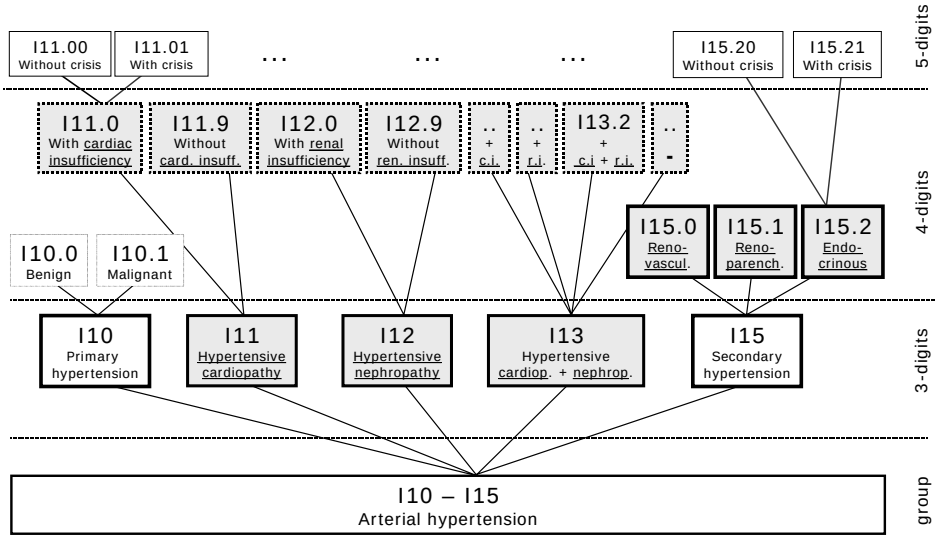

Figure 3: Arterial Hypertension in ICD-10 German Version 2006
(The five semantic dimensions are shown with different boxes)

4.2. Information loss in ICD-10 classes

Figure 3 shows what happens when we move from the fine granular to the coarse granular level: we definitely lose information. If we move from the 5-digit to the 4-digit level, we lose the information about hypertensive crisis; if we move from the 4-digit to the 3-digit level, we lose the information about specific causes of secondary hypertension, or about insufficiencies of organic functions.

This information loss is no surprise, merely the anticipated result of moving to a level with coarser granularity. We could still assume that we will have access to predictable information at the 3-digit level, and that therefore the transition from terminology to classification is not an issue. Unfortunately this is not true. A closer look at Figure 3 reveals further effects of information reduction:

1 The semantic degrees of freedom are arbitrarily placed in the coding system hierarchy. It is not decreed by nature that one degree of freedom should be effective at the 3-digit and the other at the 4-digit level and not the other way round. The sequence of semantic classes is an arbitrary characteristic of the coding system. Then again, a coding system will of course have its own characteristics. As long as we know its structure, we can make conclusions about classification from the terminology without any problems.

2. The real sticking point, however, lies in codes which represent a combination of two or more diagnoses (such occurencies are underlinded in Figure 3). If we compare the meanings of the codes I10.0 and I11.0 at the top left corner, we see that I10.0 adds the information "benign" to I10, whereas I11.0 adds the information "cardiac insufficiency" to I11. Now, unlike "benign", "cardiac insufficiency" is a diagnosis in its own right – and quite an important one, a major killer in the civilized world. "Cardiac insufficiency" therefore has an ICD-10 code of its own, namely I50.-, which can be differentiated by the addition of 2 extra digits in ICD-10 German Version 2006.
   I11.0 is in fact a combination of two diagnoses that can both stand alone, too. What's more, I11.0 is not an isolated case. In Figure 3 all code specifications which can be added to the diagnosis of hypertension, but could be diagnoses in their own right in other cases are underlined. Indeed, they all have their own ICD-10 code when they occur without hypertension. Situations of this kind, where two or more individual diagnoses (*summands*) can form one code (*sum*) when occurring together, are very frequent, not only in the case of hypertension. The problems that arise as a result of such situations are discussed in the next section.

4.3. The sum/summands problem

Figure 4 shows several diagnoses that can cause or can be caused by arterial hypertension. Each of these diagnoses can occur as a summand in a compound diagnosis (sum). This fact leads to open questions with respect to the coding system.

If I11.0 is the sum of two diagnoses, namely I11 (hypertensive cardiopathy) and I50 (cardiac insufficiency), do we have to specify the latter code as well when we code I11.0 (cardiac hypertension with cardiac insufficiency)? Do we refer to sums or summands?

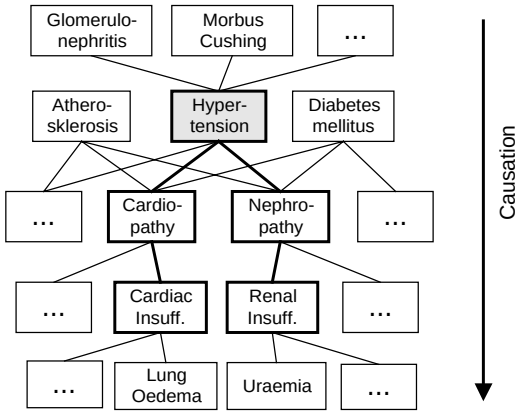

Figure 4: Causal Relationships around Arterial Hypertension

Let's examine the consequences of adding the second (summand) code.

Advantage:

If we add the code I50, we do not lose the clinically important information "cardiac insufficiency", when we move from the 4-digit (I11.0) to the 3-digit level (I11). More importantly, by adding the second code it is possible to provide additional information about the form and severity of the cardiac insufficiency. This is not possible with code I11.0, which is already complete in itself, but works well with the fourth and fifth digit of I50. The fifth digit for example provides information about the severity of the cardiac insufficiency (NYHA grading), which is not unimportant from a clinical and epidemiological viewpoint.

Disadvantage:

If we add the code, we have two codes to deal with instead of one. Which one comes first? Which one should we leave out if further coding simplifications force us to reduce the case to one code?

To give 2 codes instead of 1 is not elegant, especially as there is an obvious redundancy, as the information contained in the fourth digit of I11.0 is already given by I50.-.

Can the system to which the codes are transferred operate with two codes instead of one? The second code may just be omitted and become useless.

There is a further practical problem: how does the coder know whether he should give two codes or one? Should he use the sum or the summands or both, sum and summands?

The decision cannot be based on the hierarchal structure of the coding system, but must be specified explicitly in the coding manual or in additional guidelines. Contrary to the situation in other countries, e.g. in Switzerland, the German ICD-10 Version 2006 and the G-DRG System offer rich guideance [3,4] for coders, a wealth which must, however, be studied carefully before coding correctly.

A general rule for dealing with the sum/summands problem cannot be given; each case must be treated individually. In Figure 5 we show how to treat some hypertension combinations according to the explicit German regulations. As the sum/summands problem occurs very frequently (Figure 3), only the most obvious cases are regulated, while the rest are left to the judgement of the coders.

When coding using computers, we have to include the explicit and implicit regulations in the knowledge base for the computer system. Creating such rules is one of the major tasks facing the knowledge engineers working on the coding system.

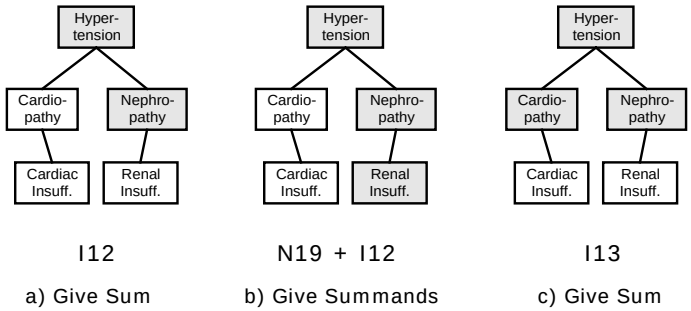

Figure 5: Sum or Summand? – Explicit Guidelines Regulate Correct Coding

4.4. Consequences for the deduction of classification codes from standardized terms

Do ICD-10 codes follow automatically from medical diagnoses? Can the information at a coarse granular level be deduced from that at a fine granular level?

As a consequence of the sum-summands problem we not only lose information of peripheral value. The root of the code hierarchy itself is affected:

If a hypertensive nephropathy with renal insufficiency has to be written as the sum N19 (renal insufficiency) + I12.0 (hypertensive nephropathy), with N19 as the leading code, as specified by the German regulations (Figure 5), and we try to move to the top level of granularity, we find an "N" as the first digit of the first code. This means that we are in the urogenital diseases section and not in the section of vascular diseases where we would expect to find arterial hypertension. The root itself is affected. (This is in fact the intention of the regulation; renal insufficiency is regarded as clinically more important than hypertension and therefore should be given priority.)

This situation is not comparable with the one in the top level in a true hierarchy (Figure 2). If we consider the elephant or any other animal, the information at the top level is always distinct and unchangeable.

As a result of the sum/summands problem, it is not possible to change from the fine granular level to the coarse granular level without doubt or additional guidelines.

These guidelines must be published. A human coder must learn the guidelines. Machine coding systems also have to "learn" them, which means that they must be capable of reading the guidelines as algorithms. Since the safe maintenance of such machine-readable algorithms is one of the major tasks facing any team operating a machine coding system, the flexibility and readability of the algorithms is a priority [14].

4.5. Are classifications necessary?

ICD-10 classifications are simplifications of the real world. Couldn't we avoid the above problems by doing without classifications altogether? Wouldn't it be enough to speak the richer, fine granular language of medical terminology instead of the poor, coarse granular language of classifications? Is coding necessary at all?

I think it is. If I want to find a book in a library, it must be categorised beforehand. If we want to learn about a new field, we must gain an overview of the field, which basically means simplification. Our heads are not capable of holding all the information of the real world. Information reduction can thus be seen as one of the major tasks of human intelligence.

Whenever we want to compare two systems, we have to group their respective cases and this means simplification and classification. Statistical comparisons, e.g. of different therapeutic regimes in distinct groups of patients in medicine, are not possible without grouping.

Generally speaking, the problems of information reduction occur whenever we move from a fine granular to a coarse granular information level, be it from terminologies to classifications or from reality to terminology. Information reduction occurs in all these transitions and the problems described are a consequence of information reduction in general, not of medical classification in particular.

If it were not difficult to obtain classification information from terminological information, classification information would be self-evident and would have no additional value. As it is *not* self-evident, it *does* have additional value.

5. Discussion

Underlying the sum/summands problem are part-of relations, which do not merely connect attributes to diagnoses, but the diagnoses *themselves*. In such cases a diagnosis can be both a summand of another sum diagnosis or an unconnected diagnosis in its own right. This leads to classification confusion.

Schulz exemplifies the same problem with procedures. In [7] he analyses the sum procedure "incision of stomach + removal of foreign body" and its representation in SNOMED CT. In this example, too, the critical relationships link top scope entities (procedures) in the form of sum/summands complexes, as is the case with the top scope entities (diagnoses) in our examples. Such a situation is most vulnerable when it comes to the number and identity of the top entities.

Schulz discusses the representation of such cases in the SNOMED CT system. Relationships are not sufficient to model the complex situation and the SNOMED relationship groups [11] are used to structure the set of relationships. Is it sufficient to provide the relationships as a set and how complex can the set structure be organized? We regard the structuring of the relationships as the crucial step when dealing with the sum/summands problem in practice. Concept molecules [12,14] are built in order to arrange complexes of several relationships in a well-defined structure and can thus be used for solving the sum/summands problem.

6. Conclusions

1. A description of a coarser granularity in a complex field cannot automatically be obtained from a description with finer granularity.
2. Thus ICD-10 codes cannot be obtained from case information in the form of standardized terms without explicit additional rules.
3. The sum/summands question must be given careful attention whenever coding systems are used.
4. A machine coding system must give the same careful attention to the problem.
5. Classifications are necessary, even if standardized terminologies offer more precise details on medical case information.

References

[1] Bodenreider O, Smith B, Burgun A. The Ontology-Epistemology Divide: A Case Study in Medical Terminology. *Proceedings of FOIS 2004:* pp. 185-195.

[2] Burgun A, Bodenreider O, Jacquelinet Ch. Issues in the Classification of Disease Instances with Ontologies. *Proceedings of MIE 2005*: pp. 695-700.

[3] Deutsche Krankenhausgesellschaft DKG, InEK gGmbH. *Deutsche Kodierrichtlinien, Version 2006.*

[4] Deutsches Institut für Medizinische Klassifikation und Information DIMDI. *ICD-10-GM Version 2006.*

[5] Fetter RB, Brand A, Dianne G (eds.). *DRGs, Their Design and Development*. Health Administration Press, Ann Arbor, 1991.

[6] MacKay DM. *Information, Mechanism and Meaning*. MIT, Cambridge, 1969.

[7] Schulz S, Hahn U, Rogers J. Semantic Clarification of the Representation of Procedures and Diseases in SNOMED CT. *Proceedings of MIE 2005*: pp .773-778.

[8] Rector AL. Clinical Terminology: Why is it so hard? *Methods Inf Med.* 1999. 38(4-5): pp. 239-52.

[9] Rector AL. *Representing Specified Valiues in OWL: "Value partitions" and "value sets"*. http://www.w3.org/TR/2005/NOTE-swbp-specified-values-20050517/ 2005.

[10] Shannon CE, Weaver W. *The Mathematical Theory of Communication*. Illinois Press, Urbana / Chicago, 1963 (1948).

[11] Spackman KA, Dionne R, Mays E, Weis J. Role Grouping as an extension to the description logic of ONTYLOG, motivated by entity modelling in SNOMED. *Proceedings of AMIA 2002:* pp. 712-716.

[12] Straub HR. *Das interpretierende System*. Z/I/M-Verlag, Wolfertswil, 2001.

[13] Straub HR. Four Different Types of Classification Models. In Grütter R. *Knowledge Media in Healthcare*. Idea Group Publishing, Hershey / London, 2002.

[14] Straub HR, Frei F, Mosimann H, et. al. Simplified Representation of Concepts and Relations on Screen, *Proceedings of MIE 2005:* pp. 799- 804.