Methods for auditing medical terminological systems
Cornet, R.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Introduction
The domain of health care is not only data intensive but is also knowledge-intensive. Consequently, ways have been sought to capture various types of knowledge with the use of computerized systems. Definitions of terms form one important type of knowledge that facilitates the exchange of information amongst computerized systems and between people and computerized systems.

When people exchange information, they use natural language terms that have meaning to them. The meaning of these terms makes it possible for humans to perform reasoning. For example, someone who knows the meaning of the term “viral hepatitis” – i.e., an inflammatory disease of the liver, caused by a virus – can infer that viral hepatitis is a liver disease, a viral disease, an inflammatory disease, and so on. This knowledge can also be used to decide whether or not an individual patient is suffering from viral hepatitis. In order to facilitate computer-based support for this kind of reasoning, the meaning of terms needs to be represented.

Terminological systems are systems that contain terms pertaining to a certain domain. Some of these systems also represent the meaning of these terms, by providing either free-text or formal definitions. Whereas a free-text definition is useful for people to understand the meaning of a term, formal representation of definitions facilitate automated support based on these definitions.

In health care many terminological systems have been developed, such as the International Classification of Diseases (ICD) [1] and the Systematized Nomenclature of Medicine (SNOMED) CT [2]. These systems are used for a variety of purposes, such as recording clinical information of patients and supporting epidemiological research. It is crucial that the quality of the knowledge represented in these terminological systems is high in order to be trusted for their intended purpose. The problem is that auditing, i.e., judging the quality of, terminological systems is hard because of their size and complexity.

The aim of this thesis is to assess and improve the quality of medical terminological systems. To this end we seek answers to these global questions of:

- how to characterize medical terminological systems,
- how to exploit the benefits of formal representation for auditing, and
- what are the merits of the characterization and auditing methods.

Our approach towards answering these questions involves two steps. The first step is to devise a method to compare terminological systems, based on the properties that characterize them. Such a method is needed as the range of terminological systems is broad, and their characteristics are often unclear or poorly defined. The most important characteristics pertain to their content [3]. The next step focuses on auditing the content of terminological systems (i.e., the concepts representing terms within the system, as well as any definitions thereof). Three types of audit are distinguished: content coverage analysis, detection and explanation of inconsistent definitions, and detection of duplicate
and equivalent definitions. Detection and explanation of inconsistent or duplicate definitions rely on the formal representation of description logics and the reasoning facilities they offer.

This thesis contributes to a better understanding of medical terminological systems and the role that description logics can play in auditing their content. It additionally introduces and describes new methods to audit terminological systems.

In the remainder of this chapter we present short introductions to the main topics of interest of this thesis. In Section 1.1 terminological systems are put in context. Their use in health care is addressed in Section 1.2. Section 1.3 provides an introduction to logic-based representation and reasoning. Auditing is covered in Section 1.4. Finally, a roadmap for this thesis is presented in Section 1.5.

1.1 Ontologies and Terminological Systems

There is nothing new in attempting to categorize the world around us. Plato (427 BC – ca. 347 BC) and Aristotle (384 BC – 322 BC) are among the first known to study logical and ontological structures. Logical structures provide a means for expressing propositions, e.g., “All Greeks are men”, “All men are mortal”, and provide a basis for reasoning, e.g., to draw the conclusion that “All Greeks are mortal”. “Ontology” originally refers to the study of being or existence as well as the basic categories thereof. This “study of being” endeavors to find out what entities and what types of entities exist. For example, in “Categoriae” Aristotle introduces 10-fold classification of that which exists. These categories consist of: substance, quantity, quality, relation, place, time, situation, condition, action, and passion. Further categorization can be done according to the principle of “definitio per genus proximum et differentia specifica” (definition by the nearest higher class and differentiating properties). An example of such a categorization from the philosopher Porphyry (ca. 232 AD — ca. 304) is shown in Figure 1.1.

The recent pervasion of computers and their interconnection through internet provide impetus for giving a new direction to this area of research; many efforts are undertaken to build systems that represent knowledge in a wide variety of areas and making them widely accessible. Tributary to ancient philosophy, such knowledge bases are commonly called “ontologies”. In [4], an ontology is defined as “an explicit specification of a conceptualization” and, more verbosely, as “a specification of a representational vocabulary for a shared domain of discourse, i.e., definitions of classes, relations, functions, and other objects.” In these definitions, a conceptualization is the set of concepts that are used to represent the part of reality or knowledge that is of interest to a community of users. Specification refers to the language and vocabulary terms that are used to specify the conceptualization [5].

Debates are ongoing on the kinds of knowledge-based systems that can be regarded an ontology, i.e., whether requirements might be posed for example
1.2. Terminological Systems in Health Care

Categorization of recorded patient information has been promoted already by Hippocrates (460 BC – 377 BC), who has been so influential that he is considered the founding father of medicine. Hippocrates stated that the medical record should describe the course as well as possible causes of disease. This
information can be used for a variety of purposes. First, it provides a means to facilitate recollecting a patient status at a later time, and to create a medical history of a patient. Next, this information can also be used for epidemiological purposes, i.e., to give insight in the incidence and prevalence of diseases in a large population.

For centuries, terminological systems have played an important role in recording and categorizing patient information. One of the earliest examples are the Bills of Mortality, that were compiled in London in the seventeenth century. These bills provided a weekly categorial overview of “diseases and casualties”. It was recognized that composing such overviews required a way of classifying patient cases. This led to the development of the first version of the International List of Causes of Death (ICD), which was published in 1893.

Over time, many terminological systems were developed, for wide ranges of domains and purposes. This is clearly demonstrated by the Unified Medical Language System (UMLS) Metathesaurus, which interrelates the concepts from over one hundred terminological systems in health care. Nowadays, the systems included cover among others anatomy, diseases, genes, and procedures, for domains such as nursing, dentistry, and primary care. This amounts to more than 1 million concepts, designated by 5 million natural language descriptions in English and other languages.

Availability of such a large number of terminological systems not only brings freedom for selecting the most appropriate system for a given (clinical) setting, but also poses a problem: how to assess suitability and quality of terminological systems for use in (clinical) practice? This thesis addresses this problem by describing and investigating a number of auditing methods and their application.

In order to be able to perform auditing, it is essential to know how terminological systems can be characterized. For such a characterization, insight in

---

Fig. 1.2: The semiotic triangle, depicting the relationships between objects (“referents”), terms (“symbols”), and concepts (“thoughts” or “references”).
the properties of terminological systems is needed. Research on characterization of terminological systems has been performed during the last decade [3, 8–10]. In Chapter 2 we build on this body of work by developing a framework that provides a means for capturing characteristics of terminological systems in a structured and standardized way.

The importance of the content of terminological systems is evident; clinicians must be able to adequately record information using these systems, and the recorded information may be used for a variety of purposes. This usage requires the content of the systems to be both complete and correct, in terms of the represented concepts, as well as the natural language representations that designate these concepts. The question is however, how the contents can be audited. In this thesis, we distinguish user-oriented auditing from knowledge modeler-oriented auditing. In Chapter 3, methods for both types of audit are described, and put into perspective. The presented measurements for content coverage provide a way to reproducibly determine the extent to which a terminological system covers the intended domain. In addition, this chapter bridges user-oriented auditing and knowledge-modeler-oriented auditing. Two methods for auditing the correctness of contents of terminological systems are presented. One method relies on domain experts judging the correctness of the contents, the other method involves an automated approach.

Development of an automated approach to supporting the auditing process of terminological systems is driven by the increasing size and complexity of these systems. Contemporary systems such as SNOMED-CT [2] typically contain tens of thousands to hundreds of thousands of concepts. As such systems are far too large to be comprehended by individuals, we resort to logic-based support that is provided by dedicated applications.

1.3 Description Logics

Terminological systems have not only increased in number and size over time, but also in complexity. Earlier systems generally presented concepts in a hierarchical ordering, possibly providing free-text definitions. The primary medium for distributing the systems was paper. So-called second and third-generation systems [11] provide multiple hierarchies and (formal) definitions of concepts. These systems are no longer distributed on paper, but via internet or on digital media such as CD-ROM or DVD.

If concepts are formally defined, the definitions can be represented using for example frames [12] or description logic [13]. Frame-based representation supports an intuitive way of knowledge modeling in which a concept is represented as a frame where its (characteristic) attributes are represented as slots of that frame. However, the semantics of frame-based expressions generally leave room for interpretation, which hinders automated reasoning. For example, a frame-based statement “telephone has color black” can be interpreted among others as: “all telephones are fully black”, “all telephones are at least partially black”, or “telephones are usually black”. Logic-based representation of knowledge has
1. Introduction

a number of prominent advantages over frame-based representation. First, the semantics of logic-based expressions are unambiguous. Second, as a result of this unambiguity, reasoning can be performed automatically, most notably satisfaction and subsumption testing. Satisfaction testing refers to the inference of logical consistency of definitions. Subsumption testing refers to determining whether one concept is a specialization of another concept, in the same vein of the earlier example, where it was inferred that all Greeks are mortal. In order to keep this kind of automated reasoning computationally tractable, restrictions have to be posed on the logic-based representation, resulting in a relatively limited expressiveness as opposed to free text. Another drawback is the inflexibility of logic-based representation in comparison to frames which allow for ad hoc special-purpose extensions and exceptions.

As logic-based representation supports inference based on formal definitions, it is essential that the definitions are correct and complete. If a terminological system contains definitions that are incorrect, this might also lead to incorrect or missed inferences. The question arises as to whether and how formal representation and reasoning can adequately contribute to auditing terminological systems.

In order to answer this question, we propose a method in Chapter 4, which is based on migration of a frame-based representation to a description-logic-based representation. Whereas the semantics of frames leave room for interpretation, description logic statements have explicit semantics. This implies that, in order to be able to migrate from frames to description logic, the semantics need to be interpreted. The proposed method is based on an interpretation that facilitates the detection of conflicting definitions by means of adding so-called closure axioms. These closure axioms state for example that Male and Female are disjoint concepts, i.e., no individual can be an instance of both concepts.

Without closure axioms, many incorrect definitions can not be detected. Using closure axioms reveals definitions that are incorrect, either due to incorrect interpretation in the migration process, or due to a truly incorrect frame-based representation.

Description logic reasoners such as FaCT++\(^2\) or RACER\(^3\) can detect definitions that are logically inconsistent, by finding so-called unsatisfiable concepts (i.e., no individual can exist that is an instance of such a concept). These reasoners, however, do not provide an explanation for unsatisfiability of a concept, hence it is left to the modeler to find the conflicting definitions, and determine the conflict within the definitions. This can be very hard, especially in large and complex terminological systems.

Chapter 5 addresses the problem of debugging terminological systems, i.e., the process of pinpointing the culprits in definitions and explaining causes for unsatisfiable concepts. A method to support this process is introduced and the results of implementation and application of this method are presented.

As mentioned above, satisfiability testing and inference of subsumption com-

\(^2\) [http://owl.man.ac.uk/factplusplus/](http://owl.man.ac.uk/factplusplus/)

\(^3\) [http://www.racer-systems.com](http://www.racer-systems.com)
prise common reasoning tasks. In order to infer subsumption, the reasoning process relies on concepts that are defined with necessary and sufficient conditions. For example, the concept MOTHER can be defined with the necessary and sufficient conditions FEMALE and PARENT. According to this definition, every concept that is defined as FEMALE and PARENT (possibly with additional conditions), can be inferred to be a specialized concept (called a subsumee) of MOTHER, or equivalent to MOTHER.

In medicine many concepts (such as “person” and “disease”) can not be completely defined with a set of necessary and sufficient conditions [14]. As a result, an assessment is required to determine whether or not concepts with necessary conditions can be further specified by adding conditions, aiming at definitions which are as complete as possible. It was shown that terminological systems suffer from large numbers of concepts for which no condition is specified at all, i.e., the genus is specified (according to the Aristotelean principle) but not the difference [15]. For example, in SNOMED CT, “Leporipoxvirus” is defined as the genus of “Squirrel fibroma virus”, but no specification of the difference(s) is given. Chapter 6 presents a method to detect concepts with equivalent definitions, as these represent potentially underspecified concepts, or, worse, different definitions of the same concept.

1.4 Auditing

Auditing means “to look over closely (as for judging quality or condition)” 4. Auditing can be performed for example on organizations, processes, or software, and the aspects to be audited can vary largely. The importance of auditing is twofold; to provide trust and to indicate any possibilities for further improvement. With the increase in number and size of terminological systems, also the attention to auditing these systems increased, as is apparent from various publications, e.g., [16–18]. These publications show that auditing medical terminological systems can be performed in many ways, focusing on various aspects, such as ambiguity, redundancy, or misclassification of concepts represented in these systems.

1.5 Roadmap

The methods described in this thesis are supplementary to existing methods. This thesis first describes a framework of methods for determining characteristics of terminological systems, both independent as well as dependent of their application. The framework contributes to understanding, applicability, mutual comparison and development of terminological systems. Thereafter, this thesis focuses on methods for auditing the knowledge that is formally represented in terminological systems. These methods are intended to maximize the use of automated reasoning and lower the interpretational burden for knowledge modelers.

4 Merriam-Webster Online Thesaurus (http://www.m-w.com)
In summary, the chapters of this thesis along with the specific research questions that they address are:

Chapter 2: How to characterize terminological systems in a structured, reproducible and comparable manner?

Chapter 3: How to assess the extent to which terminological systems cover a specific domain?

Chapter 4: How can logic-based representation be exploited to detect incorrect definitions?

Chapter 5: How can advanced reasoning support tracing and explanation of incorrect definitions?

Chapter 6: How can logic-based representation be exploited to detect duplicate definitions and underspecified concepts?

Bibliography


