Overcoming Barriers to Evaluation of Terminological Systems

Ronald Cornet[a], Nicolette F. de Keizer[a], Daniëlle G.T. Arts[a,b]

[a] *dept. of Medical Informatics, Academic Medical Center, Amsterdam, The Netherlands*
[b] *dept. of Intensive Care, Academic Medical Center, Amsterdam, The Netherlands*

Abstract

Evaluation of terminological systems has been demonstrated to be a complicated task. This is due to the broad range of terminological systems, their application, and the clinical contexts in which they can be applied. We propose an evaluation framework that explicitly distinguishes an application-independent description of terminological systems, methods to determine application-dependent requirements, and methods to assess the applicability. In order to support a systematic application-independent description of terminological systems, we present a categorization of characteristics, including explicit questions. The answers to these questions can be mapped to the requirements for a certain application of a terminological system. This framework aims at reducing the efforts for determining which terminological system is applicable for a certain clinical setting.

Keywords:

Terminology, Evaluation.

Introduction

After decades in which the number and complexity of terminological systems have grown, the topics of standardization and understanding of these systems are getting increasing attention. This has resulted in various publications that address description and evaluation of terminological systems. Among these publications is [1], in which seven barriers to evaluation have been determined:

1. *Application dependence*: quality of terminological systems is defined relative to its intended use;
2. *Empirical vs. independent assessment*: subjective examination to measure characteristics of terminological systems instead of objective measures
3. *Dichotomous vs. continuous measures of characteristics*: presence or absence of features versus the extent of use of features;
4. *Poor definition of characteristics*: different authors use the same name for different characteristics;
5. *Number of characteristics*: evaluation is hardened by the large number of characteristics;
6. *Multiple levels of significance of characteristics*: relevance of characteristics varies and is user-dependent;
7. *Interdependence of characteristics*: characteristics may influence each other.

Among the publications that aim to overcome one or more of these barriers is [2], specifying 12 desiderata that were distilled (mainly) from literature from the nineties of the previous century. The "Standards Specification for Quality Indicators for Controlled Health Vocabularies" [3] is a further step towards structured specification of terminological systems, which distinguishes between: general information, structure of the terminology model, maintenance, and (methods for) evaluation. The Object Management Group (OMG) has used a functionality-oriented approach in the Lexicon Query Service Specification (LQS) [4], which defines "methods for accessing the content of medical terminology systems". Earlier work at our department resulted in a “framework for understanding terminological systems”, which addresses the characteristics that distinguish various types of terminological systems [5-7]. In 2003, the National Committee on Vital and Health Statistics Subcommittee on Standards and Security has made an inventory of about 40 terminological systems to come to national terminology standards for Patient Medical Record Information [8]. This inventory was based on a questionnaire that contained between 40 and 100 (depending on the level of detail considered) questions regarding various aspects of terminological systems (and their developers).

This overview shows that various efforts have been made towards evaluation of terminological systems, overcoming at least some of the barriers.

In this paper we propose a general framework with methods to evaluate terminological systems that pay attention to all seven barriers. We will first focus on the first three barriers; application dependence, assessment and measures of characterization, which all involve the *evaluation methods*. Thereafter, we will elaborate on a *categorization of characteristics*, overcoming the fourth to seventh barrier, as these are all related to the characteristics of terminological systems which have to be taken into account during an evaluation. Next, the categorization of

characteristics is used to present an example of the application-independent description of terminological systems. We demonstrate how the literature discussed earlier relates to the various parts of the proposed framework, and we will address issues that require further research.

Evaluation Methods

Evaluation of the applicability of terminological systems for a specific clinical setting is a complex endeavor. This complexity originates from the lack of agreed upon yardsticks to represent the characteristics of terminological systems, and of a standardized way to elicit requirements that the terminological system should satisfy. Although terminological systems are developed with a certain aim or intended use, it is common that their use is expanding to different usages, that were unforeseen during development. This means however that for each application an assessment should be performed to determine the usefulness of a terminological system. This may lead to duplication of efforts, as some similar questions need to be answered for different applications. We therefore introduce a three-step approach to evaluate terminological systems, which is schematically shown in Figure 1. The individual steps of the method, which are directed toward overcoming the first three barriers mentioned, are discussed in more detail below.

1. Application-independent description

To reduce the duplication of efforts for evaluation as much as possible, we propose to determine application-independent descriptions of terminological systems. Elements of these descriptions are the languages in which concepts are represented, or the formalism underlying concept representation. These application-independent characteristics (which we will elaborate on in the section "categorization of characteristics") can then be used for a first selection of applicable systems, just as the characteristics of hardware devices (the processor speed of a computer, the number of pixels in a digital camera) are used to make a (first) selection. Ideally, the application-independent descriptions should be made available through a publicly available repository.

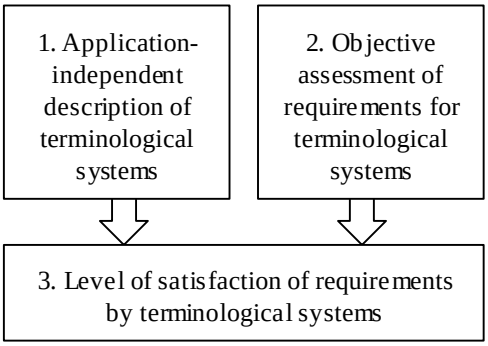

Figure 1: Three steps to evaluate applicability of terminological systems for a specific clinical setting.

2. Methods for objective assessment of requirements

The application of a terminological system is much more complicated than e.g. using a digital camera. This makes it necessary to develop methods to elicit testable requirements in a structural and objective way. These requirements elicitation methods can vary from simple questions (e.g. "in what language(s) should concepts be represented in the terminological system?") to an inventory of the content coverage. Content coverage can for example be determined by collecting a set of registered terms for a sufficiently large period (e.g. one month). This set can then be used as a reference set to be tested against various terminological systems. With this set, it needs to be determined what the required level of coverage should be. We can distinguish term coverage and concept coverage (which may differ if synonymous terms are missing). For both term and concept coverage we can distinguish the "incident" coverage, where frequently used terms are counted more than once, from the "unique" coverage. This approach has been used in [9, 10].

In order to address a broad range of requirements, we propose to distinguish three types of usage of a terminological system: registration, communication and analysis.

Registration is characterized as the task where a terminological system is used for capturing relevant patient information, such as the reason for admission, or an operative procedure. It is important to assess whether a terminological system captures as much as possible all concepts and terms that are applicable. Either these concepts must be present in the system, or it must be possible to compose them using pre-existing concepts, so-called post-coordination.

Communication is regarded as the process of information exchange. Especially when terminology-based data is used in the care process, it is essential that the semantics of the data are well understood by all people involved. In addition, it may be desirable to present information at varying levels of detail, especially when information is exchanged between various specialties.

Analysis of terminology-based data (e.g. for research or management purposes) requires the possibility of aggregation of data. It needs to be determined how concepts need to be aggregated, e.g. whether there are predefined axes (chapters), or whether it should be possible for users to freely combine concept-criteria. The latter case requires concepts to be explicitly characterized by properties and relations to other concepts.

To prioritize, requirements can be weighted on for example a 5-point scale or for example as is done in [8], distinguishing "essential technical criteria", "desired technical criteria", and "desired organizational and process criteria". Evidently, it makes sense to first assess the "essential" criteria, especially those that are relatively easy to assess. This may reduce the number of characteristics that have to be assessed.

3. Satisfaction of requirements

When application-independent descriptions of the characteristics of terminological systems are available, and an assessment of requirements is done, the third and final step is to match the

requirements to the characteristics, in order to determine to what degree systems fulfill the requirements. This can be done by assigning scores to the level of fulfillment for each requirement.

Categorization of Characteristics

The above-mentioned methods provide an outline for the evaluation process. In order to describe terminological systems in a systematic and application-independent way (step 1 of the evaluation approach in Figure 1), it needs to be determined which characteristics are of importance, and how they can be clustered. This clustering is required as the total number of characteristics can become rather large (cf. barrier 5). Whereas there are other relevant issues such as organizational issues, license policies and costs, we will focus in this paper on intrinsic characteristics to describe a terminological system. We propose a clustering that makes the meaning of characteristics more explicit (cf. barrier 4) and that helps understanding their relevance (cf. barrier 6 and 7).

Formalism, content and functions

The first dimension that we propose for clustering the characteristics is the distinction between formalism-, content-, and function-related characteristics.

Formalism-related characteristics are those that relate to the formalisms underlying the representation of terminological knowledge. For example: does the system support polyhierarchy, regardless of the fact whether there are actually any concepts in the system that have multiple parents.

Content-related characteristics describe the actual content of (a specific version/release of) a system. Examples thereof are the number of concepts, the average number of terms per concept, the covered clinical domains.

Function-related characteristics do not actually describe a terminological system (i.e. a structured collection of concepts, terms and codes) but rather a terminology service (a software module that enables navigation, manipulation and/or modification of a terminological system). Ideally, this service is separated from a terminological system, so that various services can be used for the same system. However, we take this issue into account as many contemporary terminological systems provide some default service, and as the use of a terminological system in a computerized environment is commonplace.

Types of terminological systems

The next dimension for clustering the characteristics is along the line of the various types of terminological systems. The essential features of these types, which are described in [5, 6], make it possible to distinguish between a terminology (a list of terms), thesaurus (use of indexing and synonyms), classification (containing generic relationships), vocabulary (containing definitions), nomenclature (containing composition rules), and coding system (containing codes).

Results

The steps described above together form a framework that pays attention to all of the barriers that were presented in [1]. The barriers regarding the characteristics of terminological systems are overcome by a categorization of these characteristics. We present examples of this categorization in Table 1, where characteristics are specified as questions, in order to make them as explicit as possible. Answers to these questions result in an application-independent description of terminological systems.

To specify the requirements for a terminological system, the first step is to determine the foreseen usage of the system. For each type of usage, the users involved need to be determined, along with their requirements for the system. For each of these requirements, methods are required to get more detailed insight, and to make the requirements quantifiable.

Satisfaction of requirements is determined by assigning scores to the level of fulfillment for each requirement. Combining scores and weights will lead to a ranking of terminological systems by appropriateness.

Relation between literature and barriers

We will now relate the literature presented in the introduction to the framework we have described here, although we do not claim that this bibliography provides a complete overview.

The twelve desiderata specified in [2] focus on application-independent characteristics. These mainly involve formalism-related issues (concept orientation, concept permanence, non-semantic identifiers, polyhierarchy, formal definitions, rejection of “Not Elsewhere Classified”, multiple granularities, multiple consistent views, representation of context). “Content” in itself is defined as the most important characteristic of terminological systems. “Recognition of redundancy” relates both to functionality and formalism for detection of equivalent concept definitions. “Graceful evolution” is not within the scope of this paper since it involves a formalized organization for keeping track of changes between versions of terminological systems.

The Quality Indicators from [3] cover the application-independent characteristics as mentioned in [2] but add more detail to this, plus some additional characteristics, such as: clearly stated “purpose and scope” of terminological systems, and functionality for “normalization of content and semantics”. It is furthermore stated that composition of concepts must be possible, i.e. that a terminological system is a nomenclature. A notable addition is the specification of some application-specific requirements, such as: persistence and extent of (primary) use, and the degree of automatic inference intended.

LQS [4], although intended to be a specification for implementation, can be used as a structured reference both to determine implemented and required functionality. As such it can play an important role for the application-independent description and the application-specific requirements. It can relatively easy serve to determine satisfaction of requirements.

The National Committee on Vital and Health Statistics (NCVHS) Questionnaire [8] is the first effort known to the au-

thors that delivers a structured, application-independent description of a variety of terminological systems. If the results become available electronically it will provide a valuable source of information.

Discussion

The framework that we have presented here is a next step towards further improved evaluation of terminological systems, but more research and effort are required to prove its practical utility. In practice, there may be subjective measures, for which different users may give different outcomes. Likewise, it may be hard to agree on assignment of weights to requirements.

The usefulness of the framework will depend in part on the number of people applying it. If application-independent descriptions are shared, and detailed methods are made available, the evaluation will become less complex.

This framework may also help for terminological system developers to determine in what way their system can be improved to serve more or broader needs.

We will discuss the presented results by giving additional examples on the methods and categorization.

Evaluation Methods

Application-independent description

Table 1 gives examples of how an application-independent description of terminological systems can be made. It does refer to the (technical) properties of the systems, and does not describe organizational issues such as the development of a system, the costs or the licensing model. A good overview of relevant questions for an application-independent description can be found in [8].

Methods for objective assessment of requirements

The assessment of requirements for a specific application is a regular software engineering process. As the types of usage of terminological systems can vary largely, we did not go into detail, but we specify three global types of usage: registration, communication, and analysis. The importance of requirements can be expressed by means of a weight. Note that putting weights on requirements is a rather subjective process.

Table 1: a two-axial categorization of questions to obtain application-independent characteristics of terminological systems.

	Terminology	**Thesaurus**	**Classification**	**Vocabulary**	**Nomenclature**	**Coding System**	
Formalism	Are "concepts" and "terms" explicitly distinguished?	Are synonyms allowed, i.e. can multiple terms have the same meanings? How is synonymy represented? Can multiple languages be represented? Are synonyms for fragments allowed? (e.g. cardiac ~ heart)	Can hierarchical relationships between concepts be defined? If yes, Which? Part-of? Is-a? Is multiple classification possible? Can classification be inferred based on a concept's definition?	Is the meaning of concepts represented in free text? Is the meaning of concepts represented formally? If yes: how? e.g. frames, Description Logic (DL) If DL: which DL?	Is composition of concepts possible? How is this represented? Can equivalent definitions be detected automatically? Can compositions change the meaning of a concept, or do they only specify concepts in more detail?	Are codes assigned to concepts? Is there a meaning to these codes (e.g. mnemonic)? Do the codes limit the taxonomic placement of concepts?	**Terminological System**
Content	How many total terms are in the terminology?	In which way(s) are the terms indexed? In what languages are terms described?	Can properties be inherited to subordinate concepts. What is the distribution of the number of parents per concept?	Are all concepts defined/described, or only "core concepts", e.g. diseases, but not anatomy	How many concepts can be combined or further specified?	Are all concepts coded? Are the codes proprietary or cross-mapped to another system?	
Functionality	How can terms be searched? E.g. full match, case insensitive, use of regular expressions?	Can terms be translated for one language to another?	Can all descendants of a concept be retrieved at once? Can properties of a concept be retrieved?	How is a user supported in understanding the meaning of a concept?	How is a user supported in constructing composite concepts?	Can codes be cross-mapped to codes in another coding system?	**Terminology Service**

Satisfaction of requirements

Relating the requirements to the descriptions of the terminological systems is not straightforward. The requirements have mainly been related to the use of the system, and the functionality and content that are required for making the intended use possible. However, the descriptions can be very detailed and technical. This makes it difficult to match the requirements to the descriptions. Moreover, for evaluation of items such as content coverage, the terminological system itself is needed to perform the various counts; hence just a description of the system is not sufficient.

Categorization of Characteristics

Combining the two dimensions “Formalism, content and functionality” and “Types of terminological systems” results in a matrix of characteristics as presented in Table 1. In this matrix we have specified questions as examples of the characteristics of terminological systems. These questions are examples of how to specify the application-independent description of terminological systems, the first step in the evaluation process. As the complete set of questions is currently under development, and due to length-restrictions, we do not present it here.

The content-related questions make clear whether a system actually contains specific knowledge (e.g. concepts, terms, synonyms), whereas the formalism-related questions provide insight whether a system can potentially capture and represent certain types of knowledge. This distinction helps to explain the strengths and weaknesses of systems, and the possibility to overcome the weaknesses. Generally, shortcomings in the content as it is can relatively easily be solved, whereas shortcomings in the formalism are more severe. Likewise, if a terminology server lacks functionality, this can only be implemented if the formalism underlying the terminological system provides support for such functionality. E.g. to provide word normalization, the formalism must allow to represent normal forms and inflections of terms.

Conclusion

Although work needs to be done to further specify the framework we have described in this paper, it provides insight in the way to characterize terminological systems, and to assess the requirements. In this way the framework gives insight in the completeness of assessed criteria. By making characteristics explicit and as much as possible objectively measurable, the barriers to evaluation of terminological systems can be overcome. The first three barriers mentioned by Hales [1] have been overcome by using the three-step approach as presented in Figure 1. By using a structured categorization of characteristics (table 1) in the first step of this approach we paid attention to the remaining barriers. As making an application-independent description is a one-time effort for each (version of a) terminological system, and requirements specification is a one-time effort for each application, this framework may lead to reduction of the effort to be put in evaluation, if researchers share the results.

Acknowledgments

This work is supported by the Netherlands' Organization for Scientific Research (NWO).

References

[1] Hales JW, Schoeffler KM. Barriers to Evaluation of Clinical Vocabularies. In: MedInfo1998; 1998; Vancouver; 1998. p. 680-4.

[2] Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. Methods of Information in Medicine 1998;37(4-5):394-403.

[3] ISO/TC215 WG 3. Standard Specification for Quality Indicators for Controlled Health Vocabularies; 2000 July. Report No.: TS17117.

[4] OMG. Lexicon Query Service Specification: Object Management Group; 2000 july. Report No.: 00-06-31.pdf.

[5] de Keizer NF, Abu-Hanna A, Zwetsloot-Schonk JH. Understanding terminological systems. I: Terminology and typology. Methods of Information in Medicine 2000;39(1):16-21.

[6] de Keizer NF, Abu-Hanna A. Understanding terminological systems. II: Experience with conceptual and formal representation of structure. Methods of Information in Medicine 2000;39(1):22-9.

[7] Cornet R. Towards Structured Requirements for Terminological Systems and Servers. In: et al., editor. MedInfo 2001; 2001; London; 2001. p. 295.

[8] NCVHS SSS. Summary and Analysis of Terminology Questionnaires Submitted by Developers of Candidate Terminologies for PMRI Standards. draft; 2003 April 17.

[9] Chute CG, Cohn SP, Campbell KE, Oliver DE, Campbell JR, For The Computer-Based Patient Record Institute's Work Group on Codes & Structures. The content coverage of clinical classifications. Journal of the American Medical Informatics Association 1996;3(3):224-33.

[10]Humphreys BL, McCray AT, Cheh ML. Evaluating the coverage of controlled health data terminologies: report on the results of the NLM/AHCPR large scale vocabulary test. J Am Med Inform Assoc 1997;4(6):484-500.

Address for correspondence

Ronald Cornet
Dept. of Medical Informatics, J2-256
Academic Medical Center, University of Amsterdam
P.O. Box 22700
1100 DE Amsterdam
The Netherlands

r.cornet@amc.uva.nl