# Reconciliation of Ontology and Terminology to cope with Linguistics

**Robert H. Baud[a], Werner Ceusters[b], Patrick Ruch[a], Anne-Marie Rassinoux[a], Christian Lovis[a], Antoine Geissbühler[a]**

[a] *University Hospitals of Geneva, Service d'Informatique Médicale, Switzerland*
[b] *Center of Excellence in Bioinformatics & Life Sciences, SUNY at Buffalo, USA*

**Abstract**

***Objectives:*** *To discuss the relationships between ontologies, terminologies and language in the context of Natural Language Processing (NLP) applications in order to show the negative consequences of confusing them.*

***Methods:*** *The viewpoints of the terminologist and (computational) linguist are developed separately, and then compared, leading to the presentation of reconciliation among these points of view, with consideration of the role of the ontologist.*

**Results:** *In order to encourage appropriate usage of terminologies, guidelines are presented advocating the simultaneous publication of pragmatic vocabularies supported by terminological material based on adequate ontological analysis.*

***Conclusions:*** *Ontologies, terminologies and natural languages each have their own purpose. Ontologies support machine understanding, natural languages support human communication, and terminologies should form the bridge between them. Therefore, future terminology standards should be based on sound ontology and do justice to* the diversities in *natural languages. Moreover, they should support local vocabularies, in order to be easily adaptable to local needs and practices.*

***Keywords:***

Terminology, Ontology, Natural Language Processing

## Introduction

Recently, Cimino expressed 12 new desiderata for controlled vocabularies in the twenty-first century [1]. The second desideratum says that for each object of a domain, there should be one and only one representative term, which is unique and non ambiguous. Such a term is often call the *preferred term* and the collection of all such terms is a *controlled terminology* (formerly named a *controlled vocabulary*). A revision paper adds 6 new desiderata [2], in response to the recent trend in favor of more formal ontological foundations for terminologies in which terms refer to entities in reality, rather than to concepts in the mind of domain experts [3]. Although controlled terminologies based on preferred terms may contribute to semantic interoperability amongst applications, they are also the source of a modern myth according to which the world could do without the freedom of expressivity that is possible in natural language [4]. Clearly, natural languages are not controllable. Human beings use natural language not only for professional reasons, but also for cultural and social reasons. The patient record, for example, whether paper- or computer based, is usually dictated or written directly in the care provider's natural language so that direct local expressiveness of the language and not conformity to academic desiderata or controlled terminologies is given priority.

This is not to say that Cimino is wrong, but that his paper is written from the perspective of a terminologist: the primary goal is to make data annotated by means of terminologies computer understandable and to enable interoperability between different systems. The desiderata he proposes thus encode these constraints. Despite his expertise in terminology-based lexicon development such as the MED [5], the clothes that Cimino wears in his papers are not those of the linguist. Nevertheless, he also asserts that *'synonymy is a type of redundancy which is desirable'* (desideratum 12). This statement, so we argue, supports the conclusion of this paper.

If a terminology is intended to be disseminated widely and to facilitate communication between human agents and computer systems, then it should take the modus operandi of both these players into account. Thus in order to accommodate the needs imposed by computer systems, terminologies should be based on adequate ontologies. To service human agents, terminologies should be linked to lexicons containing multiple synonyms, eponyms, acronyms, and local jargon. Perhaps, they should even come equipped with facilities to deal with common spelling errors.

This paper deals primarily with the design issues of terminologies related to human agents, and more specifically with the following paradox: on one side, the terminologists recommend unique preferred terms for referring to the entities in a domain and on the other side the linguists and the majority of users are in favor for a full diversification of natural languages, including all kinds of synonyms. To overcome this paradox this paper will first consider the two points of view (that of the terminologist and that of the linguist/end-user) and their supporting arguments. Acknowledging that each expert is guided by specific constraints and is accordingly acting coherently, the constraints are made explicit. In a second step,

conflicting constraints are examined and an enlarged perspective able to accommodate both points of view is developed. Finally, guidelines based on this enlarged perspective are formulated.

## State of the art

Numerous terminologies have been found to suffer from major inconsistencies, examples being SNOMED CT [6], the Terminologia Anatomica (TA) [7] and the NCI Thesaurus [8]. Rector, who raised the question "Clinical terminology: Why is it so hard?" [9], argues that '*clinical terminology concerns the meaning, expression and use of concepts in statements in the medical record'* and assumes the reason for the inconsistencies in terminologies to be the result of a failure in '*separating language and concept representation'* (#4 of 10 difficulties). He describes it as the '*confusion of concepts and words used to express those concepts*.'

However, others argue that the inconsistencies are introduced by the lack of a sound ontological basis for these terminologies, misled as their authors are by the concept orientation advocated by Rector [3] thereby confusing ontology with epistemology [10]. Therefore, in order to avoid such inconsistencies and to coordinate the efforts of several groups by means of commitment to an agreed upon set of principles for best ontology practices, the Open Biomedical Ontology (OBO) Foundry - a new paradigm for biomedical ontology development - has been created [11]. Currently, a set of 10 principles is published and all OBO members are committed to follow them. This initiative is clearly a collaborative experiment in the quest for discovery of best practices in ontology and terminology development. It rests on the principle that high quality representations should be built out of representational units that refer exclusively to three sorts of entities that exist in reality - universals (e.g. person, disease), defined classes (e.g. employee of a hospital, patient under treatment) and particulars (this tumor, the World Health Organization) – and that are connected by means of relationships that mirror ontological relationships [12]. In this context, a terminology is defined as a representational artifact consisting of representational units which are the general terms of some natural language used to refer to entities in some domain. As a result of this effort, quality ontologies in the biomedical domain start to exist. The Gene Ontology (GO) [13] is in constant development, following the evolution of genetic sciences and the provisions of the OBO Foundry [11]. On the side of clinical practice, the Foundational Model of Anatomy (FMA) [14], supporting the TA [15], is an example of a high quality ontology following the OBO principles.

This is in contrast with mainstream work on terminology during the last 20 years which was dominated by a linguistic and normative perspective, primarily driven by the English speaking community. This has had (at least) two inconveniences. First, language itself is not able to make predictions about what exists: although the term "king" might refer to the universal *king*, and "France" to the particular *France*, language rules allow us to create terms such as "the king of France", even though no such entity currently exists. Second, emphasis on English language terminologies hampered creating translations and this for many reasons: 1) human translation is resource dependant and time consuming; 2) the terminology is in principle subject to major revision; 3) questions arise about the role and quality of terms; 4) sets of useful synonyms are not often available and difficult to collect.

## The Terminology line

The most widely prevailing view on terminology holds that '*terms are the linguistic representation of the concepts in a particular subject field and are characterized by special reference*' as opposed to words that '*function in general reference over a variety of codes*' [16]. According to Lauriston, a term is '*the intersection between a conceptual realm (a defined semantic content) and a linguistic realm*' [17].

In order to ensure that terms are used with the correct meaning, the terminologist may provide definitions which allow the terms to be organized in a hierarchy. However, because this task is difficult, long and expensive, he may not be able to provide explicit definitions for all entities. Therefore, he uses the entity's name as subsidiary definition. But at the same time he is constrained to limit the names to short terms (say up to 5 words) for pragmatic reasons. Using terms alone as implicit definitions is a design error, leading to severe problems[1].

Most terminologies have also adopted usage of what is called *preferred terms*, although SNOMED CT, GO and FMA have added non significant numeric identifiers. Preferred terms are subject to desideratum 2 (one and only one meaning) and terminologists try to respect this constraint.

However, in our view, the primary goal of the terminologist is not to relate the terms in a domain of discourse to 'concepts', but to organize them in such a way that it is clear which terms refer to what entities in reality, and which do not. Furthermore, for the purpose of communication he has to identify the relevant entities that are not yet properly named. This means that the terminologist must play partly the role of the ontologist, or rely on services offered by ontologists. However, he must be aware that the ontology underlying any terminology is universal and language independent. There is a significant risk of mixing the ontologist's role with the terminologist's role, which is mainly language dependent. As an example, it is not because medicine identified the two distinct diseases formerly named "diabetes type 1" and "diabetes type 2", that there exists a universal referred to by the term "diabetes".

The terminologist should also be aware that in contrast to definitions in ontologies that describe the necessary and sufficient conditions for an entity to be what it is, terminological definitions should focus on the conditions under which a term may appropriately be used.

This in turn puts a different perspective on the notion of preferred terms the effectiveness of which has been questioned:

- preferred terms have been selected according to the needs of the experts, not the casual users,
- they may not be adequate in specific contexts,
- they contradict local habits and usages,

[1] In TA the code a05.6.02.010 is for *hidden part of duodenum*. Not every physicians knows about this object and the term is non existent in most atlas of anatomy!

- they are not easily accessed by non expert users,
- they vary from one language to another,
- human beings fancy to be different and are reluctant to standardize their language.

In his paper [4], Ceusters gives a voice to the users. He demonstrates that less than 50% of them are using the defined preferred terms in practice, this trend being augmented by usage of speech recognition tools. He says that '*clinicians do not face major problems in understanding terms derived from clinical narratives generated by peers*'. He also suggests that '*preferred terms are merely an academic artifact rather than a reality*'. Moreover, Ceusters shows that the mean number of variants for any term is superior to 5 in practice, in concordance with another study showing even higher figures [18]. In order to do justice to the users, it should be recognized that they '*want (and get) back the freedom of expression with all delicate yet important nuances that are required for individual patient care*' [4].

In theory, the need for preferred terms is artificial: with only numeric identifiers and explicit definitions of entities, the design and publication of a source terminology is feasible. From a pragmatic point of view, the usage of unique text identifiers (the *knowledge names* in Galen) is possibly a good choice, but their usage should be limited to the experts and discouraged to the end-users!

## The Language line

The primary goal of the computational linguist is to build applications that can analyze and understand medical texts. A first requirement for an application of this type is that it is able to recognize which terms or phrases in a text, including any linguistic variants or forms, refer to domain entities [19]. There is a practically unlimited number of variants for any given term, as shown in [18], thus the construction of an explicit list is not feasible. Therefore, the linguist tries to design intelligent algorithms that can take a short list as input, and deduce other variants based on it.

Variants come in many flavors and for various reasons, as witnessed by the TA [20]: 1) different sources for naming an entity (for a02.1.05.053 *pterigoid canal*, there is also *recurrent canal*); 2) use of Latin terms (for a02.1.06.022, there are two terms: the Latin *tegmen tympani* or simply *roof of tympanum*); 3) usage of eponyms (*Eustachian tube* for a15.3.02.073 in place of *pharyngotympanic tube*); 4) representation by different part of speech (for a15.2.07.024 *eyelids*, there is also an adjective *palpebral* and a prefix root *blepharo*); 5) use of old or layman terms (for a02.4.01.001 *scapula* there is also *shoulder blade*); 6) difficulties with orthography and/or usage of keyboard (for a11.3.00.001 *tyroid gland* in place of *thyroid gland*); 7) usage of local expressions; 8) differences between professionals, typically surgeons and anatomists; 9) order of word segments (for a03.1.02.006 both *frontosphenoid suture* and *sphenofrontal suture* are correct!); 10) morphological changes and usage of plural (terminology should avoid plural terms, but not the medical texts: a02.2.05.001 *sacral vertebrae* should preferably be *sacral vertebra*); etc.

In free text, rather than terminologies, for instance in the context of data registration in electronic health records, there are even more degrees of freedom such as the use of local idioms, non-academic sentences or orthographic errors. If NLP applications aim for total understanding of the content, these lexical aspects must be taken into account in a way that fortunately has become almost feasible today.

A second requirement for NLP applications is to identify to what domain entity a specific term refers. This often requires the resolution of ambiguities. One type of ambiguity is that the same term may refer to two or more distinct domain entities. Another type is brought about by the use of more generic terms when previously a more specific term has been used. Furthermore, algorithms that generate lexical variants to obtain better recall typically sacrifice on precision, thereby introducing more ambiguity. This calls for taking context into account. If the author of the text is willing to be non ambiguous, he could apply ambiguity free paradigms such as Referent Tracking [21], but this requires considerable future developments and fine tuning of computer applications. This is the cost for an improved situation.

## Conjunction of the needs

In [3], Smith introduced a formal definition of a terminology based on philosophical realism. Taking into account the follow-up work reported on in [12], this definition may be rewritten as:

A **terminology** T is a triple <N, L, v>

where:

- N is a set of triples <p, $S_p$, d>, called *nodes*, with p a unique *label*, $S_p$ a set of *synonyms* and d a *definition* of a node,
- L is a set of ordered pairs <r, $L_r$>, called *links*, consisting of a relation designation r (*is_a*, *part_of*, etc), together with a set $L_r$ of ordered pairs <s, s'> of those terms for which srs' represents a consensus assertion of biomedical science about corresponding universals and defined classes at the time the given terminology is prepared,
- v is a *version* number, which encodes the time.

But in order to harmonize the terminology line and the language line, an extension of the definition of synonym is necessary[2]:

This approach involves a shift from the preferred term paradigm to the *bag of terms* paradigm, a bag of terms being by definition an unordered set of terms, each being qualified by different properties governing its usage. No term is above the others and the terminology does not favor any language. Nevertheless, the parameter k below may include a tag indicating whether a string is to be considered appropriate for use in text generation applications. This will avoid that such an application would introduce terms with typographic errors or layman terms in a professional document. But from a pragmatic point of view, any implementation is free to select any term from

[2] The above definition of terminology is borrowed from Smith in [3]; the definition of synonyms' set below is original, in complement.

the set according to local preferences, for the sake of designing user friendly interfaces.

$S_p$ is a set of synonyms, where each **synonym** is a five-tuple <s, k, p, l, t>

where:

- s is a string of characters in some regimented language l,
- k specifies the sort of string (preferred term, acronym, eponym, local idiom, old term, common orthographic error, etc), governing its usage,
- p is the *part of speech* argument of the term t (noun, adjective, verb, etc, augmented by information on gender, number, and so forth)
- l is the language,
- t is the time (or time period) when the string is (or was) considered appropriate to refer to the corresponding entity.

This is in line with, Rector's *'principle of separability between clinical linguistics and clinical pragmatics'* [9]: building appropriate bags of terms is a matter of clinical pragmatics, preferred terms a matter of clinical linguistics.
The bag of term principle is not new and has been advocated by UMLS (Unified Medical Language System), where each CUI (Concept Unique Identifier) may be considered as acting as a recipient or identifier for several variant terms. But UMLS is not a terminology in itself; it is rather a collection of terminologies. The paradigm shift presented here remains a valid recommendation for any individual terminology. Wordnet [22] presents a similar approach with the synsets.

## Guidelines

Based on the above arguments, we propose the following guidelines for future terminologies (emphasizing the language point of view):

- let the label p for a node N in terminology T be a meaningless unique identifier,
- let definition d for node N be such that it reflects the necessary and sufficient conditions for a particular to be an instance of the universal, or a member of the defined class, referred to by N,
- prepare an open bag of terms $S_p$ able to contain any number of strings, preserving the principle of an unordered set (this is the main complement to the initial Smith proposal in [3]),
- qualify any string in $S_p$ for its usage, in particular specify any preferred terms depending on the contexts (be aware that terms do not strictly need to be unique),
- fill in the bag of terms for any language, corresponding to the locations where the terminology is to be disseminated,
- add for any string its parts of speech,
- develop the links of the terminology (this includes the taxonomy) in total independence of the bag of terms, in a language independent way, and on the basis of sound ontological principles,
- consider the maintenance aspect of the terminology, including free addition of new terms and their attributes at any time,
- consider quality assessment tests for the terminology, including validation of the completeness of the set of synonyms.

## Discussion

Although ontology is by definition language independent, terminology is not. We argue that terminologies should be made available for each language by using an underlying ontology as a reference, rather than by relying on a structured organization of preferred terms in a specific language.
We believe however that the ontology, terminology and language points of view can be reconciliated through a simple move: the renunciation of preferred terms and the replacing of them with a bag of representative terms, together with qualifiers defining their usage. In addition, the availability of explicit definitions is recommended. This should lead to a win-win situation: the terminology line is totally preserved and the language line is adequately presented. The multilingual aspect underlying any universal terminology is now explicitly mentioned. But of course, there is a cost: the guidelines proposed here will certainly be labor intensive if adhered to. However, we believe that the approach will help save efforts in the long run, which will be necessary anyway before widespread usage of the best terminologies will be possible. Furthermore, the open bag of terms may remain partially empty for a long time, without direct inconvenience for most of the users. The main point is that the bag remains open and is progressively filled in. Natural language generation of compound terms might be helpful here: if the ontology is expressed in some form of logic, then the ontology itself can be used to generate parts of the terminology automatically, and this in many languages. This approach should progressively become the rule for terminologies, because numerous entities are composed from more atomic entities referred to by means of single words.
In order to show how this works, consider the SNOMED CT code SN 285344007, whose preferred term in English is *viral gastritis*. A graph representation of this entity could be the following (square brackets are for objects, round brackets are for relations):

[InflammationProcess]
- (hasLocation) - [Stomach]
- (hasAgent)-[Virus]

The generator program may find in its English lexicon the following words: for [Stomach]: { gastr-, stomach, stomachal}, for [InflammationProcess]: {-ite, inflammation, inflammatory}, for [Virus]: {virus, viral}. In any other language the lexicon may be more or less extensive and the generator will have to cope with this. The generator may use any combination of words, one of them coming from each subset. However, some rules specific to the language will give more weight to some combinations and others will be excluded. Each word will be selected according to its own specificity, computed from its frequency in a large corpus of representative medical texts. At the end of the process, the generator is left with 2 combinations (the first being considered as the best

if the strategy is to prefer the short terms): *viral gastritis, inflammation of stomach caused by a virus*. Other terms are theoretically possible, they are correct but they are discarded because they are unusual: *viral stomachal inflammation, gastritis by virus*.
A successful generation experiment has already been conducted for surgical procedures [23 ,24]. The same approach is considered by WHO for preparing ICD-11.

## Conclusion

Terminologies should not be developed by reference to a system of preferred terms, rather they should be developed in such a way that their individual nodes and relations amongst these nodes are modeled on an underlying formal ontology, where the linguistic content of these nodes will be filled in based on a system of terms and synonyms (from many different languages) that is associated with each node based on the intended ontological interpretation of that node. The idea is that this will give terminologies the rigor of scientific theories, while also making them understandable in natural language.

## Acknowlegments

Thanks to James Cimino, Barry Smith and Olivier Bodenreider for their kind comments to this paper. Thanks to Andrew Spear for proof reading of the manuscript.

**Address for correspondence**

Robert H Baud, PhD
Service d'Informatique Médicale
Hôpitaux Universitaires de Genève
rue Micheli-du-Crest, 22
CH-1211 Genève Suisse

E-mail: Robert.Baud@sim.hcuge.ch

## References

[1] Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. Methods Inf Med 1998 Nov;37(4-5):394-403.
[2] Cimino JJ. In Defense of the desiderata. J Biomed Inform 2006;39(3): 299-306.
[3] Smith B. From concepts to Clinical Reality: An Essay on the Benchmarking of Biomedical Terminologies. J Biomed Inform 2006;39(3):288-298.
[4] Moerkerke C, Ceusters W. The myth of preferred terms in medical sublanguage and its impact on natural language understanding applications: an empirical study. In De Moor G, De Clercq E (eds.) Proc 18th MIC Conf, 2000: pp55-62.
[5] Cimino JJ. From Data to Knowledge through Concept-oriented Terminologies: Experience with the Medical Entities Dictionary. J Am Med Inform Assoc 2000;7(3): pp288-97.
[6] Ceusters W, Smith B, Kumar A, Dhaen C. Ontology-based error detection in SNOMED-CT. Medinfo. 2004;11(Pt 1):482-6.
[7] Rosse, C. Terminologia Anatomica; Considered from the Perspective of Next-Generation Knowledge Sources. *Clinical Anatomy* 14:pp 120-133.
[8] Ceusters W, Smith B. A Terminological and Ontological Analysis of the NCI Thesaurus. Methods Inf Med 2005; 44: pp498-507.
[9] Rector AL. Clinical Terminology : Why Is it so Hard ? Methods Inf Med 1999; 38: pp147-57.
[10]Bodenreider O, Smith B, Burgun A. The Ontology - Epistemology Divide: A case study in Medical Terminology. Procs 3rd conf on Formal Ontology in Information Systems (FOIS 2004): IOS Press; 2004. p185-195.
[11]The OBO Foundry at: http://obofoundry.org/
[12]Smith B, Kusnierczyk W, Schober D, Ceusters W. Towards a Reference Terminology for Ontology Research and Development in the Biomedical Domain. Proceedings of KR-MED 2006, Biomedical Ontology in Action, November 8, 2006, Baltimore MD, USA
[13]See: http://www.geneontology.org/ or the NLM browser GenNav: http://mor.nlm.nih.gov/perl/gennav.pl
[14]Rosse C et als. Foundational Model of Anatomy. See: http://sig.biostr.washington.edu/projects/fm/AboutFM.html
[15]Federative Committee on Anatomical Terminology. Terminologia Anatomica: International Anatomical Terminology. Thieme Ed. 1998.
[16]Sager J.C. A Practical Course in Terminology Processing. John Benjamins Publishing Company, 1990.
[17]Lauriston, A. Automatic recognition of complex terms: Problems and the TERMINO solution. In Terminology 1994, 1:;147-170.
[18]Baud RH, Ruch P, Gaudinat A, Fabry P, Lovis C, Geissbuhler A. Coping with the variability of medical terms. Medinfo. 2004;322-6.
[19]Bodenreider O. Lexical, terminological and ontological resources for biological text mining. Chap 3 in Ananiadou S et al: Text mining for biology and biomedicine; Artech House; 2006. pp43-66.
[20]Federative Committee on Anatomical Terminology. Terminologia Anatomica: international anatomical terminology. Stuttgart; New York: Thieme; 1998
[21]Ceusters W, Smith B. Strategies for Referent Tracking in Electronic Health Records. J Biomed Inform. 2006 Jun;39(3):362-78.
[22]Fellbaum C. An Electronic Lexical Database. MIT Press, 1998.
[23]Trombert-Paviot B, Rodrigues JM, Rogers JE, Baud RH et al. GALEN : a third generation terminology tool to support a multipurpose national coding system for surgical procedures. Int J Med Inform 58-59 (2000); pp71-85.
[24]Rodriges JM, Rector A, Zanstra P, Baud R & al. An Ontology driven collaborative development for biomedical terminologies: from the French CCAM to the Australian ICHI coding system. MIE 2006 Procs. A Hasman et al. (Eds), IOS Press, 2006, pp 863-8.