Ceusters W, Smith B. A Terminological and Ontological Analysis of the NCI Thesaurus. Methods of Information in Medicine 2005; 44: 498-507

# A Terminological and Ontological Analysis of the NCI Thesaurus.

*Werner CEUSTERS*

*European Centre for Ontological Research, Saarland University, Saarbrücken, Germany*

*Barry SMITH*

*Department of Philosophy, University at Buffalo, New York, USA*

*Institute for Formal Ontology and Medical Information Science, Saarland University, Saarbrücken, Germany*

*Louis GOLDBERG*

*School of Dental Medicine, University at Buffalo, New York, USA*

*Institute for Formal Ontology and Medical Information Science, Saarland, University, Saarbrücken, Germany*

**Corresponding author:**

Werner Ceusters
European Centre for Ontological Research
Universität des Saarlandes
Postfach 151150
D-66041 Saarbrücken
Germany
werner.ceusters@ecor.uni-saarland.de
Tel.: +49 (0)681-302-64770
Fax: +49 (0)681-302-64772

## **Summary**:

Objective: The National Cancer Institute Thesausus is described by its authors as "*a biomedical vocabulary that provides consistent, unambiguous codes and definitions for concepts used in cancer research*" and which "*exhibits ontology-like properties in its construction and use*". We performed a qualitative analysis of the Thesaurus in order to assess its conformity with principles of good practice in terminology and ontology design.

Materials and methods: We used both the on-line browsable version of the Thesaurus and its OWL-representation (version 04.08b, released on August 2, 2004), measuring each in light of the requirements put forward in relevant ISO terminology standards and in light of ontological principles advanced in the recent literature.

Results: We found many mistakes and inconsistencies with respect to the term-formation principles used, the underlying knowledge representation system, and missing or inappropriately assigned verbal and formal definitions.

Conclusion: Version 04.08b of the NCI Thesaurus suffers from the same broad range of problems that have been observed in other biomedical terminologies. For its further development, we recommend the use of a more principled approach that allows the Thesaurus to be tested not just for internal consistency but also for its degree of correspondence to that part of reality which it is designed to represent.

**Keywords:**

Ontology, Medical Terminology, Standardisation, Quality Control

# 1 Introduction

The automatic integration of heterogeneous information is one of the most challenging goals facing biomedical informatics today [1]. Controlled vocabularies have played an important role in realizing this goal by making it possible to draw on biomedical information deriving from divergent sources secure in the knowledge that the same terms will also represent the same entities even when used in different contexts.

Unfortunately, as has been shown in a series of recent studies, almost all existing controlled vocabularies in biomedicine have a number of serious defects when assessed in light of their conformity to both terminological and ontological principles [2, 3, 4, 5, 6, 7, 8]. The consequence is that much of the information formulated using these vocabularies remains hidden to both human interpreters and software tools. The result is that vital opportunities for enabling access to the information in such systems have been wasted, in ways which manifest themselves in difficulties encountered both by humans and by information systems when using the underlying resources in biomedical research. Such defects are destined to raise increasingly serious obstacles to the automatic integration of biomedical information in the future, and thus they present an urgent challenge to research.

In this paper, we present the results of our assessment of the conformity of the NCI Thesaurus (NCIT) to widely accepted principles in the domain of terminology development as well as to well-established principles for ontology building that have grown out of more than two millennia of philosophical research on classification and categorization.

# 2 Materials and Methods

## 2.1 The NCI Thesaurus

The NCIT is a cancer research nomenclature with features resembling those of an ontology in the sense in which this term is used in the current bioinformatics literature: thus it is a *controlled vocabulary* organized as a structured list of terms and definitions. It was created by the National Cancer Institute's Center for Bioinformatics and Office of Cancer Communications for use not only by the Institute's own researchers but also by the cancer research community as a whole. Its main goals are:

1) *to provide a science-based terminology for cancer that is up-to-date, comprehensive, and reflective of the best current understanding;*
2) *to make use of current terminology "best practices" to relate relevant concepts to one another in a formal structure, so that computers as well as humans can use the Thesaurus for a variety of purposes, including the support of automatic reasoning;*
3) *to speed the introduction of new concepts and new relationships in response to the emerging needs of basic researchers, clinical trials, information services and other users* [9].

The NCIT serves several functions, including annotation of the data in the NCI's repositories and search and retrieval operations applied to these repositories. It is also linked to other information resources, including both internal NCI systems such as caCore, caBIO and MGED and also external systems such as the Gene Ontology and SNOMED-CT. It is part of the Open Biomedical Ontologies library [10] and is also available under Open Source License on the NCI download area [11]. This makes it an important candidate for the delivery of vocabulary services in cancer-related biomedical informatics applications in the future.

NCIT is a thesaurus, and one can thus expect it to be of use to researchers engaged in biomedical database annotations. At the same time its ontological underpinnings are designed to open up the possibility of more complex uses in automatic indexing and bibliographic retrieval and in linking together heterogeneous resources created by institutions external to the NCI. It is this last potential application that is receiving most attention in the biomedical research community.

For this study we used version 04.08b of the NCIT, released on August 2, 2004 and made publicly available through the NCI website [12]. (Some of the errors identified below have been since corrected.)

## 2.2 Nature of the analysis

We have measured the NCIT's qualities along three lines: 1) conformity with relevant terminological standards put forward by ISO; 2) ontological principles; and 3) appropriateness of OWL as a knowledge exchange format.

### 2.2.1 Terminological standards:

Since the NCIT was developed using a concept-centered design, we selected as the reference for good terminological principles the standards produced by Technical Committees 37 and 46 of the International Standards Organization (ISO TC37; ISO TC46). The relevant standards are listed in Table 1.

| Standard No | Standard Title |
|---|---|
| ISO 704:2000 | Terminology work – Principles and methods |
| ISO 860:1996 | Terminology work – Harmonization of concepts and terms |
| ISO 1087-1:2000 | Terminology work – Vocabulary – Part 1: Theory and application |
| ISO 15188:2001 | Project management guidelines for terminology standardization |
| ISO 1087-2:2000 | Terminology work – Vocabulary – Part 2: Computer applications |
| ISO 12620:1999 | Computer applications in terminology – Data categories |
| ISO 16642:2003 | Computer applications in terminology – Terminological markup framework |
| ISO 2788:1986 | Documentation – Guidelines for the establishment and development of monolingual thesauri |

*Table 1: Relevant ISO standards for the evaluation of the NCI Thesaurus*

Not everything that is contained in these standards is, as we shall see, fully appropriate to the purposes of biomedical information integration. Of crucial importance in all of them, however, is the notion of definition, which in ISO 1087-1:2000 is defined as: "*a representation of a concept by a descriptive statement which serves to differentiate it from related concepts*". Only basic and familiar concepts (also called 'primitive concepts') do not need to be defined. ISO lists further a number of requirements that definitions should meet. Thus definitions must describe the *concept* – not the *words* that make up its designation. They must also describe exactly *one* concept. ISO 1087-1:2000 stipulates specifically that definitions for a concept shall not include other definitions as proper parts, and that any characteristic that requires an explanation should either be defined separately as a concept in its own right, or elucidated in a note. Another ISO requirement states that definitions should be as brief as possible and as complex as necessary. Complex definitions can contain several dependent clauses, but carefully written definitions should contain only sufficient information to ensure that the concept in question is uniquely specified. Any additional descriptive information deemed necessary should, again, be included in a note.

ISO 704:2000 lists some requirements that newly constructed *terms* should adhere to. They should be:

1. linguistically correct (i.e. they should conform to the rules of the language in question),
2. precise and motivated (i.e. they should reflect as far as possible the characteristics which are given in the definition),
3. concise.

If possible, newly introduced terms should also permit the formation of derivatives.

Every term included in a standardized terminology should be monosemic. The latter requirement is expressly laid down for those coinages designated as "preferred terms". Such terms, according to ISO, should also have the highest rating for acceptability in the relevant user community (though as a matter of fact they are often forced upon such a community with the purpose of stabilizing its terminology).

Another set of important terminological principles concern the proper use of “synonyms”. The strict definition of synonymy proposed by ISO 1087-1:2000 is: *relation between or among terms in a given language representing the same concept*, with a note to the effect that “*Terms which are interchangeable in all contexts are called synonyms; if they are interchangeable only in some contexts, they are called quasi-synonyms*”.

### 2.2.2 Ontological principles

Counterparts of ISO standards dealing with ontology development do not as yet exist. In performing the ontological part of our analysis we drew instead on the fundamental principles underlying ontology development employed in systems such as Basic Formal Ontology [4] or DOLCE [13]. The latter, which draw in their turn on a long tradition of ontological research in philosophy, distinguish between *universals* (also called kinds, species, or types) and *particulars* (individuals, instances, or tokens). Examples of universals are *cancer* as studied in medical school and each specific sort of cancer (*prostate cancer*, etc.). An example of a particular would be: this particular cancer, present in this particular patient, here before you now; or: the prostate cancer in that particular patient on the other side of the room.

Cross-cutting the distinction between universals and particulars is that between *continuants* and *occurrents*. These two sorts of entities are marked by the fact that they relate in different ways to time. Continuants endure through time, which is to say that they are wholly present at each moment of their existence. Examples of continuants are *organs*, *solid tumors*, *cutters*, *chromosomes*, and so forth. Occurrents, on the other hand, are never fully present at any given moment in time; rather they unfold themselves in their successive phases. Examples are processes such as *tumor invasion* or events such as a *surgery session*.

It is important to note that parthood relations never cross the mentioned categorial boundaries; that is, parts of continuants are always continuants and parts of occurrents are always occurrents. As an example: the tumor is not a part of the tumor invasion, nor is the surgeon a part of the surgery session. The parts of the process *removing a tumor* include: making a skin incision, draining blood, identifying the diseased tissues, and so forth. The physicians or surgeons who perform these actions are, rather, *participants* (in this case *agents*) in the corresponding processes.

A further distinction is that between *independent* entities, such as *persons* and *protein molecules*, that have the ability to exist without the ontological support of other entities, and *dependent* entities, such as *colors* and *shapes,* that require the existence of other entities – their bearers – in order to exist. Here, too, parthood relations never cross the boundaries between these two types of entities.

It is our experience that ontologies that do not respect these fundamental distinctions will contain errors of a sort which are not detected by the standard tools used for error checking in the knowledge representation field. This is because such tools focus primarily on the issue of syntactic consistency [14], rather than an ontological coherence. Typical examples of such mistakes are classes that comprehend both processes and material objects, or, even worse, classes that are defined in such a way that it is unclear whether what is meant is a process or its result. If we define, for example, the class *incision*, then we should make clear whether it is the process of making an incision that is intended or the incision itself that results therefrom. The fact that in ordinary and even in specialized languages the same word is quite often used to denote two different (albeit quite closely related) things contributes to such mistakes.

### 2.2.3 Adequacy of the OWL representation

Because the NCIT is distributed by means of OWL, we have also looked into the adequacy of this format as a knowledge representation for biomedical terminologies. We were specifically interested in the use of OWL's *complementOf* property. When applied to a target class, this defines a class whose extension is formed by the set of entities within a given domain that do not belong to the extension of this target class. Hence *complementOf* has some of the features of logical negation.

We also inspected the NCIT's usage of OWL's *someValuesFrom* and *allValuesFrom* restrictions, since there are fundamental problems associated with these restrictions. The restrictions are designed to allow an unambiguous reading of triples of the form *Class1 HasRelationshipWith Class2*, as in *Cell HasPart Cell wall*. Thus, when it is asserted that

*Class1 HasRelationshipWith someValuesFrom Class2*

this means that for any instance of Class1, there is at least one instance of Class2 to which it stands in the corresponding relationship. (It is then still allowed that an instance of Class1 may *in addition* stand in the same relationship to entities belonging to classes disjoint from Class2.)

An assertion involving the restriction *allValuesFrom*, in contrast, requires that if there are any instances that enjoy the given relationship with an instance of Class1, then all such instances must come from Class2. At the same time such an assertion is consistent with there being no instances from Class 2 at all for which the relationship holds. Thus an assertion to the effect that all middle left lobes of lung are made of green cheese using OWL's *allValuesFrom* restriction would be an allowable (indeed a true) assertion.

Ontological problems arise when these restrictions are used to capture spatial relationships. An OWL statement which (expressed in our own simplified syntax) would read:

*Human-Organ HasLocation allValuesFrom Human-Body-Region*

allows an instance of Human-Organ to have no location at all, which is clearly inconsistent with anatomical reality. But use of *someValuesFrom* here would be equally problematic, since then the OWL semantics would force any spatially located entity to be strictly located in a specific place (e.g. in the trachea, in the cranium) without the possibility of being displaced.

Note, in relation to the above, that the term ‘class’ has been used for the counterparts in the OWL representation of what are otherwise called ‘concepts’ in NCIT [15]. Classes are conceived by OWL as intensional meaning-entities; thus they are distinct from the *extensions* of concepts which are in other contexts often called ‘classes’. For the remainder of this communication – and with all due warnings [16] – we treat the two expressions ‘class’ and ‘concept’ as synonyms.

### 2.3 Analysis procedure

We used the multi-threaded SWI-Prolog version 5.4.3 [17] together with the Triple20 visualisation and editing tool [18] in order to inspect the Thesaurus in its OWL-version. We assumed that the OWL file was generated by the NCI after full classification of the system, and thus we did not ourselves reclassify the ontology. However, in light of the fact that Ontylog®, the particular Description Logic (DL) used to build the NCIT ontology [19], has a rather weak classificatory power, we used Triple20’s implementation of the full formal semantics of RDF(s) in order to gain access to those additional inferences which would not have been generated by the software used by NCIT’s developers.

Because our study envisaged a qualitative, rather than a quantitative, account of the problems encountered, we did not systematically search the NCIT for specific kinds of inconsistencies or ambiguities. Rather we started in a top-down manner, inspecting the system entry by entry until we found at least one violation for each principle. Our results are however presented in a structured fashion in order to provide a classification of the problems encountered, and we have assured ourselves by inspection that analogous problems are found also lower down the NCIT hierarchy.

## 3 Results

Triple20 was able to parse the NCIT OWL file in 212.59 sec CPU-time on a 2.5 GHz portable computer with 520 MB of RAM. 635,099 RDF triples were thereby identified (an RDF triple being the basic information unit of the RDF syntax [20]). The OWL parser rejected one concept: *Biodegradation_of _Xenobiotics_Pathway* (present in the browsable on-line version of the NCIT) because the space after the word '*of*' does not confirm to the rules governing OWL syntax. Since this concept has no children, however, the impact of this error on the generated structure is negligible.

We identified:

- 37,261 classes that represent NCI-concepts (see Fig. 1);
- 8,263 classes without assigned superclass, amongst which are 8,231 classes generated on the basis of OWL-restrictions (such as *rAnatomic_Structure_Has_Location someValuesFrom nci: Oral_Cavity*);
- 43 annotation properties providing background information related to the class described (such as *GenBank_Accession_number*, *SwissProt ID*, *synonym*),
- 90 object properties (such as *rAnatomic_Structure_Has_Location*) which are used to represent relationships among the classes defined inside the NCIT.

These figures do not relate exclusively to the RDF triples, classes, and properties defined in the NCIT. They include also triples, classes and properties that are part of the semantics of RDF, RDF(s) and OWL and which are generated by standard OWL-parsers prior to their parsing of an OWL-file. Examples are RDF(s) *container classes* such as *bags* or *sequences*, or *deprecated classes*, etc. Such classes provide a sort of upper level ontology on top of the actual content of an OWL file.

The numbers reflect the size of the NCIT ontology; they do not reflect any qualitative information.

**<Insert here Figure 1>**

### 3.1 Problems related to definitions

Many of ISO 1087-1:2000's requirements concerning definitions are frequently violated by the definitions in the NCIT.

From the total of 37,261 classes in the Thesaurus, 33,720 were stipulated to be *primitive* in the DL sense. This means that the majority of these classes are merely *described* rather than *defined*, with the consequence that only a small portion of the NCIT ontology can be used for purposes of automatic classification. In this connection one has to make a distinction between *formal* and *verbal* definitions. The former are provided (for those classes which are not primitives) for the purpose of allowing the corresponding classes to be automatically classified on the basis of an algorithm. The latter are provided to inform human users about what entities in the real world are allowable instances for the corresponding classes.

We found in the OWL file a total of 16,711 verbal definitions supplied by the NCI itself, together with some 5,368 definitions borrowed from elsewhere (primarily from MeSH or CSP). The numerical mismatch arises in virtue of the fact that some classes in NCIT are assigned more than one verbal definition (e.g. *Chromosomal Translocation* has three). On the other hand at least 55.2% of classes lack a definition in the Thesaurus, which can hardly be imagined to be the number of concepts the NCIT endorses as basic. Indeed, very many NCI concepts would benefit from a clear definition, since it is often hard to grasp what they stand for in reality and browsing the hierarchy often gives no further clues.

As an example, both *Test* and *Biological Testing* are direct subclasses of *Techniques*. *Test* is defined as: "*A procedure for critical evaluation; a means of determining the presence,* *quality, or truth of something*", while *Techniques* is defined as: "*Scientific or clinical procedures and methods*". *Biological Testing* is not defined at all. It has the subclasses *Bioassay* and *Toxicity test*, both of which seem to us to fit also the definition for *Test*. In this case therefore there is a double problem: on the one hand the definition of *Test* does not fulfill its purpose in differentiating one concept from other, related concepts; and on the other hand *Biological Testing* is clearly not a primitive concept, and so it should have a definition of its own.

Of all the many other problems encountered by users of the NCIT in virtue of its lack of definitions, we note only the puzzle raised by *Duratec*, *Lactobutyrin* and *Stilbene Aldehyde*, which are classified (!) as *Unclassified Drugs and Chemicals*. In the absence of a definition for the latter, it is hard to understand what the NCIT has in mind here. So-called residual categories ('other', 'NOS', etc.) do of course exist in many biomedical terminologies, though their inclusion has been subjected to much criticism [21]. Often, a residual class is interpreted as the complement of the union of all the non-residual siblings listed, though this interpretation causes obvious problems when a terminology is expanded to include more such siblings.

We found several entries where NCIT defines *words*, rather than *concepts*, for example in the definition of *Ontology* which reads:

> *The word ontology has a long history in philosophy, in which it refers to the study of being as such. In information science, an ontology is an explicit formal specification of how to represent the objects, concepts and other entities that are assumed to exist in some area of interest and the relationships among them.*

This definition fails to distinguish between use and mention of the word 'ontology' [22]. In addition, it runs together the specification of objects, concepts, etc., with the specification of methods for the representation thereof.

In the definition of *Iatrogenesis*, similarly, we find:

> *An iatrogenic condition is a state of ill health caused by medical treatment, usually due to mistakes made in treatment. The word literally means 'caused by a doctor'.*

This definition in fact defines *iatrogenic condition* and it tells us something about the word "iatrogenesis".

Another example of this problem is provided by the definition for *Antitubercular Agent*, which reads:

> *Drugs used in the treatment of tuberculosis. They are divided into two main classes: 'first-line' agents, those with the greatest efficacy and acceptable degrees of toxicity used successfully in the great majority of cases; and 'second-line' drugs used in drug-resistant cases or those in which some other patient-related condition has compromised the effectiveness of primary therapy.*

This contains definitions for two subordinate concepts and thus violates the principle according to which definitions should cover only one concept. Interestingly, the subordinate concepts defined do not exist in the NCIT, a fact which violates also another principle, namely that before drafting a definition for a given concept it is necessary to determine its relations to what ISO calls "other, related concepts".

Many definitions are descriptions rather than true definitions. Thus it would be better to restrict the definition for *Tuberculosis* – which currently reads:

> *A chronic, recurrent infection caused by the bacterium Mycobacterium tuberculosis. Tuberculosis (TB) may affect almost any tissue or organ of the body with the lungs being the most common site of infection. The clinical stages of TB are primary or initial infection, latent or dormant infection, and recrudescent or adult-type TB. Ninety to 95% of primary TB infections may go unrecognized. Histopathologically,*

> *tissue lesions consist of granulomas which usually undergo central caseation necrosis. Local symptoms of TB vary according to the part affected; acute symptoms include hectic fever, sweats, and emaciation; serious complications include granulomatous erosion of pulmonary bronchi associated with hemoptysis. If untreated, progressive TB may be associated with a high degree of mortality. This infection is frequently observed in immunocompromised individuals with AIDS or a history of illicit IV drug use. – 2004*

– to its initial sentence.

When the NCIT provides several definitions for the same concept these sometimes contain conflicting information. As an example, the concept *Disease Progression* enjoys three definitions:

(1) *Cancer that continues to grow or spread.*

(2) *Increase in the size of a tumor or spread of cancer in the body.*

(3) *The worsening of a disease over time. This concept is most often used for chronic and incurable diseases where the stage of the disease is an important determinant of therapy and prognosis.*

The first defines not *Disease Progression*, but rather a specific type of *Cancer*. The second conflicts with the third by limiting the concept of disease progression to neoplastic diseases. The third contains extraneous information, which should properly have been included in a note.

From the meta-tags provided in the NCIT one can infer that the three definitions come from different sources, and it would indeed be perfectly acceptable for the *term* “disease progression” to be used in one source to express the meaning that is captured by the second definition; but this is a totally different statement from the claim that the *concept: Disease Progression* is properly defined by this same definition.

To complicate matters, one of the subclasses of *Disease Progression* in the NCIT is *Cancer Progression*, which is defined as:

> *The worsening of a cancer over time. This concept is most often used for incurable cancers where the stage of the cancer is an important determinant of therapy and prognosis*.

One explanation for the etiology of the problems raised by definitions (1) and (2) above might be that they were wrongly associated at the level of the superordinate class, and that they should properly have been associated with the corresponding subordinates. The same is probably the case for the definition: *Includes both the vascular and non-vascular plants* (from the source tagged as "RAEB-2") that is currently assigned to the concept *Vascular Plant*.

### 3.2 Problems related to terms

Some of the NCIT's terms were specially created for this terminology. This group includes many which are designated 'preferred' terms, which means that the terms in question should also be such as to satisfy the principle set forth above pertaining to acceptability.

Clearly, capitalization of the first letter of all words, the standard procedure in the NCIT, is not linguistically correct for the English language. In addition, it hampers the potential use of the NCIT for text mining purposes or semantic indexing of documents (for example because of confusion with names of proprietary products).

The business of the NCIT, we are told, is to define *concepts*. Consider:

> *The National Cancer Institute has developed the NCI Thesaurus, a biomedical vocabulary that provides consistent, unambiguous codes and definitions for concepts used in cancer research* [23].

If one takes this statement seriously, then finding in the Thesaurus a concept named "*Conceptual Entities*" is worrying, to say the least. The associated definition does, it is true, provide us with some assistance in working out the proper interpretation of this concept. It reads: "*An organizational header for concepts representing mostly abstract entities*". Unfortunately, however, inspection of the subordinate classes reveals that they are mostly not abstract at all, but rather highly concrete, including: *action*, *change*, *color*, *death*, *event*, *fluid*, *injection*, *temperature* (and many others in similar vein). Moreover the definition itself contravenes the principle mentioned already above to the effect that definitions should define *concepts* and not *words*. (We hasten to point out that the NCIT is not by any means alone in having troubles with the weasel phrase "conceptual entity".)

Many terms are not precise, i.e. they do not capture the intended meaning. Imprecise terms are especially problematic in the absence of good definitions. Thus for example the term *Anatomic Structure, System, or Substance* does not give us any clue as to whether the scope of the adjective *anatomic* is restricted to *structure* or extends also to *system* and *substance*. If it is so restricted, then one may wonder why *Drugs and Chemicals* are not classified under this concept, since these are clearly substances. If, on the other hand, it is not so restricted then one may question the status of the term as preferred term. Google finds only 6 hits for the term "anatomic substance", where, according to ISO, preferred terms should be those members of groups of terms which have the highest acceptability rating.

The NCIT stretches the meaning of "synonym" in such a way that the claimed synonymy of numerous terms in the NCIT cannot be accounted for even under ISO's more relaxed definition of "quasi-synonym".

Notable examples of problematic synonyms are:

***Biological Function / Biological Process***: the problem here is that function and process are ontologically quite distinct, in a way that is crucially important especially for the purposes of our understanding of the nature and scope of clinical medicine [24]. Certainly, some biological processes are the *exercises* of biological functions. (The pumping of your heart is the exercise of the function: *to pump blood*.) Other biological processes, however – including at least the majority of pathological processes – are not. There is no organ or organ part which has the function: *to ulcerate* or: *to become cancerous*. Moreover, there are many instances of anatomical entities whose biological functions lie dormant and are thus never expressed or exercised at all. Here functions exist in the absence of any associated processes of function*ing*. Perhaps the claimed synonymy applies only within the context of the NCIT itself. This, however, would imply a serious weakness of the system, since it would mean that it is not able to differentiate between functions and processes.

***Anatomic Structure, System, or Substance / Anatomic Structures and Systems***: since a substance is not a structure, and not a system either, the two terms cannot be synonyms.

***Organism / Organisms***: this is a matter, not of synonymy, but of lexical variation. Inflected terms are not new terms that stand in a synonymy relation with an original term, but are merely variants of the same term. In addition, the claim that a plural term refers to the same real entity as does the corresponding singular term is from an ontological point of view an egregious error. A single organism cannot be the same real entity as a collection of organisms.

***Genetic Abnormality / Molecular Abnormality*** (with subclass “Molecular Genetic Abnormality”). Neither concept is provided with a definition in NCIT.

***Diseases and Disorders / Disease / Disorder***: the first term suggests that the NCIT considers diseases and disorders to be different entities; the synonymy declaration, however, suggests the opposite. The first term of the triple is thus ambiguous and is in consequence not the best choice as preferred term for the corresponding concept. Also the two definitions that are provided are not helpful. One reads:

> *A disease is any abnormal condition of the body or mind that causes discomfort, dysfunction, or distress to the person affected or those in contact with the person. Sometimes the term is used broadly to include injuries, disabilities, syndromes, symptoms, deviant behaviors, and atypical variations of structure and function*.

In addition, the second part of this definition violates the principle that only concepts should be defined, not words. The second definition:

> *A definite pathologic process with a characteristic set of signs and symptoms. It may affect the whole body or any of its parts, and its etiology, pathology, and prognosis may be known or unknown*,

is even more problematic, since it suggests that *pathologic process* is the superordinate concept, which is in contradiction to the actual place of this concept in the NCIT's own hierarchy.

### 3.3 Ontological problems

The most important part of a terminology, according to the relevant ISO standards, is the concept system that underlies it. It is the latter which forms the basis upon which concept definitions rest. Unfortunately, it is very difficult to perform an in-depth analysis of the NCIT concept system, precisely because of the many ambiguities at the level of its terms and definitions. Moreover, one can in many cases only speculate what the real entities are that are supposed to qualify as instances for the concepts which are included. It is nonetheless still possible to identify mistakes by examining a system like the NCIT from the point of view of a principles-based ontology.

No motivation is given for the NCIT's specific choice of its 21 top classes, and some of the choices made seem questionable, to say the least. As an example, we would not expect *Abnormal Cell* to be one of the top classes in a concept system, since intuitively it should be subsumed either by *Cell* or by one or other of the subclasses of the (itself somewhat incoherently constituted class of) *Diseases, Disorders and Findings*. The NCIT does indeed have a class *Cell*, but this is a subclass of another problematic class: *Other Anatomic Concept* (so that cells themselves would be concepts (!) in the eyes of the NCIT). Moreover, the NCIT also has the further class *Normal Cell*, which is a subclass of *Microanatomy*.

None of the three classes *Cell*, *Normal Cell* and *Abnormal Cell* are related to each other in any way in the NCIT. Of course, no concept system or ontology will ever be complete, and one could argue that the non-expression of such relationships in the system does not mean that their existence in reality is denied. This argument cannot be applied to the NCIT, however, because of its requirement that each class must belong to exactly one *kind* [25]. For from this it follows that neither *Normal Cell* nor *Abnormal Cells* are *Cells* within the context of the NCIT.

The same requirement forces also the class *Oncolytic Virus* to be classified as either a *Drug_or_Chemical_Kind* or as an *Organism_Kind*. It is the former choice that is made by the NCIT; yet the definition for *Oncolytic Virus* reads:

> *Manipulated or engineered* <u>*viruses*</u> *having oncolytic properties to selectively replicate in and kill targeted cancer cells, leaving normal cells unharmed*

– which is another example of a mismatch between definition and the positioning of classes within the NCIT hierarchy.

The most fundamental problem for the NCIT, however, is the unprincipled way in which its class hierarchy is built up. For this means that it ignores the basic ontological distinctions between continuants and occurrents on the one hand, and dependent and independent entities on the other. The *Conceptual Entities* class is, again, a conspicuous example of this defect, since it subsumes classes of the most diverse types:

- dependent occurrents such as *Action*, *Assignment*, *Event*, *Injection*, *Inhalation, Movement, Proliferation*. All of these comprehend entities which occur over time and depend on continuant entities as their participants;
- dependent continuants such as *Color*, *Shape* and *Temperature*, which endure through time and depend on  continuant entities as their bearers;
- independent continuants such as *Stain* (defined as "*A dye or other coloring material that is used in microscopy to make structures visible*"),
- roles, such as *Purpose*, *Reason*, *Agent*, that can be attributed to entities only in relation to some other entities.

Analogous problems are found in other hierarchies as well, though not so abundantly. Thus for example *Bilateral Disease* is subsumed by *Clinical Course of Disease* (occurrent), rather than by *Disease* (continuant).

### 3.4 Problems with NCIT's OWL representation

We noticed an inconsistent use by NCIT of the OWL-qualifiers *allValuesFrom* and *someValuesFrom*. Consider, as an example, NCIT's treatment of the class *Pleura*. Instances of this class, as well as instances of *Bronchial Tree* and *Lung*, are related spatially to instances of the class *Thoracic Cavity*. Unfortunately, however, the NCIT expresses this relation by using the property *rAnatomic_Structure_Has_Location* in two inconsistent ways. On the one hand *Pleura* is locally restricted with the qualifier *allValuesFrom* to the class *Thoracic Cavity*; on the other hand *Bronchial Tree* and *Lung* are locally restricted to the same class with the qualifier *someValuesFrom*. This means however that for all instances of the class *Pleura*, every assigned location must be an instance of the class *Thoracic Cavity*, while for all instances of *Lung*, only at least one location must be an instance of *Thoracic Cavity*. It is not clear why this difference is made.

There is also a more fundamental problem associated with the use of the *allValuesFrom* restriction: it allows instances of *Pleura* not to have a location at all, which clearly is in contradiction with physical reality. But use of *someValuesFrom* here would be equally problematic. For while its requirement that instances of *Lung* be located in at least one instance of *Thoracic Cavity* is acceptable in normal anatomical circumstances, it is not valid for lungs that have been resected in total, at least under what we would call a sensible interpretation of the *rAnatomic_Structure_Has_Location* property. Unfortunately, however, a detailed semantics for this (and all the other) relationships in the NCIT is not given.

While the *complementOf* property, which is OWL's counterpart of logical negation, is not used explicitly in the NCIT, there are a number of places in the hierarchy where its application is suggested by the use of terms of the form *non-X*, which would normally be taken to signify logical negation. If such terms are interpreted by appealing to OWL's *complementOf*, however, then this yields further inconsistencies, for example in NCIT's treatment of the subclasses of *Plant*, which are currently: *Vascular Plant, Non-Vascular Plant*, and *Valeriana Officinalis*.

Something similar holds for the top-level class *Diagnostic and Prognostic Factor*. This subsumes, as one would expect *Diagnostic Factor* and *Prognostic Factor*; but it subsumes also the additional classes *Biomarker, Risk Factor* and *Treatment Factor*.

## 4 **Discussion:**

With the development of modern formal disciplines (formal logic, and the computational disciplines which have arisen in its wake) we have learned a great deal about the criteria which must be satisfied if a terminology is to be structured in such a way that the information expressed by its means can be extracted via automatic procedures in a maximally effective way. Unfortunately, existing biomedical controlled vocabularies have been developed in large part without concern for these criteria – and this applies both to the terms they contain and to the relations and definitions associated with these terms. The NCIT, as our analysis shows, is no exception to this rule, in spite of the fact that it is described by its authors as “*a controlled terminology which exhibits ontology-like properties in its construction and use*” [26].

One of the reasons for the identified shortfalls lies in the way the NCIT was constructed:

> *by bootstrapping the initialization of NCI Thesaurus from existing terminologies, the project gained the co-operation of diverse stakeholders and avoided pitfalls associated with trying to develop a science based terminology de novo* [9, p. 36].

For by selecting this route the NCI has taken over some of the characteristic errors of the terminologies from which it draws, and especially some of the characteristic inconsistencies of the UMLS [27, 28]. Some of the sources used by NCIT are glossaries and vocabularies, and hence it is wrong to simply take over their definitions in a system of concept representations. This is because glossaries and vocabularies specify only the meanings which come to be associated with given terms in given contexts, and this implies a less rigorous requirement than that which must be met when defining concepts. For in the latter case the task is to explain what given entities in reality – the entities which fall under the concepts in question – share in common and how they relate to each other in a way which serves also to differentiate defined concepts from their neighbors.

We can distinguish three levels of organization of the terminologies and ontologies currently employed in bioinformatics. At the bottom are subject descriptor resources such as MeSH, used primarily for literature indexing purposes. Next are vocabulary resources such as SNOMED-CT or the Gene Ontology [3, 24], which enjoy a more coherent formal organization and which may involve the use of DL-based formal tools. Then come ontologies such as the Foundational Model of Anatomy [29], which manifest not only well-structured and mutually consistent hierarchies but also respect the basic ontological distinctions drawn by philosophers [30].

All of the above are designed in part to serve communicative needs, for example in clarifying and stabilizing the terms (words, names) used in a specialized domain [31]. The attempt to satisfy these needs is however often associated with a view which identifies concepts with the meanings of words and this has encouraged terminology-builders to focus on such meanings (or on the associated ideas in people's heads) rather than on the corresponding entities in reality [32]. In this way the properly ontological features of a terminology have been underemphasized, to the degree that the difference is often obscured between ontology and epistemology [33], the discipline which focuses not on objects in reality but rather on our ways of gaining or expressing our knowledge of such objects.

In resolving the resultant problems we are unfortunately not helped by the pertinent ISO standards, where the crucial definitions are themselves unclearly formulated. Thus in 1990 ISO-1087 defines a concept as: A unit of thought constituted through abstraction on the basis of properties common to a set of objects. A characteristic it defines as: A mental representation of a property of an object serving to form and delimit its concept. Ten years later ISO's definition of concept reads: A unit of knowledge created by a unique combination of characteristics. Characteristic itself is defined as: An abstraction of a property of an object or of a set of objects. Object is defined as: Anything perceivable or conceivable (a unicorn being given as a specific example of the latter). These definitions provide no help at all, for example, if we wish to know whether 'object' does or does not comprehend occurrents as well as continuants.

On behalf of the NCIT it might be argued that its purposes, too, are precisely to support communication, and that for these purposes linguistic and mental representations are precisely the proper objects of focus. The ISO Standards support a view along these lines for example in its assertion that for terminology purposes "object" means anything perceived or conceived, either existing (such as people, cars, and bridges), or non-existing and purely imagined (such as unicorns or literary characters). For what matters for the relevant ISO terminology standards is that if somebody has something in mind, whether existing in reality or not, then he must be able to convey information about it in such a way that this information is understandable to a third party. Indeed, as ISO puts it: "*In the course of producing a terminology, philosophical discussions on whether an object actually exists in reality are ... to be avoided*". [2, p. 2].

Unfortunately, however, as our arguments in the foregoing have shown – and as the NCI implicitly recognizes in its statement to the effect that the NCIT "*exhibits ontology-like properties*" – when one ignores in this spirit properly ontological principles, then the results, in terms of mis-classifications and mis-definitions, hamper the very purposes of communication for which the terminology was designed.

It is sometimes claimed that the use of a formal language, for example of one or other Description Logic, can help us to avoid mistakes of the sort described. Unfortunately, however, this is not true, at least not if such languages/logics are used naively. Certainly the semantics of representational languages such as OWL allow one to reason consistently inside the associated models; but they do not guarantee that such models do in fact amount to any sensible representation of reality. This has also been observed by other authors who subjected parts of the OWL-representation of the NCIT to a semantic analysis [34]. One should avoid, therefore, any assumption to the effect that the provision of an OWL-representation is already in and of itself sufficient to guarantee that a system like the NCIT will satisfy the requirements of a sound ontology.

The claim that terminologies encapsulated in an OWL-like formal framework are understandable both to humans and to machines should also be treated with a pinch of salt. Of course it is true that representing a terminology by means of OWL allows one to display its content in a way that makes it easier to carry out certain types of inspection and to discover certain types of inconsistencies. But as anyone who has utilized DL-based instruments for terminology and ontology management can testify, such instruments can yield valuable results only on the basis of arduous manual preparation. In the case of the NCIT, we are faced with the additional problem that its machine-readable DL-structure represents a mere fragment of the total ontological structure incorporated in the system (the additional ontological content being graspable by humans in virtue of both their informal understanding of the terms involved and of their acquaintance with the corresponding referents in reality). This means, however, that the ontological part of the NCIT does not correspond to the terminological part, so that humans and machines will understand the NCIT in different and conflicting ways. Making a system like the NCIT truly machine-interpretable, for example in the context of communication among software agents, would require a much deeper and more principled representation [35], and it would require, again, conformity to the sorts of ontological principles we have outlined above.

## 5 Conclusion

In analyzing the NCI Thesaurus we were particularly interested in how the claimed ontological features of the system work together with its terminological parts. We found that the system suffers from the same problems encountered in so many of the biomedical terminologies produced in recent years. The NCIT is, we are confident, a useful tool for the internal purposes of the NCI, which must be given credit for trying to bridge the clinical and basic biology terminology realms in a single resource. It must be given credit also for its sophisticated technology for keeping track of updates, as well as for being one of the earliest to federate its ontology operationally with another ontology system (MGED Ontology) and for trying to harmonize with external ontology modeling practices. The NCI Thesaurus is a never-ending work in progress, the content of which is dictated by the needs of its users and customers. If, however, it wants to establish itself as a useful and trustworthy terminological resource and to play the role of a reference ontology in other contexts, then a considerable effort will have to be made in order to clean up its hierarchies and to correct the definitions and ambiguous terms which they contain. We strongly suggest the use in this endeavor of a principles-based methodology that will allow the NCIT to be tested not just for internal consistency but also for consistency with that part of reality which it is intended to represent.

## 6 References

[1] Cantor MN, Lussier YA. Putting data integration into practice: using biomedical terminologies to add structure to existing data sources. In: Musen MA, editor. AMIA 2003. Proceedings of AMIA 2003 Annual Symposium; 2003 Nov 8-12, Washington D.C., USA. AMIA; 2003. p. 125-9.

[2] Kumar A, Smith B. The Unified Medical Language System and the Gene Ontology, KI 2003: Advances in Artificial Intelligence (Lecture Notes in Artificial Intelligence 2821), 2003; 135–48.

[3] Smith B, Williams J, Schulze-Kremer S. The ontology of the gene ontology. In: Musen MA, editor. AMIA 2003. Proceedings of AMIA 2003 Annual Symposium; 2003 Nov 8-12, Washington D.C., USA. AMIA; 2003. p. 609-13.

[4] Grenon P, Smith B, Goldberg L. Biodynamic ontology: Applying BFO in the Biomedical Domain, in Pisanelli DM (ed). Ontologies in Medicine. Proceedings of the Workshop on Medical Ontologies, Rome October 2003. IOS Press, Studies in Health Technology and Informatics, vol 102, 2004. p. 20-38.

[5] Ceusters W, Smith B, Kumar A, Dhaen C. Ontology-Based Error Detection in SNOMED-CT®. In: M. Fieschi, E. Coiera and Y-C.J. Li, editors. MEDINFO 2004. Proceedings of the 11th World Congress on Medical Informatics; 2004 Sep 7-11, San Francisco, CA, USA. Amsterdam: IOS Press; 2004. p. 482-6.

[6] Smith B, Rosse C. The role of foundational relations in the alignment of biomedical ontologies. In: M. Fieschi, E. Coiera and Y-C.J. Li, editors. MEDINFO 2004. Proceedings of the 11th World Congress on Medical Informatics; 2004 Sep 7-11, San Francisco, CA, USA. Amsterdam: IOS Press; 2004. p. 444-8.

[7] Ceusters W, Smith B, Kumar A, Dhaen C. Mistakes in Medical Ontologies: Where Do They Come From and How Can They Be Detected? in Pisanelli DM (ed) Ontologies in Medicine. Proceedings of the

Workshop on Medical Ontologies, Rome October 2003. IOS Press, Studies in Health Technology and Informatics, vol 102, 2004;145-164.

[8] Kumar A, Schulze-Kremer S, Smith B. Revising the UMLS Semantic Network. In: M. Fieschi, E. Coiera and Y-C.J. Li, editors. MEDINFO 2004. Proceedings of the 11th World Congress on Medical Informatics; 2004 Sep 7-11, San Francisco, CA, USA. Amsterdam: IOS Press; 2004. p. 1700-4.

[9] S. de Coronado, M. W. Haber, N. Sioutos, M. S. Tuttle, and L. W. Wright. NCI Thesaurus: Using Science-based Terminology to Integrate Cancer Research Results. In: M. Fieschi, E. Coiera and Y-C.J. Li, editors. MEDINFO 2004. Proceedings of the 11th World Congress on Medical Informatics; 2004 Sep 7-11, San Francisco, CA, USA. Amsterdam: IOS Press; 2004. p. 33-7.

[10] Open Biological Ontologies. http://obo.sourceforge.net/. Last visited 2005, Jan 24.

[11] National Cancer Institute, Office of Communications, Center for Bioinformatics. NCI Terminology browser, ftp://ftp1.nci.nih.gov/pub/cacore/EVS/. Last visited 2005, Jan 18.

[12] National Cancer Institute, Office of Communications, Center for Bioinformatics. NCI Terminology browser, http://nciterms.nci.nih.gov/NCIBrowser/Startup.do. Last visited 2005, Jan 18.

[13] Laboratory for Applied Ontology. DOLCE : a Descriptive Ontology for Linguistic and Cognitive Engineering. http://www.loa-cnr.it/DOLCE.html. Last visited 2005 Jan 18.

[14] Ceusters W, Smith B. Ontology and Medical Terminology: why Descriptions Logics are not enough. Proceedings of the conference Towards an Electronic Patient Record (TEPR 2003), San Antonio, 10-14 May 2003 (electronic publication).

[15] W3C. OWL Web Ontology Language Reference. Recommendation 10 February 2004 (http://www.w3.org/TR/owl-ref/). Last visited 2005 Jan 18.

[16] Gamper J, Nejdl W, Wolpers M. Combining Ontologies and Terminologies in Information Systems. In Proc. 5th International Congress on Terminology and Knowledge Engineering, Innsbruck, Austria.

[17] Wielemaker J., Native Preemptive Threads in SWI-Prolog, in Catuscia Palamidessi (ed.) Practical Aspects of Declarative Languages, Springer Verlag, Berlin, Germany, 2003; 331-345.

[18] Wielemaker J. Triple20: an RDF triple viewer and editor. http://www.swi-prolog.org/packages/Triple20/Triple20.html. Last visited 2005, Jan 18.

[19] Golbeck J, Fragoso G, Hartel F, Hendler J, Oberthaler J, Parsia B, The National Cancer Institute's Thesaurus and Ontology. Journal of Web Semantics, vol. 1, # 1, 75-80, 2003. (http://www.mind-swap.org/papers/WebSemantics-NCI.pdf)

[20] W3C. Resource Description Framework (RDF): Concepts and Abstract Syntax; Recommendation 10 February 2004. http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/. Last visited 2005 Jan 18.

[21] College of American Pathologists. SNOMED Clinical Terms Consultation Document; Requirements Analysis. Version 10, 2000 Oct 12.

[22] Swartz N. Use and Mention. http://www.sfu.ca/philosophy/swartz/use&mention.htm. Last visited 2005 Jan 24.

[23] de Coronado S, Fragoso G. Enterprise Vocabulary Development in Protege/OWL: Workflow and Concept History Requirements. Expanded Abstract for Protégé Workshop Jul 6-9, 2004 (http://protege.stanford.edu/conference/2004/abstracts/DeCoronado.pdf).

[24] Smith B, Williams J, Schulze-Kremer S. The ontology of the gene ontology. In: Musen MA, editor. AMIA 2003. Proceedings of AMIA 2003 Annual Symposium; 2003 Nov 8-12, Washington D.C., USA. AMIA; 2003. p. 609-13.

[25] National Cancer Institute, Office of Communications, Center for Bioinformatics NCI Thesaurus Semantics. ftp://ftp1.nci.nih.gov/pub/cacore/EVS/ThesaurusSemantics/. Last visited 2005 Jan 18.

[26] Hartel F, Warzel DB, Covitz P. OWL/RDF/LSID Utilization in NCI Cancer Research Infrastructure. W3C Workshop on Semantic Web for Life Sciences, October 27-28 2004, Cambridge, Massachusetts, USA.

[27] Hahn U, Schulz S. Towards a broad-coverage biomedical ontology based on description logics. Pac Symp Biocomput. 2003:577-88.

[28] Cimino J. Auditing the Unified Medical Language System with Semantic Methods. J Am Med Inform Assoc. 1998 January; 5 (1): 41–51

[29] Rosse C, Mejino JL Jr. A reference ontology for biomedical informatics: the Foundational Model of Anatomy. J Biomed Inform. 2003 Dec;36(6):478-500.

[30] Spyns P, De Bo J. Ontologies, a revamped cross-disciplinary buzz-word or a truly promising interdisciplanry research topic ? STAR Lab Technical Report STAR-2004-20.

[31] C.K. Ogden & I.A. Richards, The Meaning of Meaning, London (1923).

[32] Smith B. Beyond Concepts: Ontology as Reality Representation. In: Varzi AC and Vieu L, editors. FOIS 2004. Proceedings of. The International Conference on Formal Ontology and Information Systems; 2004 Nov 4-6, Turin, Italy. Amsterdam IOS Press; 2004, 73–84

[33] Bodenreider O, Smith B, Burgun A, The Ontology-Epistemology Divide: A Case Study in Medical Terminology. In: Varzi AC and Vieu L, editors. FOIS 2004. Proceedings of. The International Conference on Formal Ontology and Information Systems; 2004 Nov 4-6, Turin, Italy. Amsterdam IOS Press; 2004. 185–195.

[34] Fischer DH. Converting a Thesaurus to OWL: Notes on the Paper "The National Cancer Institute's Thesaurus and Ontology" (http://www.ipsi.fraunhofer.de/orion/pubFulltexts/NCIReview18Feb04.pdf)

[35] Schneider L, Cunningham J. Ontological Foundations of Natural Language Communication in Multiagent Systems. IFOMIS Report ISSN 1611-4019.

```
<owl:Class rdf:ID="Pleural">
      <rdfs:label>Pleural</rdfs:label>
      <code>C25223</code>
      <hasType>primitive</hasType>
      <rdfs:subClassOf rdf:resource="#Anatomy_Modifier"/>
      <Preferred_Name>Pleural</Preferred_Name>
      <Semantic_Type>Spatial Concept</Semantic_Type>
      <Synonym>PLRL</Synonym>
      <Synonym>Pleural</Synonym>
      <FULL_SYN><![CDATA[<term-name>PLRL</term-name><term-group>AB</term-
            group><term-source>CADSR</term-source>]]></FULL_SYN>
      <FULL_SYN><![CDATA[<term-name>Pleural</term-name><term-
            group>PT</term-group><term-source>CADSR</term-source>]]>
            </FULL_SYN>
      <DEFINITION><![CDATA[<def-source>NCI</def-source><def-definition>
            Pleural; of, or pertaining to, the pleura.</def-
            definition>]]></DEFINITION>
      <UMLS_CUI>C0205040</UMLS_CUI>
</owl:Class>
```

Figure 1. OWL - representation of the NCIT class “pleural”