Ontology-Based Error Detection in SNOMED-CT®

Werner Ceusters[a], Barry Smith[b], Anand Kumar[b], Christoffel Dhaen[a]

[a] *Language & Computing nv, Zonnegem, Belgium*
[b] *Institute for Formal Ontology and Medical Information Science, Leipzig, Germany*

forthcoming in *Proceedings of Medinfo* (San Francisco, September 2004)

Abstract

Quality assurance in large terminologies is a difficult issue. We present two algorithms that can help terminology developers and users to identify potential areas of improvement. We demonstrate the methodology by applying the algorithms to one of the most popular terminologies, SNOMED-CT®. Analysis of the results provides evidence for the thesis that both formal logical and linguistic tools should be used in the development and quality-assurance process of large terminologies.

Keywords

Medical natural language understanding, medical terminologies, formal ontology, quality assurance.

Introduction

The main goal of Language and Computing nv (L&C) is to deliver advanced natural language understanding (NLU) applications directed primarily towards terminology management, coding, semantic indexing, and information retrieval and extraction. NLU requires both knowledge about reality (i.e. about what is described by using language) and knowledge about language itself (so that one can assess a language user's current perspective on reality by understanding how he is using language to describe it). To achieve these ends, L&C has developed LinKBase®, a realist ontology for the healthcare domain, and LinKFactory®, an ontology authoring and management system. Since 2002 LinKBase® has been developed in close collaboration with IFOMIS, the Institute for Formal Ontology and Medical Information Science of the University of Leipzig, which is developing a framework for ontology construction and alignment in the biomedical domain based on rigorous formal definitions and axioms [1]. IFOMIS starts out from the idea that we need to understand the general structure and organization of a given domain (its ontology) before we start building software models. This means above all that the basic categories and relations structuring the domain should to the greatest possible extent be formally defined in a logically rigorous way. Such definitions and associated axioms of basic ontology should then serve as constraints on coding and on manual correction of associated ontologies and terminology systems.

In part, as a result of the collaboration with IFOMIS, both LinKBase® and LinKFactory® have now reached a level of maturity that enables them to be used to assess the quality of external systems. This document describes the results of using LinKBase® and two specific algorithms implemented in LinKFactory® to carry out a still on-going review of the January 2003 and July 2003 versions of SNOMED-CT® for possible mistakes and inconsistencies. It explains the basic mechanisms of the approach, as well as the various types of inaccuracies that can be detected, and draws conclusions for the methodology of quality assurance in large terminologies in the future.

Materials and Methods

LinKBase®

LinKBase® is a large-scale medical ontology developed by L&C using the ontology authoring environment LinKFactory® [2]. LinKBase® contains over 1.5 million language-independent medical and general-purpose concepts, associated with more than 4 million terms in several natural languages [3]. A *term* consists of one or more *words*, which may be associated with other concepts in their turn. Concepts are linked together into a semantic network in which some 480 different link types are used to express formal relationships. The latter are derived from formal-ontological theories of mereology and topology [4, 5], time and causality [6], and also from the specific requirements of semantics-driven natural language understanding [7, 8]. Link types form a multi-parented hierarchy in their own right. At the heart of this network is the formal subsumption (is-a) relationship, which in LinKBase® covers only some 15% of the total number of relationships involved. Currently, the system is being re-engineered in conformity with the theories of Granular Partitions [9] and Basic Formal Ontology [10].

LinKFactory®'s TermModelling Algorithm

The TermModelling algorithm uses conceptual and linguistic information in order to find entities in the LinKBase® ontology corresponding to terms or concept-labels drawn from external medical terminologies. The algorithm works by attempting to find entities that enjoy the closest (where possible an exact) match to given input terms. Each returned entity is accompanied by a numerical index expressing the effort exerted by the algorithm (the cost it had to pay) in order to retrieve the related entity. This index can be used as a measure of semantic distance from the input term to the retrieved entities. The higher its value the more distantly related the entity should be with respect to the input term. The algorithm makes use not only of terms in their original forms but also of linguistic variants and of ontological descriptions of the corresponding entities generated by LinKBase®. A complete analysis of the algorithm is given in [11]. When applied to the task of quality assurance for terminologies, the algorithm is used in two different settings. In the first, the ranking of the semantic distances of the various retrieved entities with respect to each given input term is assessed manually for accuracy by domain experts. If the ranking of the retrieved entities obtained by the algorithm by calculating their semantic distance to the input term is judged by an expert as inaccurate, then this is taken as a prima facie indication of inappropriate modelling in the source terminology and can rest on factors ranging from underspecification, misclassification, unresolved disambiguation (i.e. the ontology might not be aware of the various meanings for homonyms) or even plain mistakes. As an example, the semantic distance for the retrieved entity “freeing of adhesion of muscle of hand” with respect to the query term “lysis of adhesions of fascia” must be higher than the one for “lysis of adhesion of muscle”, as the second should subsume the first. This is because the path to a term always goes via the terms that subsume it. Scores for subsuming terms are therefore always lower than those for subsumed terms. The drawback of this method is its need for manual verification of the results. However, statistical methods can also be used to scan for unusual distributions of semantic distances, such as the difference in semantic distance between the Nth and (N+1)th ranked entity being further than the mean difference over all entities, or the semantic distance of the highest ranked retrieved entity for a specific query term being further than the mean semantic distance of all the highest ranked entities over all query terms.

In a second setting, the output of the algorithm for a given term is compared to the output resulting from using formal relationships without linguistic information. A different ranking of retrieved entities, which can be flagged automatically, is here a strong indication of inconsistencies, either at the level of the LinkBase® descriptions or in the terminology system from which the term is drawn. This is because formal constraints on correct coding, for example constraints relating to mereological or topological relations or to distinctions between objects and processes, always overwhelm the ways in which natural language represents the corresponding entities. To process large volumes of terms, typically deriving from third party terminologies, the TermModelling algorithm is embedded in a special component of LinkFactory® called OntologyMatcher that uses a blackboard process control system to allow analyses to be performed in background mode.

LinKFactory®’s Classifier Algorithm

Our second algorithm is a Description Logic (DL)-based classifier optimised for working with extremely large terminology systems. This algorithm not only computes subsumption relations on the basis of necessary and sufficient conditions for entities defined by the external terminology, it also proposes new entities to be added, based on the distribution of entity-characteristics as computed during the analysis. Parameters can be set for the types of entities generated according to various principles [12]. One simple (but useful) example of such a principle is: if there is a type of object that causes a specific type of infection then there are infections necessarily caused by objects of that type. With this algorithm, both under- and overspecification of entities can be identified automatically via comparison of the original subsumption hierarchy with the generated one.

SNOMED-CT®

SNOMED-CT® is a terminology system developed by the College of American Pathologists. It contains over 344,000 concepts and was formed by the merger, expansion, and restructuring of SNOMED RT® (Reference Terminology) and the United Kingdom National Health Service Clinical Terms (also known as the Read Codes).

Approach

LinKFactory®’s Ontologymatcher component used the terms of SNOMED-CT® to find related concepts in LinKBase®. The generated lists were examined manually to find superficial indicators for inconsistencies. Those SNOMED-CT® concepts deemed most prone to error were then subjected to a process of detailed examination that is still ongoing. In addition, the January 2003 version of SNOMED-CT® was processed by LinKFactory®’s classifier algorithm to find missing pre-coordinated concepts such as “abscess of central nervous system” purely on the basis of what is contained in the third-party terminologies, i.e. without taking advantage of any LinKBase® information. We established empirically that, where existing concepts are the unique children of such generated pre-coordinations, this is a good indication of questionable or incomplete modelling of the concepts involved. Formal proof of this finding is still forthcoming. Note that although the experiment involved elements of manual checking, neither the system nor the manual checkers were instructed as to the types of inconsistencies that might be detected. Thus, none of the types of inconsistencies reported here were sought out *a priori*. Rather their detection is in each case an incidental by-product of the approach to mapping external terminologies such as SNOMED-CT® into the LinKBase® environment.

Results

What follows is a brief analysis of output generated by the TermModelling algorithm when applied to the January and July 2003 versions of SNOMED-CT®. The analysis is not complete, and more work is required to yield an exhaustive list of possible inconsistencies. To assist with comprehension of the discussion section of this paper, we assign an identifying number of the form "Ja-#", "Ju-#", or "Jau-#" to each reported mistake or inconsistency, indicating presence in the January, July or in both versions of the system, respectively.

Human error

Some mistakes appear to originate from the inevitable human error that occurs with manual modelling. The following are some of the types of errors we found under this heading:

Improper assignment of ISA relationships

The concept "265047004: diagnostic endoscopic examination of mediastinum NOS" is subsumed by "309830003: mediastinoscope". Thus, a procedure is classified as an instrument (Jau-1). SNOMED-CT® marks the mentioned procedure concept as "limited", meaning that it is of limited clinical value, as it is based on a classification concept or an administrative definition. Yet SNOMED-CT® still considers concepts with this status as valid for current use and as active. Another example has a procedure wrongly subsumed by a disease: thus the concept "275240008: Lichtenstien repair of inguinal hernia" is directly subsumed by "inguinal hernia" (Jau-2).

Oversights of this type can be further divided into: 1) *Improper treatment of negation*: the concept "203046000: Dupuytren's disease of palm, nodules with no contracture" is subsumed by the concept "51370006: contracture of palmar fascia" (Jau-3); and 2) *Improper treatment of the partial/complete distinction.* This is a numerically big category. As an example, the concept "359940006: partial breech extraction" is subsumed by the concept "177151002: breech extraction" which in turn is subsumed by "237311001: complete breech delivery" (Jau-4).

Improper assignment of non-ISA relationships:

The concept "51370006: contracture of palmar fascia" is linked by means of SNOMED's Finding Site relationship to the concept "plantar aponeurosis structure". Probably as a consequence of automated classification, the concept is wrongly subsumed by "disease of foot" since "plantar aponeurosis structure" is subsumed by "structure of foot" (Jau-5). A similar phenomenon is observed in the concept "314668006: wedge fracture of vertebra", which is subsumed by "308758008: collapse of lumbar vertebra" (Ja-6). Although the wrong subsumption is no longer present in the July version, the wrong association via Finding Site "bone structure of lumbar vertebra" is still present (Jau-7). Equally the concept "30459002: unilateral traumatic amputation of *leg* with complication" is classified as an "open wound of *upper limb* with complications" due to an erroneous association with Finding Site "upper limb structure" (Jau-8).

Technology-induced mistakes

A first example of this type has been referred to already above (Jau-5): wrong subsumption because of inappropriately assigned relationships. Other errors are probably induced by tools that perform lexical or string matching. We can hardly imagine that a human modeller would allow the concept "9305001: structure of labial vein" to be directly subsumed by both "vulval vein" and "structure of vein of head". The error probably comes from an unresolved disambiguation of the word "labia" that is used for both *lip* (of the mouth) and *vulval labia* (Jau-9).

Shifts in meaning from SNOMED-RT® to CT®

In this class of errors, the meaning of specific SNOMED-CT® concepts is changed with respect to the corresponding SNOMED-RT© codes that have the same concept identifier and concept name. Above all, the adoption of [13]'s idea of SEP-triplets (structure-entire-part) led to a large shift in the meanings of nearly all anatomical concepts. One might argue that in RT anatomical terms such as "heart" were never supposed to mean "entire heart", but rather always: "heart or any part thereof"; in CT this distinction has been made explicit.

Many other concepts with the same unique ID in RT and CT also appear to have changed in meaning. A notable example is the concept "45689001: femoral flebography" that in RT only relates to ultrasound, while in CT it involves the use of a contrast medium (Jau-10). The meaning of "leg" has changed. In RT *lower leg* was invariably intended; in CT the situation is unclear. The concept "34939000: amputation of leg" means in RT: "amputation of lower leg" and in CT: "amputation of any part of the lower limb, including complete amputation" (Jau-11). We also observed numerous examples of inconsistent use within CT itself: "119675006: leg repair" refers explicitly to "lower *leg* structure", while "119673004: leg reconstruction" refers explicitly to "lower *limb* structure" (Jau-12). OntologyMatcher was able to identify these problems easily because LinKBase®, thanks to homonym processing and its mappings to UK systems such as OPCS4, is aware of differences between American and British English with respect to the meanings of "leg" and certain other words .

Redundant concepts

The TermModelling algorithm identified 8746 concepts that are the seat of redundancies, that is to say cases where no apparent difference in meaning can be detected between one concept and another. (This is in reality an underestimation because candidate-matching parameters were set very conservatively, sacrificing recall for precision.) These are all pairs or larger pluralities of terms among which differences in meaning could not be identified either conceptually or linguistically. Many of them, we believe, are the result of incomplete or inadequate integration of the Read terms into SNOMED-CT®.

An astonishing example is "210750005: traumatic unilateral amputation of foot with complication", which co-exists in SNOMED-CT® with "63132009: unilateral traumatic amputation of foot with complication". It seems that an incomplete modelling of the latter is at the origin of this mistake (Jau-13). Of the same nature is the co-existence of the concepts "41191003: open fracture of head of femur" and "208539002: open fracture head, femur" (Jau-14), concepts which are modelled entirely differently but in such a way that the technology used in the development of SNOMED-CT® was not able to find the redundancy involved: the former was modeled as directly subsumed by “fracture of femur”, while the latter by “fracture of neck of femur”. Some redundancies become overt only when a larger part of the subsumption hierarchy is examined. Thus, one can question to what extent “172044000: subcutaneous mastectomy for gynecomastia” is different from its immediate subsumer “59620004: mastectomy for gynecomastia” when the latter is itself immediately subsumed by “70183006: subcutaneous mastectomy” (Jau-15).

Table 1: Number of generated intermediate concepts per SNOMED-CT® category

SNOMED CT Concept	original number	number added	% added
ORGANISM	24768	221	0.89
PHYSICAL OBJECT	3336	69	2.07
SPECIAL CONCEPT	130	0	0.00
CONTEXT-DEPENDENT CATEGORIES	6172	233	3.78
OBSERVABLE ENTITY	6430	33	0.51
PHYSICAL FORCE	199	3	1.51
SOCIAL CONTEXT	5120	191	3.73
SPECIMEN	936	148	15.81
EVENTS	75	0	0.00
ENVIRONMENTS AND GEOGRAPHICAL LOCATIONS	1631	5	0.31
STAGING AND SCALES	1118	0	0.00
PROCEDURE	50107	4339	8.66
BODY STRUCTURE	30737	2817	9.16
PHARMACEUTICAL / BIOLOGIC PRODUCT	13623	751	5.51
FINDING	39105	2349	6.01
ATTRIBUTE	975	2	0.21
SUBSTANCE	22062	599	2.72
DISEASE	70286	5688	8.09
QUALIFIER VALUE	7963	103	1.29
Total	284773	17551	6.16

Mistakes due to problematic ontological theory

Lack of sound mereology for anatomy

It is difficult to imagine that an object can be a proper part of two regions that are mereologically disconnected. Despite this, “45684006: structure of tibial nerve” is directly subsumed by both “thigh part” and “lower leg structure”, which explicitly refer to the upper and lower parts of the lower limb, respectively (Jau-16).

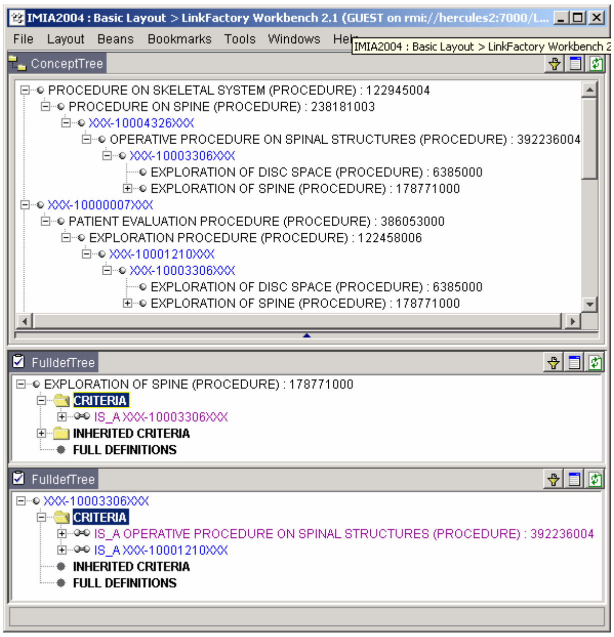
Omission of seemingly obvious relationships

Certainly no large terminology can be expected to be complete. However, one can wonder why the concept “248182008: cracked lips” is a “finding of appearance of lip” but “80281008: cleft lip” is a “disease” and has no relation to “finding of appearance of lip”. Such omissions have the consequence that many sound inferences cannot be made. As another example: “181452004: entire uterus” is part-of “362235009: entire female internal genitalia”, which itself is part-of “362236005: entire female genitourinary system”. This means, however, that SNOMED-CT® does not allow the inference to “uterus” part-of “female genital tract”, nor will it allow inferences to the effect that pregnancy involves the uterus. (Jau-17).

Problematic modeling revealed by LinkFactory’s classification algorithm

Table 1 shows the number of generated pre-coordinations using the LinKFactory®-classifier algorithm under the most conservative setting of minimal generation [12].

6,352 of the 17,551 newly generated pre-coordinations appear to be parents of concepts that they exclusively subsume, a phenomenon that, as we pointed out, is suggestive of mistakes in the neighbourhood of the concept in question. An example is shown in Fig 1, where one would expect the concept “exploration of disk space” to be subsumed by “exploration of spine”.


IMIA2004 : Basic Layout > LinkFactory Workbench 2.1 (GUEST on rmi://hercules2:7000/L...
File
Layout
Beans
Bookmarks
Tools
Windows
Hel
IMIA2004 : Basic Layout > LinkFactory Workbench 2
ConceptTree
FulldefTree
FulldefTree

```
⊟ PROCEDURE ON SKELETAL SYSTEM (PROCEDURE) : 122945004
   ⊟ PROCEDURE ON SPINE (PROCEDURE) : 238181003
      ⊟ XXX-10004326XXX
         ⊟ OPERATIVE PROCEDURE ON SPINAL STRUCTURES (PROCEDURE) : 392236004
            ⊟ XXX-10003306XXX
                 EXPLORATION OF DISC SPACE (PROCEDURE) : 6385000
               ⊞ EXPLORATION OF SPINE (PROCEDURE) : 178771000
⊟ XXX-10000007XXX
   ⊟ PATIENT EVALUATION PROCEDURE (PROCEDURE) : 386053000
      ⊟ EXPLORATION PROCEDURE (PROCEDURE) : 122458006
         ⊟ XXX-10001210XXX
            ⊟ XXX-10003306XXX
                 EXPLORATION OF DISC SPACE (PROCEDURE) : 6385000
               ⊞ EXPLORATION OF SPINE (PROCEDURE) : 178771000
```

```
⊟ EXPLORATION OF SPINE (PROCEDURE) : 178771000
   ⊟ CRITERIA
      ⊞ IS_A XXX-10003306XXX
   ⊞ INHERITED CRITERIA
   ● FULL DEFINITIONS
```

```
⊟ XXX-10003306XXX
   ⊟ CRITERIA
      ⊞ IS_A OPERATIVE PROCEDURE ON SPINAL STRUCTURES (PROCEDURE) : 392236004
      ⊞ IS_A XXX-10001210XXX
   ● INHERITED CRITERIA
   ● FULL DEFINITIONS
```

Figure 1: Algorithmically generated pre-coordinations (marked XXX) as indicators for erroneous modelling in SNOMED-CT®.

Discussion

SNOMED-CT®’s technical reference [14] describes the QA process used for developing SNOMED-CT®. Both manual and automated procedures play a role. As any medical terminology is, by definition, a constantly evolving entity, it is not reasonable to expect perfection or completion during any given release. Through the application of algorithms such as those described above, however, perfection seems that much more approachable. We noticed quality improvements in the July versus January version, as the examples Jau-7 and Ja-6 demonstrate: the wrong subsumption relation with “308758008: collapse of lumbar vertebra” has been removed, though the basic human-introduced mistake was not corrected.

The general moral of this paper is that certain families of mistakes could be prevented by using stronger logical and ontological theories implemented in powerful ontology authoring tools. Imposing restrictions to the effect that entities of disjoint top-level categories should not stand in subsumption relations would prevent mistakes like Jau-1 and Jau-2. Enforcement of mereotopological relations in accordance with an RCC-type system would prevent Jau-4 and Jau-16 and lead to the flagging of cases like Jau-8 and Jau-9 for possible error. Enforcement of logical relations would prevent cases like Jau-3.

Features such as these are the main difference between systems such as SNOMED-CT® and LinkBase®. LinKBase® incorporates strict ontological distinctions, for example between continuant and occurrent entities (i.e. between those entities, such as objects, conditions, functions – which continue to exist identically through time – and those entities, such as processes and events – which unfold themselves in successive temporal phases). When *procedures* are classified as *instruments* or as *diseases* then this reflects a conflation of high-level ontological categories that an adequate terminology system should have ways to prevent automatically. LinKBase® also incorporates formal-ontological theories of mereology and topology (theories of completeness and incompleteness, separation and connectedness, fiat and bona fide boundaries, etc.), and of other basic ontological notions in whose terms relations (link types) between general concepts can be rigorously defined. The presence of such theories results in a more accurate treatment of foundational relations such as is-a and part-of than is possible when such relations are left formally unanalyzed. Finally, it incorporates a clear opposition between *ontological* notions such as object, process, organism function, and *epistemological* notions such as concept, finding, test result, etc.

As is argued in [15] the resultant approach can be used as the basis for more rigorous but also more intuitive and thus more reliably applicable principles of manual curation than those employed in systems like SNOMED-CT® thus far.

Conclusion

Without doubt, a tremendous effort went into developing SNOMED-CT®. It is exciting to consider how much more valuable this effort would become if tools such as those described within were applied.

The LinKBase® ontology is also the result of tremendous ongoing effort, attributed to the ways the LinKBase® and LinKFactory® systems have been designed and built. L&C clearlyhas a powerful tool to detect inconsistencies not only in external systems but also in its own ontology.

References

[1] Smith B, Ceusters W. Towards industrial strength philosophy: how analytical ontology can help medical informatics. Interdisciplinary Science Reviews, 2003; 28: 106-11.

[2] Ceusters W, Martens P, Dhaen C, Terzic B. LinKBase: an Advanced Formal Ontology Management System. Interactive Tools for Knowledge Capture Workshop, KCAP-2001, October 2001, Victoria B.C., Canada (http://sern.ucalgary.ca/ksi/K-CAP/K-CAP2001/).

[3] Montyne F. The importance of formal ontologies: a case study in occupational health. OES-SEO2001 International Workshop on Open Enterprise Solutions: Systems, Experiences, and Organizations, Rome, September 2001 (http://cersi.luiss.it/oesseo2001/papers/28.pdf).

[4] Smith B. Mereotopology: a theory of parts and boundaries, Data and Knowledge Engineering 1996; 20: 287-301.

[5] Smith B, Varzi AC. Fiat and bona fide boundaries, Proc COSIT-97, Berlin: Springer. 1997: 103-119.

[6] Buekens F, Ceusters W, De Moor G. The explanatory role of events in causal and temporal reasoning in medicine. Met Inform Med 1993; 32: 274-278.

[7] Ceusters W, Buekens F, De Moor G, Waagmeester A. The distinction between linguistic and conceptual semantics in medical terminology and its implications for NLP-based knowledge acquisition. Met Inform Med 1998; 37(4/5): 327-33.

[8] Bateman JA. Ontology construction and natural language. Proc Int Workshop on Formal Ontology. Padua, Italy, 1993: 83-93.

[9] Bittner T, Smith B. A theory of granular partitions. *Foundations of Geographic Information Science*, Duckham M, Goodchild MF and Worboys MF, eds., London: Taylor & Francis Books, 2003: 117-151.

[10] Grenon P, Smith B. SNAP and SPAN: Towards dynamic spatial ontology. Spatial Cognition and Computation. In press.

[11] Ceusters W, Smith B. Ontology and medical terminology: Why descriptions logics are not enough. TEPR 2003 (electronic publication): http://ontology.buffalo.edu/medo/TEPR2003.pdf

[12] Dhaen C, Ceusters W. A novel algorithm for subsumption and classification in formal ontologies. (forthcoming).

[13] Hahn U, Schulz S, Romacker M: *Part-whole reasoning: a case study in medical ontology engineering*. IEEE Intelligent Systems & Their Applications vol 14 nr 5, 1999: 59-67.

[14] College of American Pathologists. Snomed Clinical Terms® Technical Reference Guide, July 2003 release.

[15] Smith B, Rosse C. The role of foundational relations in the alignment of biomedical ontologies. Proc. Medinfo, 2004.

Address for correspondence

Dr. Werner Ceusters, VP Research, Language & Computing nv, Het Moorhof, Hazenakkerstraat 20a, B9520-Zonnegem, Belgium. E-mail: werner@landc.be, URL: www.landcglobal.com.