KR-MED 2006 "Biomedical Ontology in Action"
November 8, 2006, Baltimore, Maryland, USA

LinKBase®, a Philosophically-inspired Ontology for NLP/NLU Applications

Maria van Gurp, PhD, Manuel Decoene, MD, Marnix Holvoet and Mariana Casella dos Santos, MD.
Language and Computing NV, Sint-Denijs-Westrem, Belgium
Tel: +32-(0)9-2808400; http://www.landc.be; {marjan, mariana}@landc.be

ABSTRACT

LinKBase® is a biomedical ontology. Its hierarchical structure, coverage, use of operational, formal and linguistic relationships, combined with its underlying language technology, make it an excellent ontology to support Natural Language Processing and Understanding (NLP/NLU) and data integration applications. In this paper we will describe the structure and coverage of LinKBase®. In addition, we will discuss the editing of LinKBase® and how domain experts are guided by specific editing rules to ensure modeling quality and consistency. Finally, we compare the structure of LinKBase® to the structure of third party terminologies and ontologies and discuss the integration of these data sources into LinKBase®.

INTRODUCTION TO LINKBASE®

To achieve full benefit of health information technology, a health information network, enabling interoperability across different facilities and countries, is essential. However, different and diverse medical information systems hamper the process of data sharing. One solution to this problem is to use a central ontology, with a strict hierarchical structure and a consistent semantic network of relationships between its types that can support NLP/NLU and data integration applications and that can serve as the link between the different medical information sources and systems. LinKBase® is such an ontology. LinKBase® has been designed with the main goal of integrating terminologies and databases with applications designed for NLP and information management and retrieval and has been built up from the ground over the past 10 years. It covers various aspects of medicine, including procedures, anatomy, pharmaceuticals and various disorders and anomalies delivering over 9 million knowledge elements making it the largest biomedical knowledge base in the world. The core ontological elements, being its types and relationships, have no embedded grammatical information and are as such language independent, but they are cross-referenced to terms and lexemes in various languages. Several features make LinKBase® the preferred ontology to eliminate some of the barriers to creating health information organizations; 1) LinKBase® is a language and application independent ontology 2) LinKBase® is integrated to and under the guidance of a formal upper level framework Basic Formal Ontology (BFO)[1], 3) LinKBase® embedded the linguistic ontology framework Generalized Upper Model (GUM)[2], 4) the types within LinKBase® are interconnected by a rich set of hierarchical relationship types, 5) LinKBase® unambiguity is supported by full definitions and 6) the LinKBase® ontology is connected to a lexicon of terms in various languages.

Inherent to the interoperability of medical information systems, is the integration of the medical data within those systems. This task turns out to be a complex endeavor, not least because the different terminologies or databases that are to be integrated are often internally and mutually inconsistent. In this respect, LinKBase® can serve as a 'translation hub' between diverse third party terminologies, based on the fact that all these terminologies essentially speak about the same reality. This makes it possible to integrate them on the basis of a sound understanding of those basic categorical distinctions that are common to them all.

STRUCTURE OF LINKBASE®

To achieve a coherent framework, able to support reasoning applications, NLP and NLU, the LinKBase® ontology is founded on philosophical and linguistic theories.

BFO[1], a philosophically inspired upper-level ontology that focuses on the entities in reality at different levels of granularity and not on the human conceptualization of this reality, was used to structure the upper level of LinKBase®. Theories of endurants and perdurants[3], mereology, topology, universals and particulars, biological classes and instantiations[4], space and time and granular partitions[5] are all included in the BFO theory. The main distinction in BFO is between the endurants (SNAP) and perdurants (SPAN). Endurants are those entities that endure through time, in contrast to perdurants, which unfold themselves through time and are never fully present at a given moment in time[3]. The LinKBase® hierarchy is integrated under and branches from the BFO upper level entities, representing general

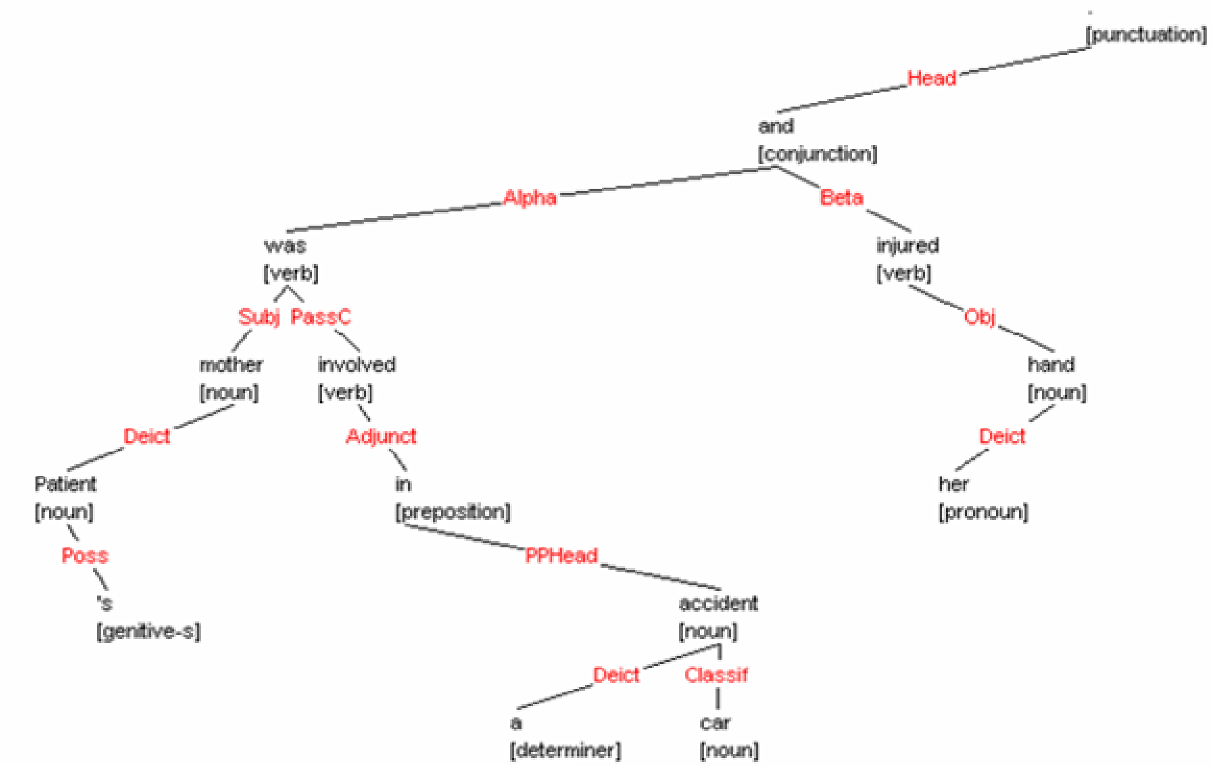

Figure 1 – Analysis of syntactical structure
Syntactic analysis of the sentence "*The patient's mother was involved in a car accident and injured her hand*".

categories such as processes, properties and objects. By using the BFO theory[1], LinKBase® is not only provided with a rigorous philosophical classification of all its entities, but is provided with the set of axioms that govern BFO entities and the relationships among them as well. These axioms are used to apply modeling restrictions and guidance to prevent erroneous editing and to maintain and improve the structure of LinKBase®. More important however, the BFO definitions of ontological entities can be used by reasoning applications, including applications designed for NLP, and aid to the filtering out of erroneous synonyms and the disambiguation of ontological structures that are inherent to the processing of free text[6]. To support correct and precise linguistic reasoning, the LinKBase® hierarchical structure is very strict and every child type is a subclass of its parent's class. Thus, the application of BFO-driven philosophical knowledge and axioms offers several advantages that are not present in application ontologies lacking a philosophical backbone.

The structure of the LinKBase® mid-layer is partially structured according to the GUM[2]. The GUM is a general task and domain independent linguistically motivated ontology intended for organizing information for expression in natural language. In LinKBase®, the "processes" are organized based on their linguistic properties. This allows us, by using a GUM-based grammatical analysis, to convert the syntactic structure of a given sentence into an 'understandable' structure of types and criteria. For example, we can determine that the actee (or object) and the actor (or subject) are identical in the sentences "The patient was treated by the doctor" and "The doctor treated the patient." In the sentence "The patient's mother was involved in a car accident and injured her hand", we deduce that "injured her hand" refers to the mother and is not referring to the patient (figure 1). In addition, we use the semantics to disambiguate the syntax by relating specific processes to specific actors and actees, e.g. a "treatment process" is related to the actee "patient" and the actor "healthcare professional". Using this strategy, LinKBase® has the capacity to support NLU applications.

TYPES

The more than 570,000 LinKBase® types represent real-world entities and not concepts in the mind of conscious beings that are abstractions of what these beings think the real-world entities are. To enable semantic reasoning, the types are hierarchically structured using a realist approach: child types represent subclasses of a given parent for 100 % of the instances (figure 2A). Using this approach, the hierarchical relationships among LinKBase® types have a consistent meaning. In LinKBase®, for example, “rash” will never be a subclass of “allergic reaction” since it is not always allergic. However, in many classification systems that lack a strict hierarchical structure, such as ICD-9[7] or MEDCIN[8], these situations do occur, hampering the use of algorithms in reasoning. Conflicts arise when analyzing the sentence “the patient was diagnosed with meningitis that was not due to infection” using an ontology in which “meningitis” is modeled as “a-consequence-of infection”. Following the realist approach, the creation of an additional subclass of “meningitis”, namely “infective meningitis”, forms a solution to this problem. “Infective meningitis is-a meningitis” ànd “a-consequence-of infection”. However, “aseptic meningitis”, the illness of the above mentioned patient, is ‘only’ “meningitis” and does not have a relationship, direct or inherited, to “infection”. Thus, the principle of 100 % criteria allows LinKBase® to support NLU applications where other ontologies fail.

LinKBase® is a “living” ontology and types and subsequent relationships are added and edited on a daily basis by the modeling team. Although it is not required for types to be perfectly modeled from the beginning, the creation of new types and subsequent relationships is strictly regulated and new types can only be added if specific criteria are met[9].

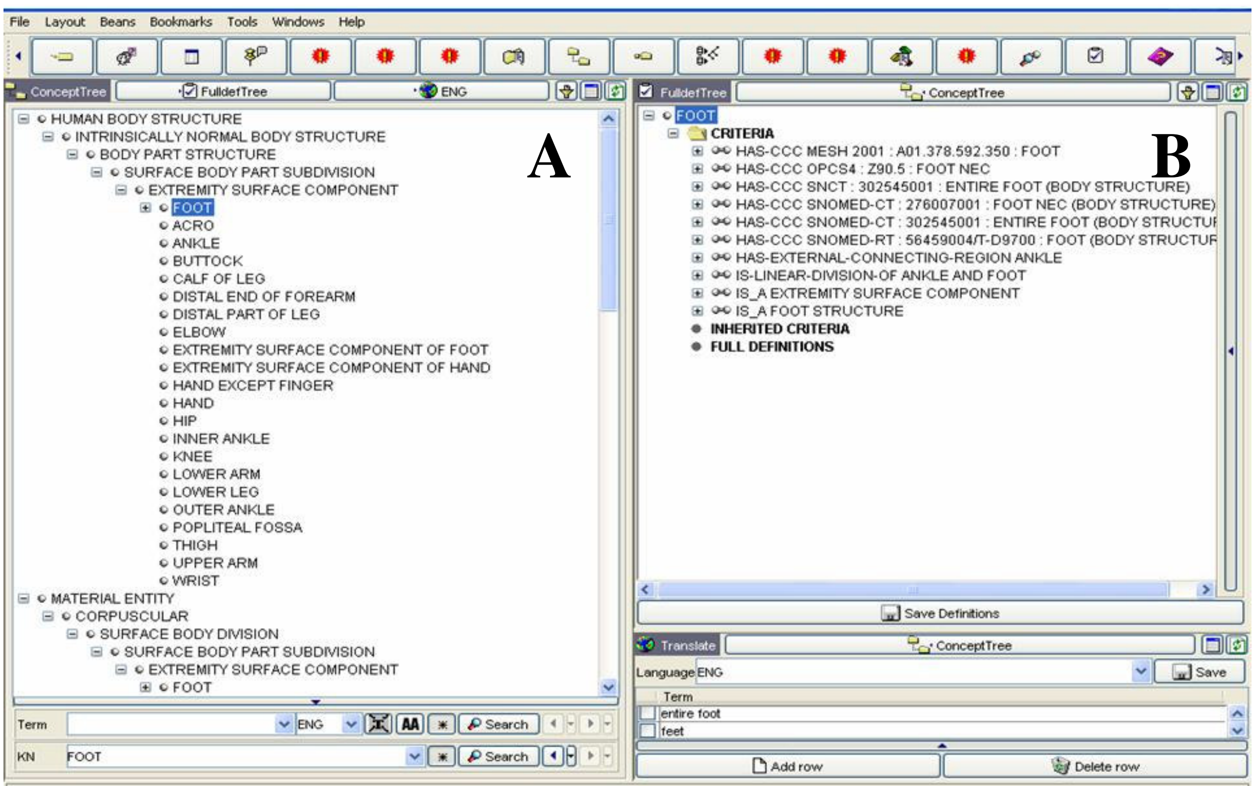

Figure 2 - LinKBase® structure
Screenshot of LinKBase® structure showing:
A) hierarchical structure; all child types are a representation of their parent(s), B) several types of relationships to 3rd party terminologies (the HAS-CCC relationships) as well as to other LinKBase® types and C) the terms that are assigned to, in this example, the type FOOT.

RELATIONSHIPS

The types in LinKBase® are linked into a semantic network by a rich set of relationship types (figure 2B). Most relationships are based on theories, including BFO[1], that deal with topics such as mereology and topology[10,11], time and causality[12] and

models for semantics driven natural language understanding[13,14]. In addition, LinKBase® contains relationships that fall out of any theory but are essential to express important notions in the medical world. One example is "absence of entity", considered a lack of entity and not a real entity in most theories, but needed to represent types such as "anuria", "absence of blood" or "noninvasive". Since it is not possible to consider absences as processes[6], absences are represented as a relation between the "absent entity" and the "entity from which the related entity is absent". This avoids the creation of "absent processes" and keeps the distance between the related types to a minimum, which is relevant to many LinKBase applications[15].
The LinKBase® relationship types are structured in a multi-parented hierarchy, taken into account both the formal realistic ontological implications and the linguistic aspects of the relationships. LinKBase® contains 383 different relationship types, covering many, often subtle, semantic differences; including spatial, temporal and process-related relationship types. New relationship types are added when the existing relationships are not capable to represent the semantics of new types or when new insights justify the creation of new relationship types that might provide a better quality assurance or are needed for certain applications. Within LinKBase®, we are currently revising the framework of "function" and "dysfunction". New relationship types are needed to relate, for example, "function", the function that the body part is supposed to perform, with "functioning process", the body process that it is really performing at a given point in time. For this purpose, the relationship type "has-realisation" was created, going from "function" to the "functioning process". The reverse relationship type is "is-realisation".

Within LinKBase®, formal or full definitions are created by those criteria, whether direct or inherited, that are necessary and sufficient to uniquely define the type (figure 3).

The formal logic used by LinKBase® is an important prerequisite for an ontology with the ability to support reasoning applications[16], since the system automatically infers that, if a real-world entity satisfies the full definition of a given domain-entity, it is an instance of that domain-entity.

Only around 15 % of the total number of relationships within LinKBase® is covered by formal subsumption relationships. As a consequence, the structure of LinKBase® is much richer compared to

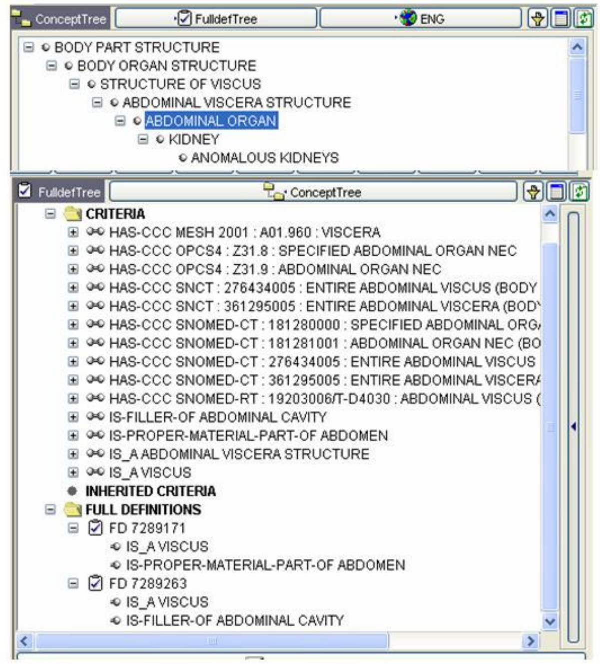

Figure 3 - Formal or full definitions in LinKBase® Within LinKBase®, formal or full definitions are created by those criteria, which are necessary and sufficient to uniquely define the type. In this example, two full definitions are defined for the type ABDOMINAL ORGAN.

other ontologies and terminology systems, in which type-relationships are often expressed as "narrower" or "broader", as is the case for the Unified Medical Language System (UMLS)[16].

TERMS

The LinKBase® ontology is connected to a lexicon of approximately 1.5 million terms. Terms are signs or symbols that are used to represent types in the real world. Terms can be synonyms, symbols, translations or, for example, singular or plural forms of the type name (figure 2C). In LinKBase®, the assignment of terms depends on the meaning of the types. Terms can only be assigned to types when they express exactly the same meaning in natural language. Bad synonym assignment often occurs because conditions are tightly connected in a medical cause-effect or symptom-disorder relation, as is the case for the SNOMED[19, 20] type "viral gastroenteritis (disorder)" that is linked to the terms "viral diarrhea", "viral vomiting" and "viral gastroenteritis". Although this example of SNOMED term assignment might be correct from a medical point of view and is suited for terminology standardization/coding applications it is not compatible with NLP/NLU applications and is thus avoided in LinKBase®.

Next to types, criteria and relationships can receive terms as well; the criterion "has-happening-earlier-than systole" has the term presystolic and the relationship "is-part-of" has the German term "ist-ein-Teil". Unlike types and relationships, terms can be stored in different languages. Thus, although LinKBase® itself is language independent, the assignment of multi-lingual terms and lexemes to its ontological elements allow the analysis of text in any European language.

EDITING/MODELING PROCESSES

An accurate and consistent modeling is not always obvious when dealing with a large and complex ontology as LinKBase®. To overcome this problem and to guide and assist the modelers, several mechanisms have been developed. These tools include management issues, such as hierarchical user privileges and log file reviews, and modeling guidance in which the BFO theory[1] is used as automatic error detection. Both the BFO subsumption and disjoint axioms were implemented in LinKBase®. Of the BFO relationship axioms, only the domain-range restrictions were used. The axioms on the level of inference are not applied, but future work involves the application of these and other BFO axioms, to allow for further levels of inference. In addition, the BFO framework and the BFO partition theory are used as guidelines for the modelers to follow.

hierarchical user privileges

Hierarchical user privileges is a mechanism that assigns types to the modeler that created them. The users are organized in a hierarchical structure according to their skills and experience. Elements can only be modified by the ontologist who created the item or by a user at a higher level in the hierarchy. In this way, erroneous modeling of an already correctly modeled type is prevented as well as repetitive modeling of a certain type by different modelers.

log files

Every action performed by a modeler is stored in a log file. In the case of erroneous modeling, one can go back to the log files and check what went wrong, in order to be able to correct their mistake(s). In addition, the log files can be used for training purposes, in which the work of an ontologist is reviewed by an experienced ontologist and the performed actions are discussed.

relationship type domain-range restrictions

One method enforced by LinkFactory®[21], the ontology management system used to edit, store and maintain LinKBase®, in order to limit the amount of modeling errors is domain-range restriction. A domain-range restriction on a relationship type limits the amount of types to which the relationship can refer, since that specific relationship type can only relate types that are located within its domain. These domain-range attributes have values corresponding to the SNAP and SPAN entities of BFO[3] between which they apply. In addition, the embedded GUM theory[2] and the linguistically structured processes allow the further refinement of domain-range restrictions to the mid-layer and linguistic layer of LinKBase® as well. For example, the relationship type "has-theme" holds between an endurant and a motion process and the theme is the entity that is displaced in the motion process (e.g. "excision of kidney has-theme kidney"). The source of the relationship type "has-actee", an actee is someone or something that passively undergoes, is changed by, or is directly affected by a predicate, is always a-kind-of the linguistic process "directed action" (e.g. "treatment of acne has-actee acne"). Since both the relationship types and the types within LinKBase® are hierarchically structured, the relationship type domain-range restriction applies to the subtypes of the relationship type and type(s) in question as well. The relationship type "has-theme", is a further refinement of the "has-participant" relationship type, valid between processes and endurants, of which it is a subclass.

If a modeler tries to link a type to another type that is not within the domain of the specific relationship type used, the modeler receives a warning that a restriction is violated and has to revise his modeling.

disjoint restrictions

Another method enforced by LinkFactory®[21] to avoid modeling errors and to enhance the quality of LinKBase® is disjoint restriction. When two types are made disjoint, this implies that no type can be a subclass of both disjoint types. These checks are performed in real-time and the modeler receives a disjoint violation warning whenever he wants to make a type a subclass of both disjoint types. In addition, when (re)structuring the ontology, disjoint violations support the creation of a valid model of reality. Examples of disjoints in LinKBase® are the endurants (SNAP) and perdurants (SPAN) and the categories Corpuscular (e.g. organisms and organs) and Non-Corpuscular (e.g. tissues and liquids).

THE META- AND DOMAIN-MAPPING FRAMEWORK

In LinKBase®, the domain-entity is defined as the set of types and their relationships that always have a consistent meaning. Outside this domain, in an area called the meta-entity, the 3rd party terminologies are located, standard classification systems such as ICD9[7] and SNOMED[19, 20]. The external ontologies are stored in their exact original style and structure and are linked to the LinKBase® domain-entity by specific formal relationship types[22]. This framework of a central domain-ontology linked to external (medical) information sources is called the "meta- and domain-mapping framework". Table 1 contains an overview of some of the most important 3rd party terminologies that are linked to LinKBase®.

Table 1 - Absolute number of meta-entity type names appended and subsequently processed within LinkBase®.

Meta-entity versus LinkBase®	METN[a] appended	LBKN[b] in support to METN	LBKN defined	English terms for LBKN	Non-English terms for LBKN
ICD-9-CM	27659	50105	20406	138186	223110
CPT-4	10202	8366	404	11121	2806
ICPC-2	746	3491	2297	14978	28823
MEDCIN	208981	208400	5450	243218	61432
MEDDRA 6.0	60616	51572	10401	111042	110221
MESH 2001	35780	19984	2997	54960	71349
SNCT	275222	322360	25121	539031	308786

[a] *Meta-entity type name*
[b] *LinkBase® knowledge name*

The "meta- and domain-mapping framework" has several advantages compared to a direct integration of external ontologies, such as the reusability of existing mappings, the ability to cross map several data sources and the ability to transpose divergent levels of granularity between external information sources. However, it also requires a careful mapping procedure to the central domain ontology LinKBase®, since the different information sources often have internally and mutually inconsistent structures[22]. Through the implementation of the "meta- and domain-mapping framework" LinKBase® becomes the ontology of choice to serve as a "translation-hub" between diverse 3rd party terminologies. Indeed, other ontologies that integrate several different 3rd party terminologies do exist, such as the UMLS[16]. Why then, do we claim that LinKBase® is the preferred ontology for data integration? Is the UMLS®[16], for this application, not a useful source? A comparison between LinKBase® and the UMLS®[16] will shed a light on the differences in structure and potential applications.

LINKBASE® VERSUS THE UMLS®

Within the Metathesaurus of the UMLS®[16], a large number of different source vocabularies and classification systems, e.g. ICD9[7], Meddra[23] and SNOMED[19, 20], are integrated with the purpose to facilitate the development of NLP/NLU computer systems and to overcome disparities in language, granularity and perspective. When integrating different vocabularies, it is important to respect the original structure and granularity of the source vocabularies. If not, circular hierarchical relationships might occur, as has been described in Bodenreider[24]. For example, in the UMLS® Metathesaurus, "maduromycosis" is related to "mycetoma of foot" in one vocabulary and to "eumycotic mycetoma" in another one. In LinKBase®, however, "eumycotic mycetoma" (mycetoma caused by fungi) and "mycetoma of foot" are child types of "mycetoma" (synonym of maduromycosis). The types are modeled according to their meaning and linked to their respective information sources, thus keeping a consistent and realistic view of the world (see figure 4).

A second distinction between the UMLS®[16] and LinKBase® are the relationship types and more specific the hierarchy within. Whereas LinKBase® follows a realist approach resulting in relationship types with a consistent meaning and child types that represent subclasses of a given parent for 100 % of the instances, this is not the case for the UMLS®. The hierarchical relationship types of the UMLS® can be both parent-child relationship types, comparable to the ones used in LinKBase®, or broader/narrower-than relationship types. An example of the latter is "toe is-a foot". Although a toe is part of the foot, it certainly is not a kind-of foot and hence should not be placed as a subclass of "foot".

Within LinKBase®, this problem is solved by creation of the type "foot structure" with the subclasses "foot", referring to the extremity foot, and "foot part". "Foot part", in turn, contains the subclasses "toe part" and "toe", which refers to the digit toe[25] (figure 5). This consistent class-subclass hierarchy of LinKBase® is a huge asset compared to the UMLS ® hierarchy when considering NLP/NLU applications, since it avoids misclassification and allows clear and correct crossmapping.

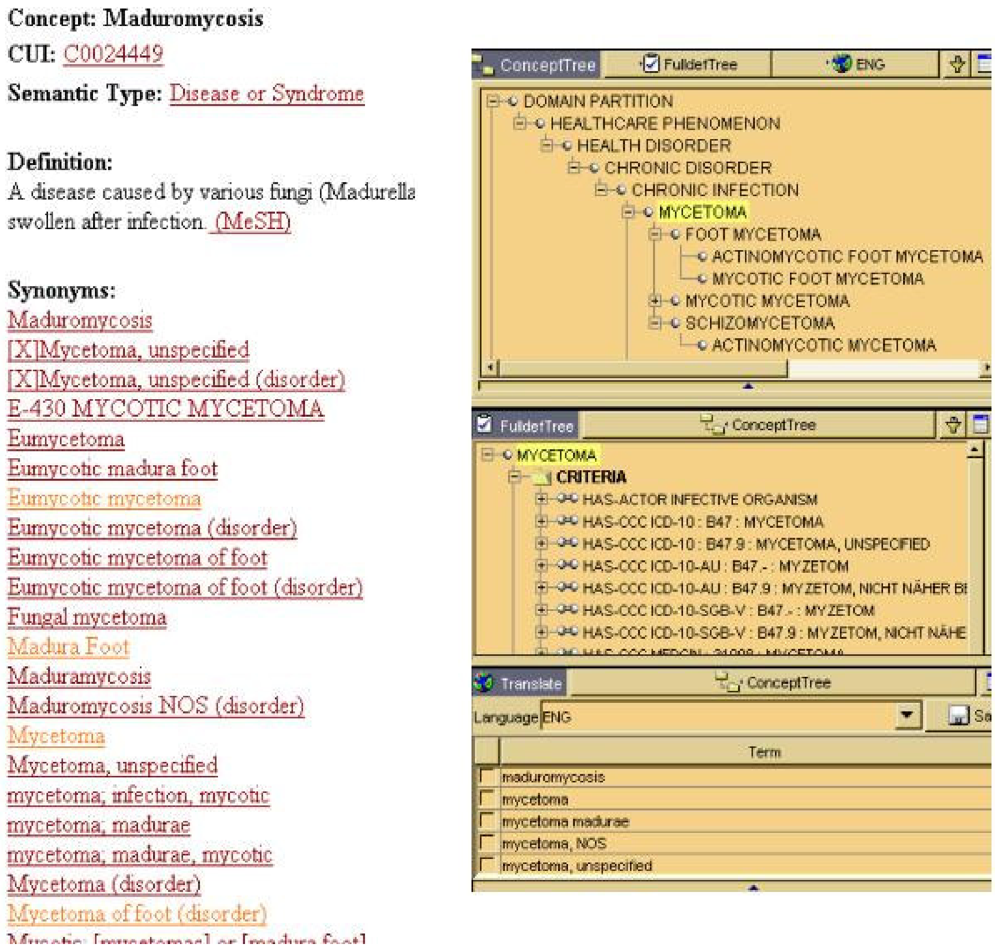

Figure – 4 Comparison between LinKBase® (right panel) and the UMLS (left panel, see text for details)

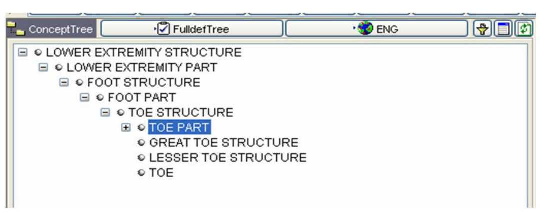

Figure 5 - LinKBase® class-subclass structure

When comparing LinKBase® to the UMLS®, we can conclude that LinKBase® is more suited for NLP/NLU applications. Conflicting relationships and the lack of a consistent hierarchy makes the mapping of free text to UMLS® a highly error-prone task. An example of a LinKBase®-based NLP/NLU application is the development of an information extraction application for extraction of findings and procedures and their related context information, encoded into SNOMED according to the SNOMED Context Model guidelines[26]. Another example of a LinKBase®-based NLP/NLU application is the extraction of patient-related suicide- and self-harm behavior from medical reports that were generated during clinical trials. This aim of this project was to enhance data retrieval and to decrease manual review. In a first pilot study, based on 153 documents, the accuracy was more than 99 % (based on precision and recall against manual annotations).

CONCLUSION

The novelty of LinKBase® compared to other terminologies is the LinKBase® "meta- and domain-mapping framework". This framework of 3rd party terminologies, linked to the LinKBase® domain-entity, makes exchange, management and integration of data possible. The application-independency of LinKBase®, its strong framework based on established ontological theories, combined with a rich set of hierarchical relationship types, without any doubt, creates a flexible yet powerful ontology.

References

1. B. Smith, Basic formal Ontology (BFO) http://ontology.buffalo.edu/bfo/

2. J.A. Bateman, R. Henschel, and F. Rinaldi. The generalized upper model 2.0. Technical report, GMD/Institut für Integrierte Publikations- und Informationssysteme, Darmstadt, Germany, 1995. http://www.gmd.darmstadt.de/publish/komet/gen-um/newUM.html

3. P. Grenon, B. Smith, SNAP and SPAN: Prolegomena to Geodynamic Ontology, Hornsby and Worboys Spatial Cognition and Computation, forthcoming, http://www.ontology.buffalo.edu/smith/courses03/tb/SNAP_SPAN.pdf

4. B. Smith, Westerstahl (ed), Invited Papers from the 10th Int Conf in Logic Methodology and the Philosophy of Science Oviedo, http:// ontology. buffalo.edu/bio/logic_of_classes.pdf, 2003

5. Smith B. The Logic of Biological Classification and the Foundations of Biomedical Ontology. Westerstahl (ed), Invited Papers from the 10th International Conference in Logic Methodology and the Philosophy of Science, Oviedo, Spain, 2003.

6. M. Casella dos Santos, C., Dhaen, and J.M. Fielding, Philosophical scrutiny for Run-Time support of Application Ontologies Development, International Conference on Formal Ontology in Information Systems (FOIS), Torino, Italy, 2001.

7. International Classification of Diseases, Ninth Revision, with Clinical Modifications, Washington DC: U.S. National Center for Health Statistics, 1980

8. MEDCIN, http://www.medicomp.com/.

9. A. Flett, M. Casella dos Santos, W. Ceusters, Some Ontology Engineering processes and their Supporting technologies. Gomez-Perez, A, Benjamins VR (eds), Springer, 2002: 154-165.

10. B. Smith, A. C. Varzi,, Fiat and Bona Fide Boundaries, Proc COSIT-97 Springer-Verlag, 1997: 103-119.

11. B. Smith, Data and Knowledge Engineering 20, http://ontology.buffalo.edu/smith/articles/mereotopology.htm, 1996.

12. F. Buekens, W. Ceusters, G. De Moor, The Explanatory Role of Events in Causal and Temporal reasoning in Medicine, Met Inform Med 32, 1993: 274-278.

13. W. Ceusters, F. Buekens, T. Deray, A. Waagmeester, The distinction between linguistic and conceptual semantics in medical terminology and its implications for NLP-based knowledge acquisition, Met Inform Med 37, 1998: 327-333.

14. J. Bateman, Ontology construction and natural language. Proc International Workshop on Formal Ontology. Padua(Italy), 1993: 83-93.

15. W. Ceusters, B. Smith and J. Fielding. LinkSuite™: formally robust ontology-based data and information integration,. In DILS (Database Integration in the Life Sciences), Berlin: Springer, 2004.

16. B.L. Humphreys and D.A. Lindberg, The UMLS project: making the conceptual connection between users and the information they need, Bull Med Libr Assoc. 1993; 81(2): 170–177,

17. T.C. Rindflesch, B. Libbus, D. Hristovski, A.R. Aronson, and H. Kilicoglu, Semantic relations asserting the etiology of genetic diseases, AMIA Symposium Proceedings, 2003: 554-558.

18. G. Schadow and C.J. McDonald, Extracting structured information from free text pathology reports, AMIA Symposium Proceedings, 2003: 584-588.

19. R.A. Côté and S, Robboy, Progress in medical information management: Systematized Nomenclature of Medicine (SNOMED), JAMA, 1980, 243; 756-762

20. K.A. Spackman, K.E. Campbell, R.A. Cote, SNOMED RT: a reference terminology for healthcare, proc AMIA Annu Fall Symp, 1997; 640-644

21. W. Ceusters, P. Martens, C. Dhaen, B. Terzic, Linkfactory: an advanced formal ontology management system, Victoria, Canada K-CAP 2001, http://www .landcglobal.com/images/linkfactory.pdf, 2001.

22. M. Casella dos Santos, C. Dhaen, D. Decraene, M. van Gurp, T. Deray, The methodology behind the military health system conceptual framework and core ontology, http://www.landcglobal.com/images/TSB_Methodology.pdf, 2005.

23. MEDDRA, http://www.meddramsso.com/MSSOWeb/index.htm.

24. O. Bodenreider, Circular hierarchical relationships in the UMLS: etiology, diagnosis, treatment, complication and prevention, Proc of AMIA Annual Symposium, 2001, 57-61.

25. U. Hahn, S. Schulz, M. Romacker, An ontological engineering methodology for part-whole reasoning in medicine, 1998, http://citeseer.ist.psu.edu/hahn98ontological.htm.

26. HL7 Term Info working group. Using Snomed CT in HL7 Version 3. http://www.hl7.org/