Third International Conference on Knowledge Representation in Medicine

KR-MED 2008

Representing and sharing knowledge using SNOMED

Editors: Ronald Cornet, Kent A. Spackman

Arizona Grand Resort, Phoenix, Arizona, May 31-June 2, 2008

Co-organized by:
Working Group on Formal (Bio-)Medical Knowledge Representation
of the American Medical Informatics Association (AMIA) and the International Health
Terminology Standards Development Organisation (IHTSDO)

Organization

Scientific Program Committee

Chairs:

- Ronald Cornet (Academic Medical Center, Amsterdam, The Netherlands)
- Kent Spackman (International Health Terminology Standards Development Organisation)

- Hans Åhlfeldt (University of Linköping, Sweden)
- Franz Baader (University of Dresden, Germany)
- Robert Baud (University Hospital Geneva, Switzerland)
- Olivier Bodenreider (National Library of Medicine, USA)
- Anita Burgun (University of Rennes, France)
- James Cimino (Columbia University, USA)
- Maureen Donnelly (State University of New York at Buffalo, USA)
- Kathy Giannangelo (Language & Computing, USA)
- Josef Ingenerf (University of Lubeck, Germany)
- Marie-Christine Jaulent (Université P.M. Curie, France)
- Nicolette de Keizer (Academic Medical Center, The Netherlands)
- Yves A. Lussier (University of Chicago, USA)
- David Markwell (Clinical Information Consultancy, UK)
- Onard Mejino (University of Washington, USA)
- Peter Mork (Mitre Corporation, USA)
- Erik van Mulligen (Erasmus Medical Center, The Netherlands)
- Jon Patrick (University of Sydney, Australia)
- Alan Rector (University of Manchester, UK)
- Jean Marie Rodrigues (University of St. Étienne, France)
- Daniel Rubin (Stanford University, USA)
- Patrick Ruch (University Hospital Geneva, Switzerland)
- Barry Smith (State University of New York at Buffalo, USA)
- Lowell Vizenor (National Library of Medicine, USA)

Organizing Committee

- Stefan Schulz (Freiburg University Hospital, Germany)
- Ulrich Andersen (Sorano, Denmark)

Acknowledgements

The help of AMIA staff in putting together this conference is greatly appreciated. Much work has been performed by them. We especially want to thank Tia Abner, Dasha Cohen, and Jason Harbonic for their efforts
We furthermore express our thanks to Peggy LaJoie from SNOMED Terminology Solutions (CAP-STS) for attracting sponsors.
Thanks to Stefan Schlachter (Freiburg University Hospital, Germany), the webmaster of our website: http://www.kr-med.org.

Representing and sharing knowledge using SNOMED

These are the proceedings of KR-MED 2008, the Third International Conference on Formal Biomedical Knowledge Representation, held in Phoenix, Arizona on May 31st – June 2nd, 2008. The conference is co-organized by the Working Group on Formal (Bio-)Medical Knowledge Representation of the American Medical Informatics Association (AMIA) and the International Health Terminology Standards Development Organisation (IHTSDO), and collocated with the 2008 AMIA Spring Congress.

This 3rd edition of KR-MED was inspired by the vision of a universal clinical terminology, covering a broad range of health-related domains and meeting the needs of all health professionals. In the last two decades numerous health informatics research activities have been performed to turn this vision into reality. This conference highlights the progress made in research as well as application, challenges the current state, and reflects on future work from a perspective of knowledge representation and formal ontologies.

KR-MED 2008 focuses on the Systematized Nomenclature of Medicine – Clinical Terms: SNOMED CT®. SNOMED CT is emerging as a comprehensive, multilingual clinical healthcare terminology, which is under a new international ownership since 2007.
Focusing on SNOMED CT, KR-MED 2008 follows up on the successful first Semantic Mining Conference on SNOMED (SMCS) organized in Copenhagen in October 2006.

KR-MED 2008 will bring 2 tutorials, 3 invited speakers, a panel on the state of affairs in member countries, poster presentations and product presentations by some of the sponsors. The scientific sessions address the full spectrum from theory to practice. Theoretical issues include representation, formalization, classification, and the integration of SNOMED CT with information models. Sessions geared toward application cover mapping, subsetting, interface terminologies, and applications for the use of SNOMED CT.

We like to point out that the one-page abstracts in the paper section represent papers that are currently under review for journal submission. Depending on the outcome of the review process, either a reference to the full publication will be added, or the full paper will be included later.

We are confident that these proceedings represent the aim of the conference: sharing knowledge about using SNOMED CT.

Wishing you all an informative and enjoyable conference,

Ronald Cornet, Academic Medical Center, Amsterdam, The Netherlands
Kent Spackman, International Health Terminology Standards Development Organisation
Chairs of the KR-MED 2008 Scientific Program Committee

Table of Contents

Papers

Posters

Representing and sharing knowledge using SNOMED
Proceedings of the 3rd international conference on Knowledge Representation in Medicine (KR-MED 2008)
R. Cornet, K.A. Spackman (Eds)

Debugging SNOMED CT Using Axiom Pinpointing in the Description Logic $\mathcal{EL}^+$

Franz Baader and Boontawee Suntisrivaraporn
Institute for Theoretical Computer Science, TU Dresden, Germany

Abstract

SNOMED CT *is a large-scale medical ontology, which is developed using a variant of the inexpressive Description Logic* $\mathcal{EL}$*. Description Logic reasoning can not only be used to compute subsumption relationships between* SNOMED *concepts, but also to pinpoint the reason why a certain subsumption relationship holds by computing the axioms responsible for this relationship. This helps developers and users of* SNOMED CT *to understand why a given subsumption relationship follows from the ontology, which can be seen as a first step toward removing unwanted subsumption relationships.*

In this paper, we describe a new method for axiom pinpointing in the Description Logic $\mathcal{EL}^+$*, which is based on the computation of so-called reachability-based modules. Our experiments on* SNOMED CT *show that the sets of axioms explaining subsumption are usually quite small, and that our method is fast enough to compute such sets on demand.*

Introduction

Description Logics (DLs) [1] are a family of logic-based knowledge representation formalisms, which can be used to develop ontologies in a formally well-founded way. This is true both for expressive DLs, which are the logical basis of the Web Ontology Language OWL [2], and for inexpressive DLs of the $\mathcal{EL}$ family [3], which are used in the design of large-scale medical ontologies such as SNOMED CT[1] and the National Cancer Institute's ontology.[2]

One of the main advantages of employing a logic-based ontology language is that reasoning services can be used to derive implicit knowledge from the one explicitly represented. DL systems can, for example, classify a given ontology, i.e., compute all the subsumption (subconcept–superconcept) relationships between the concepts defined in the ontology. The advantage of using an inexpressive DL of the $\mathcal{EL}$ family is that classification is tractable, i.e., $\mathcal{EL}$ reasoners such as CEL [4] can compute the subsumption hierarchy of a given ontology in polynomial time.

Similar to writing large programs, building large-scale ontologies is an error-prone endeavor. Classification can help to alert the developer or user of an ontology to the existence of errors. For example, the subsumption relationship between "amputation of finger" and "amputation of upper limb" in SNOMED CT is clearly unintended [6, 7], and thus reveals a modeling error. However, given an unintended subsumption relationship in a large ontology like SNOMED CT with almost four hundred thousand axioms, it is not always easy to find the erroneous axioms responsible for it by hand. To overcome this problem, the DL community has recently invested quite some work on automating this process. Given a subsumption relationship or another questionable consequence, axiom pinpointing computes a minimal subset (all minimal subsets) of the ontology that have this consequence (called MinAs in the following). Most of the work on axiom pinpointing in DLs was concerned with rather expressive DLs (see, e.g., [8, 9, 10]). The only work that concentrated on pinpointing in the $\mathcal{EL}$ family of DLs was [11]. In addition to providing complexity results for pinpointing, [11] introduces a "pragmatic" algorithm for computing one MinA, which is based on a modified version of the classification algorithm used by the CEL reasoner [4]. Though this approach worked quite well for mid-size ontologies (see the experiments on a variant of the GALEN medical ontology described in [11]), it was not efficient enough to deal with large-scale ontologies like SNOMED CT.

[1] http://www.ihtsdo.org/our-standards/
[2] http://www.nci.nih.gov/cancerinfo/terminologyresources

In the present paper, we describe a new method for axiom pinpointing in the Description Logic $\mathcal{EL}^+$, which is based on the computation of so-called reachability-based modules [5]. Our experiments on SNOMED CT show that the sets of axioms explaining a given subsumption are usually quite small (78% of the MinAs we computed were of size ten or less), and that our method is fast enough (on average, it took one second to obtain a MinA) to compute these sets on demand, i.e., whenever the user asks for a MinA for a suspect subsumption relationship.

Axiom pinpointing in $\mathcal{EL}^+$

In this section, we first introduce the DL $\mathcal{EL}^+$, which is an extension of the DL $\mathcal{EL}$ used to define SNOMED CT. Then, we define minimal axiom sets (MinAs) for subsumption, and recall some of the known results about computing MinAs in $\mathcal{EL}^+$.

Syntax	Semantics
$\top$	$\Delta^{\mathcal{I}}$
$C \sqcap D$	$C^{\mathcal{I}} \cap D^{\mathcal{I}}$
$\exists r.C$	$\{x \in \Delta^{\mathcal{I}} \mid \exists y \in \Delta^{\mathcal{I}} : (x,y) \in r^{\mathcal{I}} \wedge y \in C^{\mathcal{I}}\}$
$C \sqsubseteq D$	$C^{\mathcal{I}} \subseteq D^{\mathcal{I}}$
$r_1 \circ \cdots \circ r_n \sqsubseteq s$	$r_1^{\mathcal{I}} \circ \cdots \circ r_n^{\mathcal{I}} \subseteq s^{\mathcal{I}}$

Table 1: Syntax and semantics of $\mathcal{EL}^+$.

Starting with a set of concept names CN and a set of role names RN, $\mathcal{EL}^+$ *concept descriptions* can be built using the constructors shown in the upper part of Table 1, i.e., every concept name $A \in \mathsf{CN}$ and the top concept $\top$ are $\mathcal{EL}^+$ concept descriptions, and if C, D are $\mathcal{EL}^+$ concept descriptions and $r \in \mathsf{RN}$ is a role name, then $C \sqcap D$ (conjunction) and $\exists r.C$ (existential restriction) are $\mathcal{EL}^+$ concept descriptions. Role chains of the form $r_1 \circ \cdots \circ r_n$ for $n \geq 0$ are called *role descriptions*. An $\mathcal{EL}^+$ *ontology* is a finite set of axioms of the form shown in the lower part of Table 1, where axioms of the form $C \sqsubseteq D$ are called general concept inclusions (GCIs) and of the form $r_1 \circ \cdots \circ r_n \sqsubseteq s$ role inclusions (RIs). An $\mathcal{EL}$ *ontology* is an $\mathcal{EL}^+$ ontology that does not contain RIs. We use $C \equiv D$ as an abbreviation for the two GCIs $C \sqsubseteq D, D \sqsubseteq C$.

The semantics of $\mathcal{EL}^+$ is defined in terms of *interpretations* $\mathcal{I} = (\Delta^{\mathcal{I}}, \cdot^{\mathcal{I}})$, where the domain $\Delta^{\mathcal{I}}$ is a non-empty set of individuals, and the interpretation function $\cdot^{\mathcal{I}}$ maps each concept name $A \in \mathsf{CN}$ to a subset $A^{\mathcal{I}}$ of $\Delta^{\mathcal{I}}$ and each role name $r \in \mathsf{RN}$ to a binary relation $r^{\mathcal{I}}$ on $\Delta^{\mathcal{I}}$. The extension of $\cdot^{\mathcal{I}}$ to arbitrary concept descriptions is inductively defined, as shown in the semantics column of Table 1. An interpretation $\mathcal{I}$ is a *model* of an ontology $\mathcal{O}$ if, for each inclusion axiom in $\mathcal{O}$, the conditions given in the semantics column of Table 1 are satisfied.

α_1	$\mathsf{AmpOfFinger}$	$\equiv$	$\mathsf{Amp} \sqcap \exists \mathsf{site}.\mathsf{Finger}_S$
α_2	$\mathsf{AmpOfHand}$	$\equiv$	$\mathsf{Amp} \sqcap \exists \mathsf{site}.\mathsf{Hand}_S$
α_3	$\mathsf{InjToFinger}$	$\equiv$	$\mathsf{Inj} \sqcap \exists \mathsf{site}.\mathsf{Finger}_S$
α_4	$\mathsf{InjToHand}$	$\equiv$	$\mathsf{Inj} \sqcap \exists \mathsf{site}.\mathsf{Hand}_S$
α_5	Finger_E	$\sqsubseteq$	Finger_S
α_6	Finger_P	$\sqsubseteq$	$\mathsf{Finger}_S \sqcap \exists \mathsf{part}.\mathsf{Finger}_E$
α_7	Hand_E	$\sqsubseteq$	Hand_S
α_8	Hand_P	$\sqsubseteq$	$\mathsf{Hand}_S \sqcap \exists \mathsf{part}.\mathsf{Hand}_E$
α_9	ULimb_E	$\sqsubseteq$	ULimb_S
α_{10}	ULimb_P	$\sqsubseteq$	$\mathsf{ULimb}_S \sqcap \exists \mathsf{part}.\mathsf{ULimb}_E$
α_{11}	Finger_S	$\sqsubseteq$	Hand_P
α_{12}	Hand_S	$\sqsubseteq$	ULimb_P

Figure 1: Ontology $\mathcal{O}_{\mathsf{Amp}}$ illustrating a faulty SEP-triplet encoding in SNOMED CT.

The main reasoning problem in $\mathcal{EL}^+$ is the *subsumption problem*: given an $\mathcal{EL}^+$ ontology $\mathcal{O}$ and two $\mathcal{EL}^+$ concept descriptions C, D, check whether C is *subsumed* by D w.r.t. $\mathcal{O}$ (written $C \sqsubseteq_{\mathcal{O}} D$), i.e., whether $C^{\mathcal{I}} \subseteq D^{\mathcal{I}}$ holds in all models of $\mathcal{O}$. The computation of all subsumption relationships between the concept names occurring in $\mathcal{O}$ is called *classification* of $\mathcal{O}$.

Figure 1 shows a small $\mathcal{EL}$ ontology defining concepts related to amputation/injury of hand and finger. It uses the so-called SEP-triplet encoding [12], in which anatomical concepts (like hand) are represented by three concepts: the structure concept (e.g, Hand_S, which stands for the hand and its proper parts), the part concept (e.g, Hand_P, which stands for the proper parts of the hand), and the entity concept (e.g, Hand_E, which stands for the entire hand). The axioms α_5–α_{10} constitute a completed SEP-triplet encoding for finger, hand, and upper limb. For example, axiom α_8 says that proper parts of the hand belong to the structure concept Hand_S, and they are parts of hand (i.e., linked with the role part to the entity concept Hand_E). Given this encoding, the fact that the finger is part of the hand can be expressed using axiom α_{11}. The main reason for using this encoding in SNOMED CT is that it can simulate transitivity reasoning for the role part, although transitivity of part cannot be expressed in $\mathcal{EL}$. For example, it is easy to see that the ontology $\mathcal{O}_{\mathsf{Amp}}$

implies that the finger is part of the upper limb, i.e., $\mathsf{Finger}_E \sqsubseteq_{\mathcal{O}_{\mathsf{Amp}}} \exists \mathsf{part}.\mathsf{ULimb}_E$. As a side-effect, the SEP-triplet encoding can also be used to simulate so-called right-identity rules [13], which allow to inherit properties along the part role. Consider the following subsumption relationships that hold in our example ontology:

$$\begin{aligned} \mathsf{AmpOfFinger} &\sqsubseteq_{\mathcal{O}_{\mathsf{Amp}}} \mathsf{AmpOfHand}, && (1) \\ \mathsf{InjToFinger} &\sqsubseteq_{\mathcal{O}_{\mathsf{Amp}}} \mathsf{InjToHand}. && (2) \end{aligned}$$

While subsumption (2) actually makes sense (it is sensible to say that an injury to the finger is an injury to the hand), subsumption (1) is clearly undesirable. Subsumption (1) is an example of a false positive subsumption, which does indeed occur in SNOMED CT. It has been argued [6, 7] that this subsumption is due to a faulty SEP-triplet encoding. In fact, using the entity concepts instead of the structure concepts in the axioms α_1 and α_2 would have avoided this problem.

In $\mathcal{EL}^+$, one could actually dispense with the SEP-triplet encoding altogether since both transitivity and right-identity rules can be expressed using RIs. For example, $\mathsf{part} \circ \mathsf{part} \sqsubseteq \mathsf{part}$ expresses transitivity of the role part. An alternative and direct representation of anatomical concepts, as well as referring concepts like clinical findings and procedures, based on this additional expressive power of the DL $\mathcal{EL}^+$ is proposed in [6]. The new modeling is succinct and also avoids the above false positive subsumption (1).

For a small ontology like $\mathcal{O}_{\mathsf{Amp}}$, it is not hard to do the subsumption reasoning manually, and thus also to find the axioms responsible for a given subsumption relationship by hand. For a very large ontology like SNOMED CT, this manual approach to pinpointing the responsible axioms is very time-consuming, and thus should be automated. First, we give a formal definition of what automated pinpointing is actually supposed to compute.

Definition 1 (MinA). Let $\mathcal{O}$ be an $\mathcal{EL}^+$ ontology, and A, B concept names such that $A \sqsubseteq_{\mathcal{O}} B$. The set $\mathcal{S} \subseteq \mathcal{O}$ is a *minimal axiom set (MinA) for* $A \sqsubseteq_{\mathcal{O}} B$ if, and only if, $A \sqsubseteq_{\mathcal{S}} B$ and, for every $\mathcal{S}' \subset \mathcal{S}$, $A \not\sqsubseteq_{\mathcal{S}'} B$. ◇

In our example, $\{\alpha_1, \alpha_2, \alpha_8, \alpha_{11}\}$ is the only MinA for subsumption (1), whereas $\{\alpha_3, \alpha_4, \alpha_8, \alpha_{11}\}$ is the only MinA for subsumption (2). As shown in [11], a given subsumption relationship w.r.t. an $\mathcal{EL}^+$ ontology may have exponentially many MinAs, and even deciding whether there is a MinA of cardinality $\leq k$ is an NP-complete problem. In contrast, one MinA can always be extracted in polynomial time. In [11], this was shown using the simple Algorithm 1, which requires linearly many (polynomial) subsumption tests. For a large ontology, however, this naive approach is not feasible. For example, for SNOMED CT it would require almost half a million subsumption tests for each MinA extraction.

Algorithm 1 Naive linear extraction of a MinA.

```
function lin-extract-mina($A, B, \mathcal{O}$)
$\mathcal{S} := \mathcal{O}$
for each axiom $\alpha \in \mathcal{O}$ do
    if $A \sqsubseteq_{\mathcal{S} \setminus \{\alpha\}} B$ then
        $\mathcal{S} := \mathcal{S} \setminus \{\alpha\}$
return $\mathcal{S}$
```

Algorithm 2 Logarithmic extraction of a MinA.

```
function log-extract-mina($A, B, \mathcal{O}$)
return log-extract-mina-r($A, B, \emptyset, \mathcal{O}$)
function log-extract-mina-r($A, B, S, \mathcal{O}$)
if $|\mathcal{O}| = 1$ then
    return $\mathcal{O}$
$S_1, S_2$ := halve($\mathcal{O}$)
if $A \sqsubseteq_{S \cup S_1} B$ then
    return log-extract-mina-r($A, B, S, S_1$)
if $A \sqsubseteq_{S \cup S_2} B$ then
    return log-extract-mina-r($A, B, S, S_2$)
$S'_1$ := log-extract-mina-r($A, B, S \cup S_2, S_1$)
$S'_2$ := log-extract-mina-r($A, B, S \cup S'_1, S_2$)
return $S'_1 \cup S'_1$
```

We can do much better by adopting the algorithm for computing prime implicates described in [14] to our problem. Basically, this algorithm applies binary search to find a MinA. Instead of taking out one axiom at a time, it partitions the ontology into two halves, and checks whether one of them entails the subsumption. If yes, it immediately recurses on that half, throwing away half of the axioms in one step. Otherwise, essential axioms are in both halves. In this case, the algorithm recurses on each half, using the other half as the "support set". Algorithm 2 describes this approach in more detail, where the function halve($\mathcal{O}$) partitions $\mathcal{O}$ into $S_1 \cup S_2$ with $||S_1| - |S_2|| \leq 1$. It follows from the results in [14] that computing a MinA $\mathcal{S}$ for a given subsumption $A \sqsubseteq_{\mathcal{O}} B$ with Algorithm 2 requires $O\left((|\mathcal{S}| - 1) + |\mathcal{S}| log(|\mathcal{O}|/|\mathcal{S}|)\right)$ subsumption tests. This greatly improves on the naive algorithm. For instance, computing a MinA consisting of nine axioms for SNOMED CT requires about one hundred subsumption tests. Though this is much better than the almost half a million required by the naive algorithm, it is still not good enough to compute MinAs on demand (see below).

Modularization-based axiom pinpointing in $\mathcal{EL}^+$

Instead of applying Algorithm 1 or 2 directly to the whole ontology $\mathcal{O}$, one can first try to find a non-minimal (but hopefully small) subset $\mathcal{S} \subseteq \mathcal{O}$ with $A \sqsubseteq_{\mathcal{S}} B$ (called *nMinA* in the following), and then apply Algorithm 1 or 2 to this subset to obtain a MinA. In [11], we have sketched a modified version of the classification algorithm for $\mathcal{EL}^+$ [3, 4] that extracts such nMinAs. In the experiments on a version of GALEN described in [11], Algorithm 1 was then used to minimize these sets. Whereas the nMinA extraction was fast and produced quite small sets for GALEN, it crashed after a few hours because of space problems when applied to SNOMED CT.

To overcome this problem, we propose an algorithm for extracting nMinAs that is based on modularization. In the following, we introduce only those notions regarding modularization that are strictly necessary in the context of this paper. More details regarding the reachability-based modularization approach from which these notions are derived, as well as its connection to other work on modularization, can be found in [5].

Let $\mathcal{O}$ be an $\mathcal{EL}^+$ ontology, and A a concept name occurring in $\mathcal{O}$. We say that $\mathcal{S} \subseteq \mathcal{O}$ is a *subsumption module for A in $\mathcal{O}$* whenever $A \sqsubseteq_{\mathcal{O}} B$ if, and only if, $A \sqsubseteq_{\mathcal{S}} B$ holds for all concept names B occurring in $\mathcal{O}$. Obviously, if $\mathcal{S}$ is a subsumption module for A in $\mathcal{O}$ and $A \sqsubseteq_{\mathcal{O}} B$, then $\mathcal{S}$ is an nMinA for this subsumption, and Algorithm 1 or 2 can be used to compute a MinA $\mathcal{S}' \subseteq \mathcal{S}$ from $\mathcal{S}$. Thus, we know that a subsumption module for A contains a MinA for every valid subsumption relationship $A \sqsubseteq_{\mathcal{O}} B$. The reachability-based modules introduced below satisfy an even stronger property: they contain *all* MinAs for all valid subsumptions.

Definition 2. Let $\mathcal{O}$ be an $\mathcal{EL}^+$ ontology and A a concept name occurring in $\mathcal{O}$. The subsumption module $\mathcal{S}$ for A in $\mathcal{O}$ is called *strong* if the following holds for all concept names B occurring in $\mathcal{O}$: if $A \sqsubseteq_{\mathcal{O}} B$, then every MinA for $A \sqsubseteq_{\mathcal{O}} B$ is a subset of $\mathcal{S}$. ◇

Obviously, $\mathcal{O}$ itself is a strong subsumption module for every concept name A occurring in $\mathcal{O}$. The following definition (first introduced in [5]) yields strong subsumption modules that are usually much smaller than the whole ontology. For an $\mathcal{EL}^+$ entity X—i.e., either a (concept or role) description, a (concept or role) inclusion axiom, or an ontology—we write $\mathsf{Sig}(X)$ to denote the set of concept and role names occurring in the entity X.

Definition 3 (Reachability-based modules). Let $\mathcal{O}$ be an $\mathcal{EL}^+$ ontology and A a concept name occurring in $\mathcal{O}$. The *set of A-reachable names in $\mathcal{O}$* is the smallest set N of concept and role names such that

- A belongs to N;
- for all (concept/role) inclusion axioms $\alpha_L \sqsubseteq \alpha_R$ in $\mathcal{O}$, if $\mathsf{Sig}(\alpha_L) \subseteq \mathsf{N}$ then $\mathsf{Sig}(\alpha_R) \subseteq \mathsf{N}$.

We call an axiom $\alpha_L \sqsubseteq \alpha_R$ *A-reachable in $\mathcal{O}$* if every element of $\mathsf{Sig}(\alpha_L)$ is A-reachable in $\mathcal{O}$. The *reachability-based module for A in $\mathcal{O}$*, denoted by $\mathcal{O}_A^{\mathsf{reach}}$, consists of all A-reachable axioms from $\mathcal{O}$. ◇

In [5], it has been shown that $\mathcal{O}_A^{\mathsf{reach}}$ is indeed a subsumption module for A in $\mathcal{O}$. Here, we show the following stronger results.

Theorem 4. *Let $\mathcal{O}$ be an $\mathcal{EL}^+$ ontology and A a concept name. Then $\mathcal{O}_A^{\mathsf{reach}}$ is a* strong *subsumption module for A in $\mathcal{O}$.*

Proof. The fact that $\mathcal{O}_A^{\mathsf{reach}}$ is a subsumption module was already shown in [5]. To show that it is strong, assume that $A \sqsubseteq_{\mathcal{O}} B$ holds, but there is a MinA $\mathcal{S}$ for $A \sqsubseteq_{\mathcal{O}} B$ that is not contained in $\mathcal{O}_A^{\mathsf{reach}}$. Thus, there is an axiom $\alpha \in \mathcal{S} \setminus \mathcal{O}_A^{\mathsf{reach}}$. Let $\mathcal{S}_1$ be the subset of $\mathcal{S}$ that contains the A-reachable axioms. Note that $\mathcal{S}_1$ is a strict subset of $\mathcal{S}$ since $\alpha \notin \mathcal{S}_1$. We claim that $A \sqsubseteq_{\mathcal{S}} B$ implies $A \sqsubseteq_{\mathcal{S}_1} B$, which contradicts the assumption that $\mathcal{S}$ is a MinA for $A \sqsubseteq_{\mathcal{O}} B$.

To show the claim, we assume to the contrary that $A \not\sqsubseteq_{\mathcal{S}_1} B$, i.e., there is a model $\mathcal{I}_1$ of $\mathcal{S}_1$ such that $A^{\mathcal{I}_1} \not\subseteq B^{\mathcal{I}_1}$. We modify $\mathcal{I}_1$ to $\mathcal{I}$ by setting $y^{\mathcal{I}} := \emptyset$ for all (concept or role) names that are not A-reachable. It is easy to see that $A^{\mathcal{I}} \not\subseteq B^{\mathcal{I}}$. In fact, we have $A^{\mathcal{I}} = A^{\mathcal{I}_1}$ (since A is A-reachable), and $B^{\mathcal{I}} = B^{\mathcal{I}_1}$ or $B^{\mathcal{I}} = \emptyset$.

It remains to show that $\mathcal{I}$ is indeed a model of $\mathcal{S}$, i.e. satisfies all axioms $\beta_L \sqsubseteq \beta_R$ in $\mathcal{S}$. If β_L contains a name that is not A-reachable, then $(\beta_L)^{\mathcal{I}} = \emptyset$, and the axiom is trivially satisfied. Otherwise, this axiom belongs to $\mathcal{S}_1$, and the definition of A-reachability implies that all names in β_R are A-reachable as well. Consequently, $\mathcal{I}_1$ and $\mathcal{I}$ coincide on the names occurring in $\beta_L \sqsubseteq \beta_R$. Since $\mathcal{I}_1$ is a model of $\mathcal{S}_1$, we thus have $(\beta_L)^{\mathcal{I}} = (\beta_L)^{\mathcal{I}_1} \subseteq (\beta_R)^{\mathcal{I}_1} = (\beta_R)^{\mathcal{I}}$. ❑

As an immediate consequence of this theorem, instead of extracting a MinA for $A \sqsubseteq_{\mathcal{O}} B$ from $\mathcal{O}$, it is sufficient to extract a MinA for $A \sqsubseteq_{\mathcal{O}_A^{\mathsf{reach}}} B$ from $\mathcal{O}_A^{\mathsf{reach}}$. This is what the function $\mathsf{extract\text{-}mina}$ in

Algorithm 3 Modularization-based extraction of a MinA

```
function extract-mina($A, B, \mathcal{O}$)
    $\mathcal{O}_A^{\mathsf{reach}} \leftarrow$ extract-module($\mathcal{O}, A$)
    return log-extract-mina($A, B, \mathcal{O}_A^{\mathsf{reach}}$)
function second-mina?($A, B, \mathcal{O}_A^{\mathsf{reach}}, \mathcal{S}_1$)
    for each axiom $\alpha \in \mathcal{S}_1$ do
        $\mathcal{O}' \leftarrow \mathcal{O}_A^{\mathsf{reach}} \setminus \{\alpha\}$
        if $A \sqsubseteq_{\mathcal{O}'} B$ then
            return "second MinA exists"
    return "MinA unique"
function extract-module($\mathcal{O}, A$)
    $\mathcal{O}_A \leftarrow \emptyset$
    queue $\leftarrow$ active-axioms($\{A\}$)
    while not empty(queue) do
        $(\alpha_L \sqsubseteq \alpha_R) \leftarrow$ fetch(queue)
        if $\mathsf{Sig}(\alpha_L) \subseteq \{A\} \cup \mathsf{Sig}(\mathcal{O}_A)$ then
            $\mathcal{O}_A \leftarrow \mathcal{O}_A \cup \{\alpha_L \sqsubseteq \alpha_R\}$
            queue $\leftarrow$ queue $\cup$
                        (active-axioms($\mathsf{Sig}(\alpha_R)$) $\setminus$ $\mathcal{O}_A$)
    return $\mathcal{O}_A$
```

Algorithm 3 does. Note that, instead of the logarithmic extraction algorithm (Algorithm 2), we could also use the linear extraction algorithm (Algorithm 1). Since reachability-based modules are usually quite small, it is not a priori clear whether using the more complicated logarithmic algorithm really pays off (see the results of our experiments below). The function second-mina? in Algorithm 3 takes the extracted module and the first MinA as input, and checks if the subsumption in question still holds in the absence of one of the axioms in the MinA. In this case, this subsumption obviously must have more than one MinA. Note that, for this function to be correct, we really need to know that $\mathcal{O}_A^{\mathsf{reach}}$ is a *strong* subsumption module.

The function extract-module in Algorithm 3 realizes one way of computing reachability-based modules. The function call active-axioms used there yields, for a given set of names, all axioms that contain at least one of these names in their left-hand side. It is not hard to show that the call extract-module($\mathcal{O}, A$) indeed computes the reachability-based module for A in $\mathcal{O}$ (see [5] for more details). The experiments described in [5] show that extraction of reachability-based modules in Snomed ct is usually quite fast, and the modules obtained this way are quite small. In the next section, we show that these positive results extend to the modularization-based extraction of MinAs.

Experimental Results

We have implemented the three algorithms described in this paper, using CEL [4] to compute subsumption. Our experiments use the January/2005 release of the DL version of Snomed ct, which contains 379,691 concept names, 62 role names, and 379,704 axioms.[3] In the following, we call this ontology $\mathcal{O}^{\text{Snomed}}$. The experiments were carried out on a PC with 2.40 GHz Pentium-4 processor and 1 GB of memory.

As stand-alone algorithms for computing a MinA, we applied Algorithm 1 and 2 only to the false positive subsumption $\mathsf{AmpOfFinger} \sqsubseteq_{\mathcal{O}^{\text{Snomed}}} \mathsf{AmpOfHand}$. Algorithm 1 did not terminate on this input after 24 hours, whereas Algorithm 2 required 26:05 minutes (1,565 seconds) to compute a MinA of cardinality 6. (Note that the actual modelling of "amputation of finger" and "amputation of hand" in Snomed ct differs from the one given in Fig. 1 due to the use of role groups and of two different roles to express location in Snomed ct. Thus, the computed MinA also differs from the one given above. However, it also shows that the reason for the unintended subsumption is the incorrect use of the SEP-triplet encoding.)

Algorithm 3 performs much better for the amputation example. The reachability-based module $\mathcal{O}^{\text{Snomed}}_{\mathsf{AmpOfFinger}}$ contains 57 axioms, and was computed in 0.04 seconds. Extracting a MinA for $\mathsf{AmpOfFinger} \sqsubseteq_{\mathcal{O}^{\text{Snomed}}_{\mathsf{AmpOfFinger}}} \mathsf{AmpOfHand}$ from $\mathcal{O}^{\text{Snomed}}_{\mathsf{AmpOfFinger}}$ using the logarithmic minimization algorithm then took only half a second. An application of second-mina? then showed that the extracted MinA is the only one for this subsumption.

We have also applied Algorithm 3 to a large number of subsumption relationships that follow from $\mathcal{O}^{\text{Snomed}}$. Since there are more than five million such subsumptions, testing the algorithm on all of them was not feasible: assuming an average extraction time of 1 second, this would have required 58 days. For this reason, we sampled 0.5% of all concepts in each top-level category C in Snomed ct. Let us denote the set of samples for category C by c-samples(C). For each sampled concept A, all positive subsumptions $A \sqsubseteq_{\mathcal{O}^{\text{Snomed}}} B$ with A as subsumee were considered.

The first column of Table 2 shows the top-level categories and the second the number of sampled subsumption relationships with the subsumee in this category. The next four columns give the time needed to compute and the size of the corresponding modules and MinAs. The values in square brackets give the time required by the

[3] The *DL version* is also known in the SNOMED lingo as the 'stated form,' while *axioms* here boil down to (primitive) concept definitions.

SNOMED category C $A \in \textsf{c-samples}(C)$	♯Subs. samples	Time to extract module $\mathcal{O}_A^{\text{SNOMED}}$ (avg/max)	Size of $\mathcal{O}_A^{\text{SNOMED}}$ (avg/max)	Time to extract MinA for $A \sqsubseteq_{\mathcal{O}_A^{\text{SNOMED}}} B$ (avg/max)	Size of MinA for $A \sqsubseteq_{\mathcal{O}_A^{\text{SNOMED}}} B$ (avg/max)	%Subs. with one MinA
Attribute	25	< 0.01 / < 0.01	5.12/8	0.05 / 0.09 [0.15 / 0.18]	3.16/7	100
Body structure	4 738	< 0.01 / 0.01	40.76 / 76	0.41 / 4.19 [0.63 / 2.24]	5.54 / 18	64.16
Clinical Finding	11 112	0.03 / 3.97	71.50 / 143	1.66 / 9.58 [1.15 / 5.04]	9.00 / 34	63.00
Context-dependent categories	208	0.01 / 0.03	0.14 / 108	0.37 / 1.43 [0.63 / 1.77]	4.10 / 13	95.67
Environments & geographical locations	51	< 0.01 / < 0.01	7.65 / 9	0.07 / 0.12 [0.17 / 0.19]	3.82 / 8	100
Events	28	< 0.01 / < 0.01	4.64 / 6	0.04 / 0.08 [0.13 / 0.16]	2.32 / 5	100
Observable entity	253	< 0.01 / < 0.01	8.26 / 12	0.08 / 0.18 [0.18 / 0.24]	3.68 / 8	90.12
Organism	1 429	< 0.01 / 0.01	13.03 / 21	0.09 / 0.20 [0.25 / 0.36]	4.72 / 13	65.01
Pharmaceutical/biologic product	1 233	< 0.01 / 0.01	31.41 / 60	0.16 / 0.47 [0.50 / 0.91]	3.68 / 10	81.51
Physical force	6	< 0.01 / < 0.01	7.00 / 7	0.38 / 0.58 [0.16 / 0.17]	2.83 / 5	50.00
Physical object	166	< 0.01 / < 0.01	9.35 / 12	0.09 / 0.19 [0.19 / 0.24]	4.18 / 11	93.98
Procedure	5 183	0.02 / 0.05	71.89 / 146	0.62 / 5.21 [0.15 / 4.38]	8.65 / 36	66.29
Qualifier value	216	< 0.01 / 0.01	6.68 / 11	0.06 / 0.13 [0.16 / 0.22]	2.67 / 7	87.96
Social context	204	< 0.01 / 0.01	10.11 / 15	0.18 / 0.47 [0.21 / 0.28]	3.59 / 9	77.45
Special concept	1 272	< 0.01 / 0.01	5.00 / 5	0.05 / 0.09 [0.14 / 0.14]	2.5 / 4	100
Specimen	38	0.01 / 0.02	67.74 / 127	0.19 / 0.59 [1.06 / 2.11]	4.55 / 13	81.58
Staging and scales	20	< 0.01 / < 0.01	5.60 / 8	0.04 / 0.08 [0.14 / 0.18]	2.8 / 7	100
Substance	1 295	< 0.01 / 0.01	14.76 / 32	0.16 / 0.41 [0.19 / 0.51]	4.17 / 12	72.82
Overall in SNOMED CT	27 477	0.02 / 3.97	53.21 / 146	1.03 / 9.58 [0.67 / 5.04]	7.11 / 36	68.26

Table 2: Empirical results of the modularization-based axiom pinpointing on SNOMED CT (time in seconds; size in number of axioms).

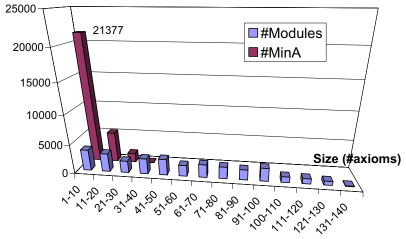


Figure 2: Module and MinA size distribution.

modularization-based pinpointing algorithm, but with the naive linear minimization algorithm instead of the logarithmic one. In all four columns, we give both average and maximum values. The last column shows the percentage of subsumptions that have only one MinA. Interestingly, more than two thirds of all subsumptions have only one MinA. The overall empirical results for the 27,477 sampled subsumptions (about 0.5% of all subsumptions) are given in the last row of the table. These results show that, on average, a MinA can be computed within one second, and its size is smaller than 10. Thus, MinAs can indeed be computed on demand, and their size is small enough such that they can then be inspected by hand.

Surprisingly, the linear minimization algorithm performed better in our experiments than the logarithmic one. An explanation for this is probably that, unlike the experiments of Algorithm 1 and 2 on the whole ontology, the modules are already quite small, and thus the overhead required by the logarithmic algorithm does not pay off. Figure 2 depicts the size distribution of our sampled modules and MinAs. As easily visible from the chart, the modules are quite small, but the MinAs are even smaller. In fact, the majority of all subsumptions (78%) have a MinA of size ten or less.

Conclusions

We have introduced a new method for axiom pinpointing in the DL $\mathcal{EL}^+$ that is based on the computation of reachability-based modules. The experiments carried out on SNOMED CT show that this method is fast enough to extract a minimal axiom set (MinA) for a given subsumption on demand. In addition, the extracted MinAs are usually quite small and can therefore be inspected by users and designers of SNOMED CT by hand. In the future, we will extend the approach such that it can (i) extract all MinAs, (ii) provide natural language explanations for subsumption, and (iii) give suggestions for how to revise the ontology to get rid of an unwanted subsumption.

Acknowledgements

The first author was partially supported by NICTA, Canberra Research Lab, and the second by the German Research Foundation (DFG) under grant BA 1122/11-1.

Address for Correspondence

Franz Baader and Boontawee Suntisrivaraporn
TU Dresden, Theoretical Computer Science,
01062 Dresden, Germany
{baader,meng}@tcs.inf.tu-dresden.de

References

[1] F. Baader, D. Calvanese, D. McGuinness, D. Nardi, and P. F. Patel-Schneider, editors. *The Description Logic Handbook: Theory, Implementation, and Applications.* Cambridge University Press, 2003.

[2] I. Horrocks, P. F. P.-Schneider, and F. van Harmelen. From SHIQ and RDF to OWL: The making of a web ontology language. *Journal of Web Semantics*, 1(1):7–26, 2003.

[3] F. Baader, S. Brandt, and C. Lutz. Pushing the $\mathcal{EL}$ envelope. In *Proc. IJCAI 2005.*

[4] F. Baader, C. Lutz, and B. Suntisrivaraporn. CEL—a polynomial-time reasoner for life science ontologies. In *Proc. IJCAR 2006*, Springer LNAI 4130, 2006.

[5] B. Suntisrivaraporn. Module extraction and incremental classification: A pragmatic approach for $\mathcal{EL}^+$ ontologies. In *Proc. ESWC 2008*, Springer LNCS, 2008. To appear.

[6] B. Suntisrivaraporn, F. Baader, S. Schulz, and K. Spackman. Replacing SEP-triplets in SNOMED CT using tractable description logic operators. In *Proc. AIME 2007*, Springer LNCS 4594, 2007.

[7] U. Hahn S. Schulz, K. Mark. Spatial location and its relevance for terminological inferences in bio-ontologies. *BMC Bioinformatics*, 2007.

[8] S. Schlobach and R. Cornet. Non-standard reasoning services for the debugging of description logic terminologies. In *Proc. IJCAI 2003.*

[9] B. Parsia, E. Sirin, and A. Kalyanpur. Debugging OWL ontologies. In *Proc. WWW 2005.*

[10] T. Meyer, K. Lee, R. Booth, and J. Z. Pan. Finding maximally satisfiable terminologies for the description logic $\mathcal{ALC}$. In *Proc. (AAAI 2006).* AAAI Press/The MIT Press, 2006.

[11] F. Baader, R. Peñaloza, and B. Suntisrivaraporn. Pinpointing in the description logic $\mathcal{EL}^+$. In *Proc. KI 2007*, Springer LNAI 4667, 2007.

[12] S. Schulz, M. Romacker, and U. Hahn. Part-whole reasoning in medical ontologies revisited: Introducing SEP triplets into classification-based description logics. *JAMIA*, 1998.

[13] K.A. Spackman. Managing clinical terminology hierarchies using algorithmic calculation of subsumption: Experience with SNOMED-RT. *JAMIA*, 2000. Fall Symposium Special Issue.

[14] A. R. Bradley and Z. Manna. Checking safety by inductive generalization of counterexamples to induction. In *Proc. FMCAD 2007.*

Representing and sharing knowledge using SNOMED
Proceedings of the 3rd international conference on Knowledge Representation in Medicine (KR-MED 2008)
R. Cornet, K.A. Spackman (Eds)

Exploiting Fast Classification of SNOMED CT for Query and Integration of Health Data

Michael J. Lawley
Queensland University of Technology, Faculty of Information Technology, Brisbane, (Queensland), Australia
E-Health Research Centre, CSIRO ICT Centre, (Queensland), Australia

Abstract

By constructing local extensions to SNOMED *we aim to enrich existing medical and related data stores, simplify the expression of complex queries, and establish a foundation for semantic integration of data from multiple sources.*

Specifically, a local extension can be constructed from the controlled vocabulary(ies) used in the medical data. In combination with SNOMED, *this local extension makes explicit the implicit semantics of the terms in the controlled vocabulary. By using* SNOMED *as a base ontology we can exploit the existing knowledge encoded in it and simplify the task of reifying the implicit semantics of the controlled vocabulary. Queries can now be formulated using the relationships encoded in the extended* SNOMED *rather than embedding them ad-hoc into the query itself. Additionally,* SNOMED *can then act as a common point of integration, providing a shared set of concepts for querying across multiple data sets.*

Key to practical construction of a local extension to SNOMED *is appropriate tool support including the ability to compute subsumption relationships very quickly. Our implementation of the polynomial algorithm for $\mathcal{EL}+$ in Java is able to classify* SNOMED *in under 1 minute.*

INTRODUCTION

Experience with integrating medical and related data [1] shows that the use of controlled vocabularies successfully modulates the amount of noise in the data. However, when querying the collected data, any semantic relationships between the terms that are relevant to the query (for example, specialisation/generalisation or part-of relationships) need to be explicitly encoded in the query and/or accounted for in the interpretation of the query results.

These kinds of implicit relationships are especially common in the health domain where terms often involve an implicit context of usage (e.g., *lobe* in the context of lung cancer) or implicit references to anatomical structures (e.g., colorectal cancer) or related classes of diseases, injuries, or procedures. Accurately and consistently encoding these relationships in queries relies on the person formulating the queries to understand them, thus creating many opportunities for errors, omissions, and inconsistencies to occur. When multiple people are constructing queries these risks are further exacerbated.

By constructing the vocabularies so as to explicitly represent the relationships between terms, queries can directly and consistently exploit the relationships. Using an ad-hoc explicit representation of these relationships helps, but may introduce new problems in terms of consistency of usage and how the relationships are interpreted (see, for example, the Radiological Electronic Atlas of Malformation Syndromes and Skeletal Dysplasias (REAMS) [2]). Instead, using a well-understood formal mechanism for representing the relationships, such as Description Logic, can avoid these problems.

However we still have two problems to solve:

1. how do we deal with all the existing data sets that do not do this; and
2. how do we mitigate the, potentially quite high[1], cost of explicitly representing all the relationships?

We can deal with both these problems by extending (as needed) an existing standard ontology, such as the Systematized Nomenclature of Medicine (SNOMED) [3], that already embodies

[1] Getting the modelling right, from scratch, requires not only an excellent understanding of the concepts involved as well as their relationships, but also an understanding of how best to represent them in a particular Description Logic formalism.

many of the relationships we need. However, one of the main difficulties with this approach is that building an extension to SNOMED is not dissimilar to maintaining and developing SNOMED itself. That is, the sheer size of SNOMED has meant that, until recently, very few tools could compute all of its subsumption relationships, and even those that could would reportedly take several hours.

Fortunately, recent work by Baader et al. [4, 5] on the tractable family of description logics $\mathcal{EL}$ has shown that polynomial time classification algorithms exist and are practical. Moreover despite their relatively low expressive power, the $\mathcal{EL}$ family of description logics is suitable for representing such real-world ontologies as SNOMED and offer additional expressiveness suitable for properly representing partOf relationships and sufficient conditions.[2] Their implementation of this algorithm in Lisp is able to classify SNOMED in 1,782 seconds [5] (approx. 30 minutes) which suggests an optimised implementation in a lower-level language may be fast enough for near real-time feedback in an editing tool.

Thus, our goal is to provide tool support for defining a *local extension* to an existing standard formal ontology; a mapping from an existing set of terms that characterise an informal ontology to concepts in the formal ontology. In doing so we effectively realise latent semantics in the existing medical data via the standard ontology. This should facilitate simpler and more robust queries and in turn aid data integration, a special-case application of querying where related medical data sets use semantically overlapping, but distinct term sets.

RELATED WORK

There is a great deal of published work on using ontologies for data integration (see Wache et al. [6] for an overview), but it is mostly focussed on their use at the meta-data level; ontologies are used to describe, reason about and integrate database schemas. While related to our goals, we are addressing the more specific problem of semantic data integration or semantic translation. Stuckenschmidt et al. [7] discuss an approach to this problem in the context of their Ontology Interchange Language (OIL) [8]. In particular they raise the question of whether it is feasible to find or create a sufficient shared terminology. In our domain of medical data we believe that SNOMED represents such a shared terminology. A possibly more important problem, and one identified in our work with skeletal dysplasias [2], is how to cope with errors in the shared terminology.

Wade and Rosenbloom [9] report on the manual construction of what is almost a local extension to SNOMED (they conceived the task as a semi-formal mapping). In this work 2002 terms were mapped to combination of single and post-coordinated concepts of which about 75% were equivalencies (20% of these were to single concepts) and only 1% (26) were, in their words, *"not mappable"*. It is unclear why these terms were categorized as such since they include, for example, *presyncope* which could reasonably be related to 3006004|disturbance of consciousness|, but it may be that the context of use of the terms was unavailable in order to properly discern their meaning. However, their work does demonstrate that the goal of producing a local extension to SNOMED is feasible.

PROBLEM DESCRIPTION

The problem of embedding domain semantics such as specialisation/generalisation or part-of relationships into queries is illustrated in the following. For example, a query to find all performed procedures involving a colectomy might enumerate all such procedures:

```
SELECT S.*
FROM Surgery S
WHERE S.procedure = '32003-00'
   OR S.procedure = '32003-01'
   OR S.procedure = '32012-00'
   ...
```

which has the potential to accidently omit certain codes and will require updating if the terminology is updated with additional forms of colectomy.

Alternatively, some kind of heuristic query could be used:

```
SELECT S.*
FROM Surgery S, ProcedureCodes C
WHERE S.procedure = C.code
  AND C.text LIKE '%colectomy%';
```

which has the potential to miss a term that doesn't follow the expected naming pattern (e.g., epiploectomy) or provide false matches where a compound or composite name does not reflect a valid specialisation.

If, however, the terms were encoded as concepts in an ontology, the query is simple[3]:

[2] See also `http://webont.org/owl/1.1/tractable.html#2`

[3] We envisage that the complete set of subsumption relationships would be stored in a database table to support fast subsumption-based queries using only two

```
SELECT S.*
FROM Surgery S, Ontology O
WHERE O.ancestor = 23968004
  AND S.procedure = O.descendant;
```

Note also that SNOMED, unlike classification schemes such as ICD-9 and ICD-10, support a multi-parented generalisation hierarchy.

CONSTRUCTING LOCAL EXTENSIONS

In order to construct an ontology from an existing terminology (or collection of terminologies) we take a multi-step approach:

1. Map each term from the controlled vocabulary to a concept, factoring out any synonyms, to produce $\mathcal{P}$.

   This is often a simple one-to-one mapping, but it may be necessary to extend the mapping to include disambiguating data values when the same term is used to mean different things in different contexts.

2. Make any simple implicit relationships explicit, adding them to $\mathcal{P}$.

   For example, generalisation, *partOf*, or *hasLocation* relationships. It may be necessary to introduce new concepts to act as the generalisation of two or more sibling concepts.

3. Specify relationships between these (local) concepts and those in the chosen standard ontology $\mathcal{Q}$, adding them to $\mathcal{P}$.

To be able to answer queries involving our new ontology we first need to classify $\mathcal{Q} \cup \mathcal{P}$ to identify all the subsumption relationships it entails.
Note that, we should be careful that $\mathcal{Q} \cup \mathcal{P}$ represents a conservative extension [10] of $\mathcal{Q}$. That is, $\mathcal{Q} \cup \mathcal{P}$ produces the same consequences over the set of concepts in $\mathcal{Q}$ as $\mathcal{Q}$ does by itself. We also need to ensure various integrity constraints (such as disjointness) are preserved in $\mathcal{Q} \cup \mathcal{P}$. Thus we would like to be able to interactively edit $\mathcal{P}$ while exploiting the consequences of $\mathcal{Q} \cup \mathcal{P}$ in live feedback through the mapping tool. These kinds of checks can be performed by classification of $\mathcal{Q} \cup \mathcal{P}$ but this may not be viable if $\mathcal{Q} \cup \mathcal{P}$ is large, as is the case when $\mathcal{Q}$ is SNOMED.

Colorectal Cancer Example

In this section we consider a sample set of ICD-10-AM [11] terms for procedures relating to colorectal cancer, shown in Figure 1. We can map these, one-to-one, to a set of concepts for a local ontology.

Procedure Code (ICD-10-AM)	Meaning
32000-00	Sig colectomy *with stoma formation*
32003-00	Sig colectomy *with anastomosis*
32003-01	Right hemicolectomy
32005-00	Subtotal colectomy
32005-01	Ext right hemicolectomy
32006-00	Left hemicolectomy
32012-00	Total colectomy
32024-00	High anterior resection
32025-00	Low anterior resection *extraperitoneal*
32026-00	Low anterior resection *coloanal anastomosis*
32028-00	Ultra low anterior resection
32030-00	Hartmann's procedure
32039-00	Abdomino-perineal excision
32051-00	Total proctocolectomy with ileo-anal anastomosis

Figure 1: A Term-Set of Colorectal Cancer Procedures

The next step is to make any simple relationship explicit. In our case there are none that can be expressed using just the concepts we have currently identified.

Figure 2 describes the identified relationships between these terms and selected SNOMED concepts as per step 3. Note that several concepts (for example, 32028-00|ultra low anterior resection|), have no exact equivalent in SNOMED, and that one, 32051|total proctocolectomy with ileo-anal anastomosis| implies a composite of concepts.

Figure 3 shows a visualisation of the results of classifying SNOMED augmented with the ontology from Figure 2. As can be seen, unifying generalisation concepts such as 84604002|sigmoid colectomy| have been identified, and thus provide a strong foundation for constructing queries that span the various procedures. Additionally, since SNOMED includes detailed anatomical concepts, queries can now be composed in terms of anatomical features even though they did not exist in the original terminology.

joins.

Procedure	Relation	SNOMED
32000-00	$\equiv$	315327002
32003-00	$\equiv$	315326006
32003-01	$\equiv$	235326000
32005-00	$\equiv$	43075005
32005-01	$\equiv$	174071004
32006-00	$\equiv$	82619000
32012-00	$\equiv$	26390003
32024-00	$\equiv$	400988008
32025-00	$\sqsubseteq$	314592008
32026-00	$\sqsubseteq$	314592008
32028-00	$\sqsubseteq$	314592008
32030-00	$\equiv$	16564004
32039-00	$\equiv$	265414003
32051-00	$\sqsubseteq$	174059005 $\sqcap$ 70172002

Figure 2: Identified Relationships with SNOMED Concepts

COMPLEX QUERIES AND CONTEXT

So far we have only considered simple query scenarios where a single database column represents the concept we wish to query (e.g., `Surgery.procedure`) and there already exists a concept that characterises the bound of the query (e.g., 2396804).

Consider instead a table, as shown in Figure 4, that stores both scheduled and performed procedures while using another column to distinguish them, and which encodes laterality, if any, of the procedure in yet another column. Now imagine we wish to query for all patients who have had an amputation including the left hand.

Patient	Date	Status
...	...	...

	Procedure	Laterality
	...	...

Figure 4: Table storing records with contextual information split across columns

Patient	Date	...	Laterality	Code
...		...		...

Code	Equivalent SNOMED Expression
...	...

Figure 5: Augmented table for representing contextualised concepts

To support this kind of problem with reasonable generality and decent query speed, we need to generate a new column containing codes that are mapped to the set of compound concepts that correspond to the contextualised meaning of each database row. Hence, as shown in Figure 5, the table from Figure 4 would be extended with a `Code` foreign-key column, and an additional table containing the SNOMED expressions of the form[4]:

$$\begin{aligned}&\exists\ \text{associatedProcedure}.\langle \text{P}\rangle && \sqcap\\ &\exists\ \text{laterality}.\langle \text{L}\rangle && \sqcap\\ &\exists\ \text{procedureContext}.\langle \text{S}\rangle &&\end{aligned}$$

which gives us another ontology extension that we can add to SNOMED.

Finally, in order to be able to pose a subsumption-based complex query involving composite concepts and have it evaluated at database join speeds, we can employ the same strategy: extend the ontology with a new fully-defined concept corresponding to our query expression, re-classify, and perform a join-based query using the new concept.

The need to construct compound expressions that explicitly represent the context associated with a record in a database occurs any time the data needs to be queried outside its original context. This may happen in as trivial a case as when one table in a database is joined with another, but the more general scenario occurs when integrating data from multiple data sources.

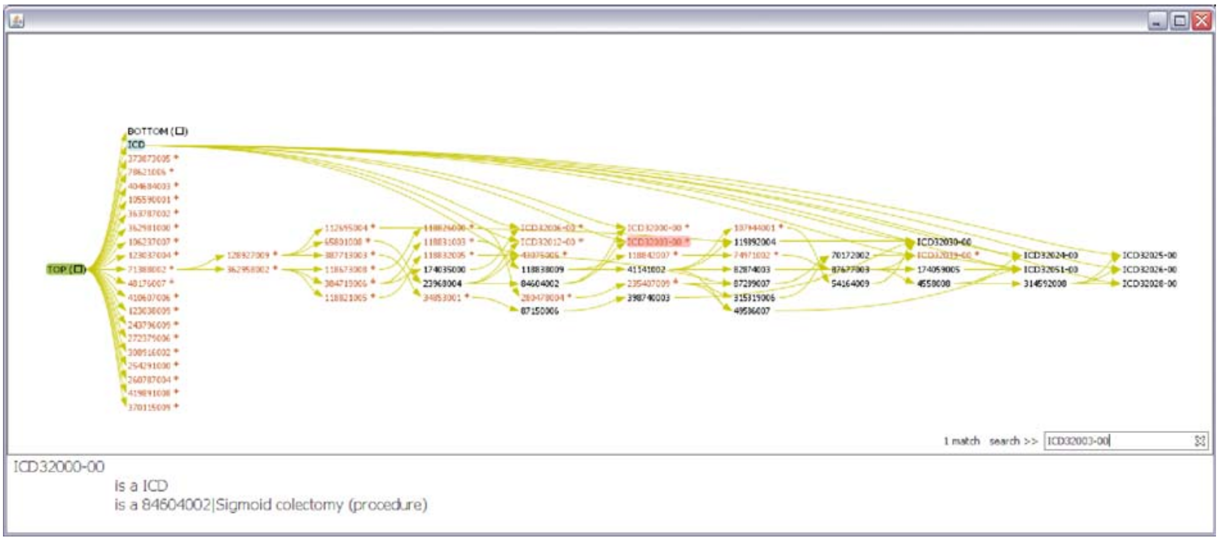
RESULTS

Classifying SNOMED

The practicality of creating local extensions of SNOMED is dependent on sufficient tool support and, as mentioned previously, a cornerstone of this is fast classification. Indeed we believe that near real-time feedback in an editing environment, be it an IDE for programming or a 3D architectural modelling tool, can have a transformational effect on the authoring and editing process.

To this end, we have implemented *snorocket*, using a slightly altered form of the algorithm in [5] written in Java. We use several optimised Map and Set data-structures tailored for ontologies with roughly the same number of concepts and roles as SNOMED. This implementation is able to classify

[4] Note that considerable experience with SNOMED and all its documentation may be required to construct suitable valid post-coordinated expressions like those above. Tool support for this is clearly an important issue and recent work in the IHTSDO Concept Model SIG on producing a Machine Readable Concept Model will be valuable for this.



1 match search >> ICD32003-00

ICD32000-00
is a ICD
is a 84604002|Sigmoid colectomy (procedure)

Figure 3: Visualisation of part of an extended SNOMED ontology

SNOMED in 54 seconds on a modern 2.4GHz Intel Core 2 Duo running Windows XP and Sun's Java 1.6.0_03.

For a fairer comparison with CEL, which only runs under Linux, we ran both snorocket and CEL on an older four-CPU Xeon 3.6GHz machine running RedHat Linux 2.6.9 and Sun's Java 1.6.0_04. The results, for several of the ontologies available from `http://lat.inf.tu-dresden.de/~meng/toyont.html`, are in Table 1.

Clearly, being able to classify SNOMED in close to a minute is a substantial improvement over roughly 23 minutes and brings us much closer to the near real-time feedback we are seeking.

Incremental Classification

In our mapping scenario we observe that SNOMED ($\mathcal{Q}$) is unchanging while the local extension ($\mathcal{P}$) is modified. If we can classify $\mathcal{Q}$ once and record the result $C(\mathcal{Q})$ then, due to the monotonicity of the description logic, the classification of $\mathcal{Q} \cup \mathcal{P}$, $C(\mathcal{Q} \cup \mathcal{P})$, is a superset of $C(\mathcal{Q})$. The goal is then to derive $C(\mathcal{Q} \cup \mathcal{P})$ given $C(\mathcal{Q})$ (and, of course, $\mathcal{Q}$ and $\mathcal{P}$) which should be much faster than deriving $C(\mathcal{Q} \cup \mathcal{P})$ from scratch.

Suntisrivaraporn [12] calls this Duo-Ontology Classification and presents a variation of the algorithm in [5] to do just this. We have independently derived our own variant of this algorithm along similar lines; the queue-processing core is essentially unchanged but the initialisation of the queues is different to account for the work that has already been done.

Currently this work is in a preliminary state and the correspondence with the variant described in [12] is unknown. However the performance of this incremental algorithm is very promising. With $\mathcal{P}$ consisting of the 14 new concepts as defined as in Figure 2, incremental classification takes around 0.9s using our un-optimised implementation.

DISCUSSION

Ideally, as a term set is developed, it would be explicitly constructed as an ontology and, to avoid re-invention and promote interoperability, could be developed as an extension of an existing standard ontology such as SNOMED. These extension ontologies could then be shared and evolved within their specialist community while still being useful and usable in more general communities. One such example is an ontology for skeletal dysplasias extracted from REAMS [13].

It is thus useful to be able to represent these ontologies in a standard format such as OWL so that they can be shared or manipulated using existing toolsets. Currently we use the OWL 1.1 proposal [14] rather than OWL 1.0 since it supports the expression of the role axioms (to describe role transitivity and right-identity). The particular subset we use is characterised by the description logic $\mathcal{EL}{+}^{\perp}$. OWL 1.1 is supported by, for example, the latest development-release of Protégé (4.0 alpha).

Unfortunately, OWL is not practical for representing large ontologies like SNOMED where an OWL

	SNOMED	FULL-GALEN	NOT-GALEN	NCI
CEL	1391.9	368.9	5.4	1.8
snorocket	72.8	15.1	0.4	0.4

Table 1: Comparison of classification time for snorocket and CEL running on the same hardware.

XML representation is approximately 240MB [15], about eight times the size of the equivalent KRSS representation. Moreover, due to the complexities inherent in parsing XML, it is much slower to load and parse than a simpler format such as KRSS.

One work-around for this, and something that would greatly benefit the e-health community, would be for the International Health Terminology Standards Development Organisation, the newly formed governing body of SNOMED, to formally publish URIs for the concepts in SNOMED. This would allow tool vendors to "bake in" SNOMED to their tools, while still allowing other OWL-based ontologies to reference SNOMED concepts in a consistent and interoperable manner in order to describe extensions to SNOMED.

CONCLUSION

Our preliminary work on producing local extensions to SNOMED for semantic data integration is promising as is the performance of our classifier. The current implementation is single-threaded and we anticipate a further speed increase from a multi-threaded implementation running on a multi-core CPU.

We are currently integrating snorocket with a 3rd-party SNOMED editing tool which requires specific support for SNOMED's use of role grouping and the ability to distinguish between stated and inferred relationships in the output of the classifier, although this adds little overhead to the classification time. In addition, we are prototyping mapping tools specifically targeting the task of constructing local extensions of SNOMED from existing data.

Finally, we are continuing work on our incremental form of the algorithm but have not yet tuned or verified the implementation. Preliminary results indicate that this approach should be very useable when integrated with our mapping tool.

Acknowledgements

The work described in this paper was carried out while on secondment to the CSIRO's E-Health Research Centre and the author would like to gratefully acknowledge the support of David Hansen and the other members of the Health Data Integration team.

Address for Correspondence

Michael J. Lawley, Faculty of Information Technology, University of Queensland, 126 Margaret Street Brisbane Qld 4000, Australia
m.lawley@qut.edu.au

References

[1] D. Hansen, C. Daly, K. Harrop, M. O'Dwyer, C. Pang, and J. Ryan-Brown. HDI: Research Software To Commercial Product. *ASWEC 2005 Industry Experience Papers*, 2005.

[2] I. Jakobsen, M.J. Lawley, A. Zankl, and D. Hansen. Ontologies for Skeletal Dysplasias. *MedInfo 2007 Workshop: MedSemWeb 2007*, 2007.

[3] SNOMED *Clinical Terms*. College of American Pathologists, 2006. http://www.snomed.org.

[4] F. Baader, C. Lutz, and B. Suntisrivaraporn. CEL—a polynomial-time reasoner for life science ontologies. In U. Furbach and N. Shankar, editors, *Proceedings of the 3rd International Joint Conference on Automated Reasoning (IJCAR'06)*, volume 4130 of *Lecture Notes in Artificial Intelligence*, pages 287–291. Springer-Verlag, 2006.

[5] F. Baader, C. Lutz, and B. Suntisrivaraporn. Efficient Reasoning in $\mathcal{EL}^+$. In *Proceedings of the 2006 International Workshop on Description Logics (DL2006)*, CEUR-WS, 2006. http://lat.inf.tu-dresden.de/research/papers/2006/BaaLutSun-DL-06.pdf.

[6] H. Wache, T. Vögele, U. Visser, H. Stuckenschmidt, G. Schuster, H. Neumann, and S. Hübner. Ontology-based integration of information-a survey of existing approaches. *IJCAI-01 Workshop: Ontologies and Information Sharing*, 2001:108–117, 2001.

[7] H. Stuckenschmidt. *Catalogue Integration: A Case Study in Ontology-based Semantic Translation.* Vrije Universiteit, Faculteit der Exacte Wetenschappen, Divisie Wiskunde & Informatica, 2000.

[8] Dieter Fensel, Ian Horrocks, Frank van Harmelen, Deborah L. McGuinness, and Peter F. Patel-Schneider. OIL: An Ontology Infrastructure for the Semantic Web. *IEEE Intelligent Systems*, 16(2), 2001.

[9] G. Wade and S.T. Rosenbloom. Experiences Mapping a Legacy Interface Terminology to SNOMED CT. In *Proceedings of the SMCS 2006 - Semantic Mining Conference on SNOMED CT*, 2006. http://www.hiww.org/smcs2006/proceedings/9WadeSMCS2006final.pdf.

[10] S. Ghilardi, C. Lutz, and F. Wolter. Did I Damage my Ontology? A Case for

Conservative Extensions in Description Logics. In Patrick Doherty, John Mylopoulos, and Christopher Welty, editors, *Proceedings of the Tenth International Conference on Principles of Knowledge Representation and Reasoning (KR'06)*, pages 187–197. AAAI Press, 2006. `http://lat.inf.tu-dresden.de/~clu/papers/archive/kr06a.pdf`.

[11] *International Statistical Classification of Diseases and Related Health Problems, Tenth Revision, Australian Modification (ICD-10-AM).* National Centre for Classification in Health, 5th edition, 2006. `http://www3.fhs.usyd.edu.au/ncch/4.2.1.1.htm`.

[12] Boontawee Suntisrivaraporn. Module extraction and incremental classification: A pragmatic approach for $\mathcal{EL}^+$ ontologies. In Sean Bechhofer, Manfred Hauswirth, Joerg Hoffmann, and Manolis Koubarakis, editors, *Proceedings of the 5th European Semantic Web Conference (ESWC'08)*, Lecture Notes in Computer Science. Springer-Verlag, 2008. To appear.

[13] C. Hall and J. Washbrook. Radiological Atlas of Malformation Syndromes and Skeletal Dysplasias (REAMS) [software]. Oxford University Press, CD-ROM, 1999.

[14] B. Motik, P.F. Patel-Schneider, and I. Horrocks. OWL 1.1 Web Ontology Language. World Wide Web Consortium, W3C Member Submission, 2006. `http://www.w3.org/Submission/2006/SUBM-owl11-owl_specification-20061219/`.

[15] K. Spackman. An Examination of OWL and the Requirements of a Large Health Care Terminology. In *Proceedings of the OWL: Experiences and Directions Third International Workshop (OWLED 2007), CEURWS*, June 2007. `http://owled2007.iut-velizy.uvsq.fr/PapersPDF/submission_26.pdf`.

Why do it the hard way? The Case for an Expressive Description Logic for SNOMED

Alan Rector, Sebastian Brandt
School of Computer Science, University of Manchester, Manchester M13 9PL,
(rector | brandt @cs.manchester.ac.uk

Since SNOMED-RT/CT was originally formulated in the early to mid 1990s, there have been major developments in logic-based formalisms, ontology design and associated tools. Combined with the increase in computing power in the past two decades, these developments mean that many of the restrictions that limited SNOMED's original formulation and schemas no longer need apply. We contend that future development of SNOMED would be made easier if a more expressive formalism and more modern tools were adopted.

The difficulties in the existing structure of SNOMED have been well documented. For example, Bodenreider (1) examined the specialization hierarchy of SNOMED classes. Schulz discussed 'relationship groups' (2) and a broad range of other ontological problems along with potential remedies (3). Schulz suggested a modest extension of SNOMED's formalism to one with more clearly defined semantics (EL+) but which still lacks true negation and disjunction. We argue here that judicious use of a more expressive language, OWL 1.1[1], is now practical and would bring great benefits including:

- A uniform, clear and understandable schema for all concepts used in clinical records, including context and negation.
- Elimination of the need for special mechanisms to deal with context, partonomy, and role groups.
- More effective leveraging of the underlying logical representation to organise and quality assure the SNOMED hierarchies.
- Improved ability to recognise semantic equivalence between post-coordinated and pre-coordinated expressions and between "observables" with "values" and the corresponding "findings."
- Improved ability to modularise and segment SNOMED for specific purposes
- Access to the tools and techniques being developed by the wider Semantic Web and OWL communities.

In outline, the proposals are:

- To represent all concepts used in clinical records (findings, observables, and procedures) uniformly as fully defined "situations" that include any context required and that deal with negation explicitly and formally.
- To represent all sites explicitly as to whether they refer to the site in its entirety or to the disjunction of the site and its parts.
- To define observables and related findings in such a way that the classifier can be used to recognise the equivalence between a situation involving an observable with a given value and the corresponding finding of the observable with that value – *e.g.*, between an observable of "blood pressure" qualified by "elevated" and a finding of "elevated blood pressure".
- To organise the stated form as a set of modules that can be separated for specific applications.

Details of the proposed mechanisms are described in the extended version of this paper and in (4, 5).

Although the effort to migrate any large software object should not be underestimated, most of the proposed changes would cause few changes to the schemas except for "Situations with specific context," which are known to be problematic. (However, the proposed analysis would identify many errors to be corrected.) The effort would be more than repaid by providing a more regular and consistent system that would improve usability and simplify software development and query formulation. We argue that a feasibility study using a modest subset of around 25K concepts should be an urgent priority for the SNOMED community.

References

1. Bodenreider O, Smith B, Kumar A, Burgun A. Investigating subsumption in SNOMED CT: An exploration into large description logic-based biomedical terminologies. *AI in Medicine*. 2007;39:183-195.
2. Schulz S, Hanser S, Hahn U, Rogers J. The semantics of procedures and diseases in SNOMED CT. *Meth Inf Med*. 2006;45:354-358.
3. Schulz S, Suntisrivaraporn B, Baader F; SNOMED CT's Problem List: Ontologists' and logicians' therapy suggestions.; Medinfo 2007: IOS Press; 802-806.
4. Rector A, Qamar R, Marley T; Binding ontologies & coding systems to electronic health records and messages. 2006; Formal Biomedical Knowledge Representation (KR-MED 2006
5. Rector AL; What's in a code: Towards a formal account of the relation of ontologies and coding systems. 2007; Medinfo 2007: Brisbane, Australia: IOS Press; 730-734.

[1] http://www.webont.org/owl/1.1/

Representing and sharing knowledge using SNOMED
Proceedings of the 3rd international conference on Knowledge Representation in Medicine (KR-MED 2008)
R. Cornet, K.A. Spackman (Eds)

Leveraging SNOMED CT with a General Purpose Terminology Server

R. Weida, PhD, J. Bowie, ScD, R. McClure, MD, D. Sperzel, MD
Apelon, Ridgefield, CT, USA
weida@apelon.com

General purpose terminology server software facilitates coordinated use of multiple standard medical terminologies for diverse healthcare applications. SNOMED CT is an important clinical reference terminology, whose size and scope make advanced terminology server capabilities particularly useful. Moreover, capabilities tied to SNOMED CT's special features and requirements can result in substantial further benefits. Enhancements to a general purpose terminology server have been developed to facilitate the tailored creation, validation, organization, deployment, distribution, submission and maintenance of (post-coordinated) extensions to SNOMED CT.

INTRODUCTION

Standard medical terminologies are vital to all sorts of contemporary healthcare information technology endeavors, ranging from encoding and exchanging information in electronic health record (EHR) systems to facilitating outcomes analysis and decision support. However, effective integration of terminologies into clinical applications poses substantial challenges. These applications generally require multiple terminologies since each terminology has been designed for different purposes by different healthcare constituencies, e.g., SNOMED CT for representation of clinical data; ICD-9-CM, ICD-10-CM and CPT-4 for reimbursement; LOINC for laboratory test results; and HL7 for application interfaces. Drug nomenclatures such as RxNorm and NDF-RT, device taxonomies such as UMDNS, specialty ontologies, and others are also important, as are enterprise-specific terminology enhancements. Terminologies employ different data models and they are delivered in different data formats. Finally, terminologies are constantly evolving, so they must be regularly updated in clinical and other applications. However, revision schedules and processes vary widely and are often inconsistent. Such challenges can be effectively met with a comprehensive, general purpose *terminology server,* defined as a networked software component that centralizes and integrates terminology content and reasoning to provide (complete, consistent, effective) terminology services for users and other network applications. Earlier terminology servers[1,2,3,4] did not provide the modular classification, subset, template or SNOMED-specific features described here.

Terminology servers support diverse applications. For example, they are used by informaticists to create, maintain, localize and map terminologies; by clinical applications and their users to select and record standardized data; and by software integration engines to map data elements between applications. SNOMED CT is of special interest due to its broad clinical scope, extensive detail, formal structure, and international standing.[5,6] This paper describes some ways that one general purpose terminology server has been enhanced and applied to support SNOMED CT within the context of a full complement of other healthcare terminologies.

DISTRIBUTED TERMINOLOGY SYSTEM

Apelon's Distributed Terminology System (DTS) is an open source terminology software suite whose key component is a terminology server. DTS is robust and mature, benefiting from years of production deployment in diverse healthcare industry settings. It has been used by software and content vendors, pharmaceutical companies, government agencies, universities and research institutions, healthcare delivery systems, and standards development organizations around the world.

DTS Architecture

DTS employs typical three-tier architecture, as illustrated in Figure 1. Multi-tier architectures offer many well known advantages, including the ability to support highly flexible, easily scalable, and extremely dependable deployment solutions.

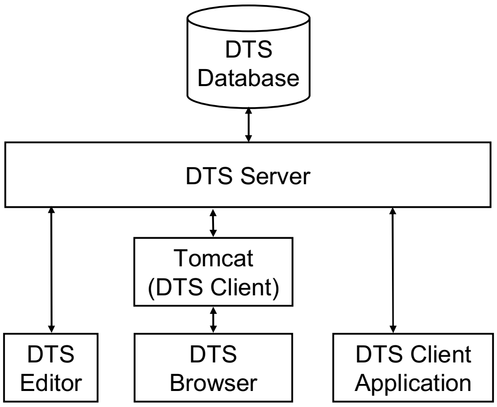


Figure 1 – DTS Architecture.

The DTS client tier (below the DTS Server in Figure 1), provides both Java and .Net APIs for developing custom terminology applications. DTS comes with packaged client applications such as an extensible desktop (fat client) terminology editor, the DTS Editor. There is also a web-based (thin client) terminology browser, the DTS Browser, which requires an Internet browser and an intermediary Apache Tomcat (or equivalent) web server. The middle tier of DTS consists of the DTS Server, a terminology-focused application server which supports highly concurrent, authenticated access to terminology services via the APIs. It features numerous performance optimizations, logging, tracing, remote monitoring, etc. The APIs support browsing, navigation, search, query, editing, localization, mapping, subsetting and other common terminology operations. A relational database comprises the third – or data – tier of DTS, shown at the top of Figure 1. In addition, DTS supplies various utilities for software and content management, including content subscription updates. Readers interested in DTS features outside the scope of this paper are referred to the DTS White Paper[7].

DTS Namespaces

DTS employs a unified content model for uniform access to diverse terminologies, including ones based on Description Logic (DL) such as SNOMED CT, the NCI Thesaurus and NDF-RT, as well as non-DL terminologies like CPT, ICD, and LOINC. A subscription service is available for all major medical terminologies (plus cross-terminology mappings) formatted for easy loading into DTS, ensuring that the latest versions of the terminologies are always available. A DTS *namespace* is the unit of management for content delivery (and access control). Thus, each standard terminology resides in a separate namespace so it can be independently updated and versioned. A *mapping* between (elements of) a pair of terminologies, e.g., from CPT to SNOMED CT, is also typically delivered in its own separate namespace. DTS also supports an unlimited number of *local namespaces* enabling users to create and maintain user- or organization-specific terminology data. These local terminologies are also housed in distinct namespaces, as are the local extensions to standard terminologies described below.

DESCRIPTION LOGIC

Description Logic (DL) is a well known field of study within the area of knowledge representation.[8] DL is a type of formal logic focused on creating definitions of concepts and reasoning about them effectively. Thus, DL is well suited for expressing precise descriptions of medical concepts, including anatomy, diseases, drugs, procedures, and so on. DL enables clear and unambiguous *formal definition* of a concept's meaning, primarily in terms of its relationships with other concepts. A given concept (e.g., representing a class of drugs) can be described succinctly by naming the concepts it specializes (more general classes of drugs) and introducing distinguishing characteristics (e.g., relationships to its ingredients). The logical consistency of an entire set of concepts, such as those comprising a medical terminology, is automatically tested and enforced. Moreover, logical consequences that are implicit in the given descriptions are automatically made explicit.

A particular DL provides a language for describing concepts and a repertoire of logical inferences for reasoning about them. SNOMED CT uses the Ontylog DL[9], which is also used for the US Veterans Health Administration's NDF-RT (National Drug File – Reference Terminology) and the National Cancer Institute's NCI Thesaurus, all standards of the US Government's Consolidated Health Informatics (CHI) Initiative[10]. Ontylog syntax and semantics have been published in connection with the NCI Thesaurus.[11] Among the most powerful aspects of DL are its facilities for reasoning about relationships among concepts and thus automatically managing a logically consistent taxonomy (i.e., generalization hierarchy or "is-a" hierarchy) of concepts.

The DL *classification* operation automatically organizes concepts into a taxonomy based on their logical descriptions. Software that implements classification is called a *classifier*. As a simplified expository example, a set of concepts { *A, B, C, D, E, F, G, H, I, J* } might be classified into the taxonomy shown in the top portion of Figure 2, where *A* is a generalization of *B*, *C* and *D*; *B* is a generalization of *E*, *F* and *G*, etc. We will use this taxonomy in subsequent examples. Extant classifiers generally create an explicit representation of a taxonomy, including explicit information corresponding to each of the lines shown between pairs of linked concepts. The Apelon classifier generates a very high performance, in-memory "classification graph" which includes all information necessary to continue classifying additional concepts in the future.

As a result of classification, each concept in the taxonomy is guaranteed to be more specific than its parents and all other ancestors (directly or indirectly connected concepts above), as well as more general than its children and all other descendants (directly or indirectly connected concepts below). Therefore, concepts are always found in predictable locations. That makes it easier to envision relationships among concepts and to recognize unintended results. Well-organized taxonomies allow medical knowledge (e.g., advice, rules, warnings, arbitrary codes, etc.) to be associated with concepts at the most appropriate level in the taxonomy (neither too general nor too specific) and appropriately inherited by (implicitly associated with) descendant concepts.

A *terminology* is a collection of presumably related concepts. In DTS, a namespace is a set of concepts that are managed as a group. Thus, one can classify the set of concepts comprising a namespace into a taxonomy. Ordinarily, an entire terminology is contained – and thereby managed – in one namespace, e.g., all the concepts shown in the top portion of Figure 2 might comprise a single namespace. (For authoring purposes, some DLs allow terminologies to be composed by "importing" (the concepts of) one terminology into another, but the entire result is still classified monolithically.)

MODULAR EXTENSION

DTS terminology extension features are motivated largely by the existence of SNOMED CT and the desire of users to adapt it in diverse ways. SNOMED CT contains hundreds of thousands of concepts. New versions of SNOMED have been released twice yearly. Many different users (persons or organizations) may wish to extend SNOMED by adding their own concepts. The SNOMED data model provides for this possibility. Indeed a single user may be interested in extending SNOMED several different ways. However, it is important to clearly distinguish the authoritatively published core of SNOMED from any extensions thereof. Furthermore, it is important to classify terminology extensions, including post-coordinated expressions, as rapidly as possible. Traditional classifiers organize an entire set of concepts into a taxonomy by "starting from scratch" and classifying (processing) each and every concept in turn.

Modular Classification

DTS uniquely facilitates multiple independent extensions of a concept taxonomy based on DL. Separate classification operations determine how one or more distinct sets of additional concepts, each comprising an extension, fit in with the original taxonomy while leaving the original taxonomy intact and without copying it. Classification results are recorded so that the original taxonomy as well as every extension thereof can be independently browsed, searched, queried and retrieved on demand. As a result, DL taxonomies such as SNOMED CT can be extended easily and accurately, using the same language as the original, in multiple independent ways, to meet local and/or specialized needs in a timely manner. We call this process *modular classification.* Thus, DTS introduces effective means for working with multiple independent extensions of an existing taxonomy while preserving the integrity of the original. Indeed, DTS uses the same classification software used in the creation of SNOMED CT.

We will refer to an existing, self-contained namespace, e.g., a namespace containing SNOMED CT, as a *base namespace*. Concepts therein are referred to as *base concepts*. Then, an *extension namespace* contains one or more additional concepts to be classified, viewed, and otherwise used *as if* they were also part of the base namespace, but without altering and without copying the base namespace. Concepts within an extension namespace are referred to as *extension concepts*.

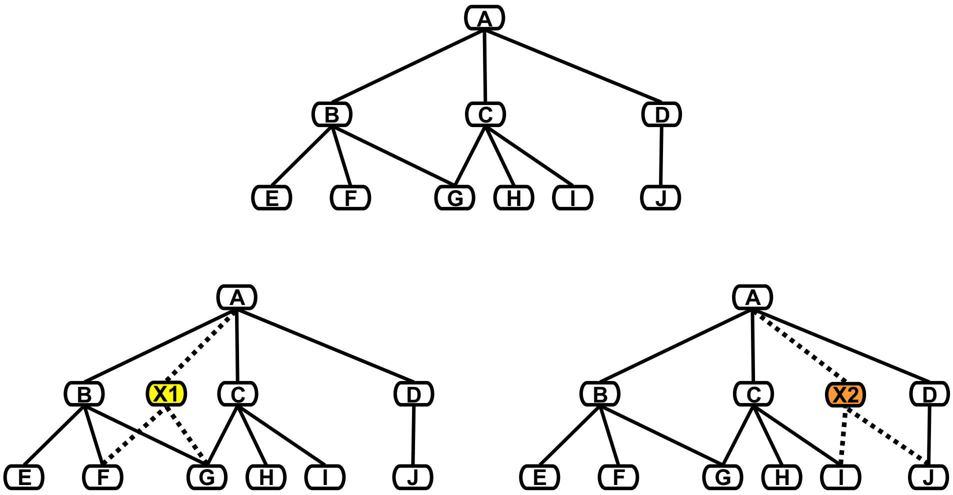


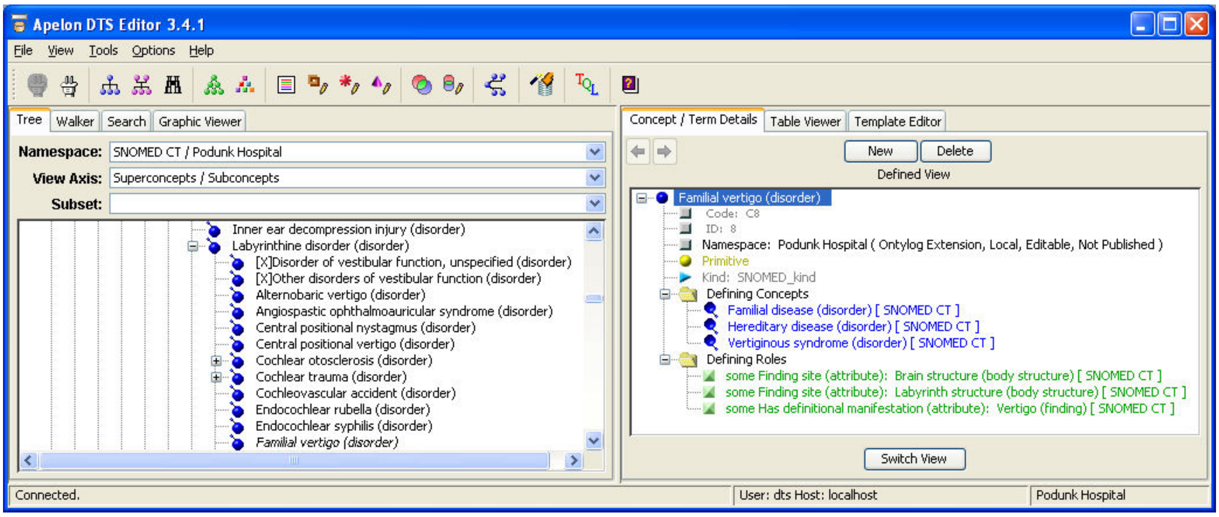
Figure 2 - Base Namespace Taxonomy (top) with Multiple Independent Extended Taxonomies (bottom).

The modular classifier operates on DL elements of SNOMED extension concepts defined in extension namespaces. These concepts are linked by SNOMED relationships to other concepts in the base namespace and/or the same extension namespace. DTS extension namespaces can also contain other local information about core SNOMED concepts. Examples include additional local synonyms; local associations connecting them to or from other concepts, e.g., to represent mappings from a local terminology; and local properties (attribute value pairs, e.g., to indicate that a procedure is performed locally, or that a certain person last edited the concept). In all cases, extensions to SNOMED CT could become problematic if a base SNOMED concept is later retired. Reports detailing any such connections are available, thus allowing for remediation.

As an example, the fictitious Podunk Hospital may wish to extend the *SNOMED CT* base namespace with a *Podunk Hospital* extension namespace. That extension namespace may include an extension concept for a disorder, *Familial vertigo*, with definitional relationships to several base concepts in *SNOMED CT***.** In general, an extension concept can be defined in terms of its relationships to base concept(s) and/or fellow extension concept(s). The user's definition of *Familial vertigo* is shown on the right in Figure 3. This definition was created interactively within the DTS Editor, drawing from concepts and relationships (roles) in the standard SNOMED CT Namespace. Following modular classification, the position of *Familial Vertigo* with respect to one branch of the SNOMED taxonomy is shown on the left. Of note, the classifier has inferred the position of *Familial Vertigo* directly under a concept *Labyrinthine disorder* not mentioned explicitly in its definition. The DTS Editor italicizes extension concepts in the context of a base namespace for emphasis.

In the interest of clarity and brevity (SNOMED CT has hundreds of thousands of concepts), the upper portion of Figure 2 shows a much simpler sample taxonomy for a base namespace. Beneath that are two independent extensions, one where the taxonomy is extended with a namespace consisting of the concept *X1*, and another where the taxonomy is extended with a namespace consisting of the concept *X2*. Notice that an extended taxonomy effectively contains the entire set of concepts from the base namespace augmented with additional concept(s) from the extension namespace. The dashed lines are intended to suggest that while the relationships of the extension concepts to the base taxonomy have been determined, they are not (destructively) spliced into the original base taxonomy (shown with solid lines). While these simple illustrations show only one concept per extension, an extension can of course contain an arbitrary number of concepts. We have used the DTS modular classifier with an extension namespace that (experimentally) extends SNOMED with LOINC laboratory concepts, and another extension namespace containing the US Drug Extension[12], each containing well over 15,000 concepts.

A base namespace may have multiple extensions which depend on it; extensions are mutually independent. Multiple independent namespaces extending SNOMED CT might have a variety of custodians and purposes, including a person (for learning and testing), a project (for research and development), an organization (for specific institutional needs), a specialty society (for terminology related to their practice area), national



Namespace: SNOMED CT / Podunk Hospital
View Axis: Superconcepts / Subconcepts
Subset:

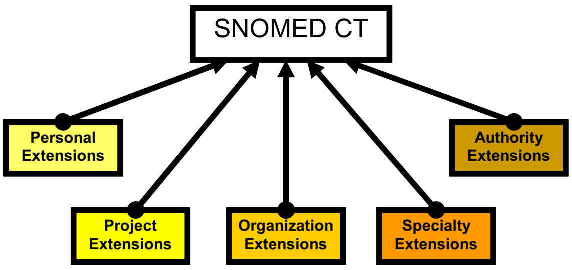
Inner ear decompression injury (disorder)
Labyrinthine disorder (disorder)
[X]Disorder of vestibular function, unspecified (disorder)
[X]Other disorders of vestibular function (disorder)
Alternobaric vertigo (disorder)
Angiospastic ophthalmoauricular syndrome (disorder)
Central positional nystagmus (disorder)
Central positional vertigo (disorder)
Cochlear otosclerosis (disorder)
Cochlear trauma (disorder)
Cochleovascular accident (disorder)
Endocochlear rubella (disorder)
Endocochlear syphilis (disorder)
Familial vertigo (disorder)

Familial vertigo (disorder)
Code: C8
ID: 8
Namespace: Podunk Hospital (Ontylog Extension, Local, Editable, Not Published)
Primitive
Kind: SNOMED_kind
Defining Concepts
Familial disease (disorder) [SNOMED CT]
Hereditary disease (disorder) [SNOMED CT]
Vertiginous syndrome (disorder) [SNOMED CT]
Defining Roles
some Finding site (attribute): Brain structure (body structure) [SNOMED CT]
some Finding site (attribute): Labyrinth structure (body structure) [SNOMED CT]
some Has definitional manifestation (attribute): Vertigo (finding) [SNOMED CT]

Connected.
User: dts Host: localhost
Podunk Hospital

Figure 3 – DTS Editor with Extension Concept (right) and Extended Taxonomy (left).

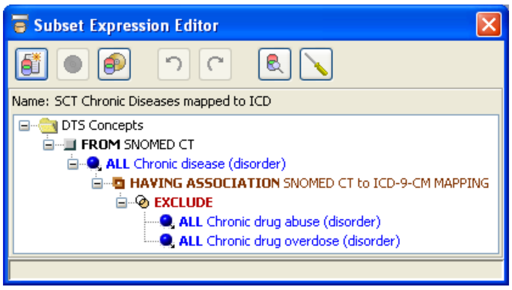
authorities, or even the creators of the base namespace themselves (e.g., to preview possible future enhancements to the base):



So far, we have focused on authoring sets of concepts covering a unified extension of interest. However, it is important to note that modular classification is equally adept at "on the fly" post-coordination of new concepts in accord with the SNOMED model, e.g., to help populate EHRs at run-time using the DTS API. Logical equivalence (hence redundancy) with a base concept or another extension concept is always detected and reported by the modular classifier.

SUBSETS

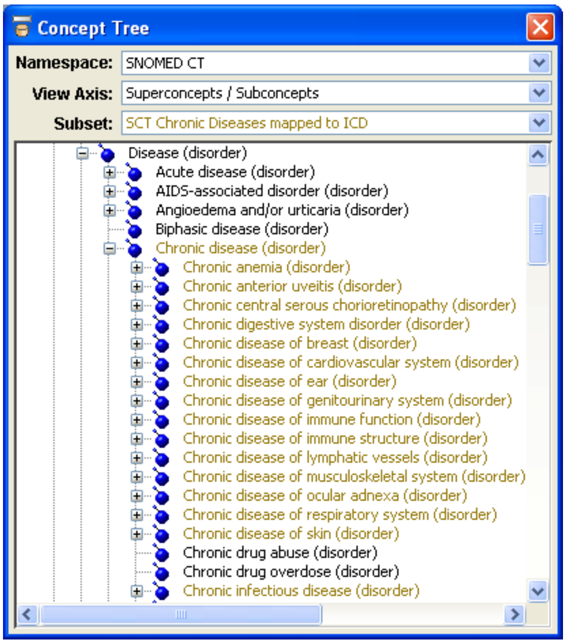
Considering the large size and broad scope of SNOMED CT and other contemporary medical terminologies, it can be extremely helpful to work with smaller, more focused subsets of terminologies when populating pick lists in EHR systems or fields in HL7 messages (HL7 value sets), constraining searches to pertinent concepts for data matching and analysis, etc. Subsets of interest can themselves be large and therefore challenging to maintain when the underlying terminologies are revised, e.g., concepts that are members of the subset may be retired and new concepts that should become members may be introduced. Enumerating each element of a large subset is tedious, opaque and often highly inefficient. Therefore, DTS takes a constructive approach to subset specification: a concise *subset expression* compositionally defines an arbitrary subset by specifying member concepts according to their names, synonyms, other properties, and relationships. Subset expressions can specify inclusion or exclusion of identified concepts and/or all of their descendants in the (base or extended) taxonomy. Moreover, subset expressions can be arbitrarily nested to include sub-taxonomies, exclude portions thereof, etc. Subset expressions can use various concept attributes, even those that refer to other namespaces, e.g., we can specify all SNOMED chronic diseases mapped to ICD-9-CM but (strictly for illustration) excluding chronic drug abuse and chronic drug overdose:



Name: SCT Chronic Diseases mapped to ICD

```
DTS Concepts
  FROM SNOMED CT
    ALL Chronic disease (disorder)
      HAVING ASSOCIATION SNOMED CT to ICD-9-CM MAPPING
        EXCLUDE
          ALL Chronic drug abuse (disorder)
          ALL Chronic drug overdose (disorder)
```

Visualization of subsets greatly aids review and revision. The DTS Editor (and likewise the web-based browser) can highlight subset members in the larger context of the entire SNOMED taxonomy; note the subset member concepts highlighted in gold:



Namespace: SNOMED CT
View Axis: Superconcepts / Subconcepts
Subset: SCT Chronic Diseases mapped to ICD

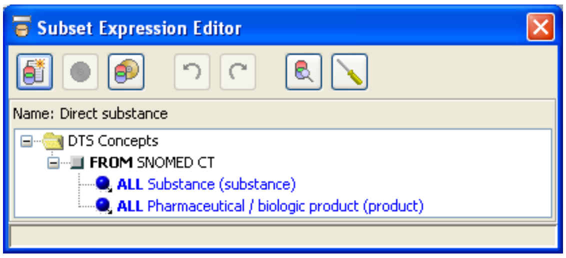
```
Disease (disorder)
  Acute disease (disorder)
  AIDS-associated disorder (disorder)
  Angioedema and/or urticaria (disorder)
  Biphasic disease (disorder)
  Chronic disease (disorder)
    Chronic anemia (disorder)
    Chronic anterior uveitis (disorder)
    Chronic central serous chorioretinopathy (disorder)
    Chronic digestive system disorder (disorder)
    Chronic disease of breast (disorder)
    Chronic disease of cardiovascular system (disorder)
    Chronic disease of ear (disorder)
    Chronic disease of genitourinary system (disorder)
    Chronic disease of immune function (disorder)
    Chronic disease of immune structure (disorder)
    Chronic disease of lymphatic vessels (disorder)
    Chronic disease of musculoskeletal system (disorder)
    Chronic disease of ocular adnexa (disorder)
    Chronic disease of respiratory system (disorder)
    Chronic disease of skin (disorder)
    Chronic drug abuse (disorder)
    Chronic drug overdose (disorder)
    Chronic infectious disease (disorder)
```

The DTS Editor can also render and browse the hierarchical structure of the subset members alone, just as if all non-members were spliced out of the original taxonomy (not shown for brevity). Of course, DTS can also enumerate and export subsets, test for subset membership, search within subsets, etc. All of these features are available in the DTS Editor GUI application and also via the DTS APIs for runtime application integration.

TEMPLATES

Since DTS is a general purpose system for arbitrary terminologies, the DTS Editor enables unconstrained editing using generic terminology constructs. However, the SNOMED model carefully constrains concept definitions. Particular types of concepts (within a particular SNOMED hierarchy) are to be defined using particular SNOMED relationships to target concepts chosen from particular portions of SNOMED. The DTS *Template Builder* (a DTS Editor "plug-in") has been developed to specify templates for context-dependent editing in compliance with such a model. Due to space restrictions, the following example is necessarily very abbreviated and simplified but conveys the gist.

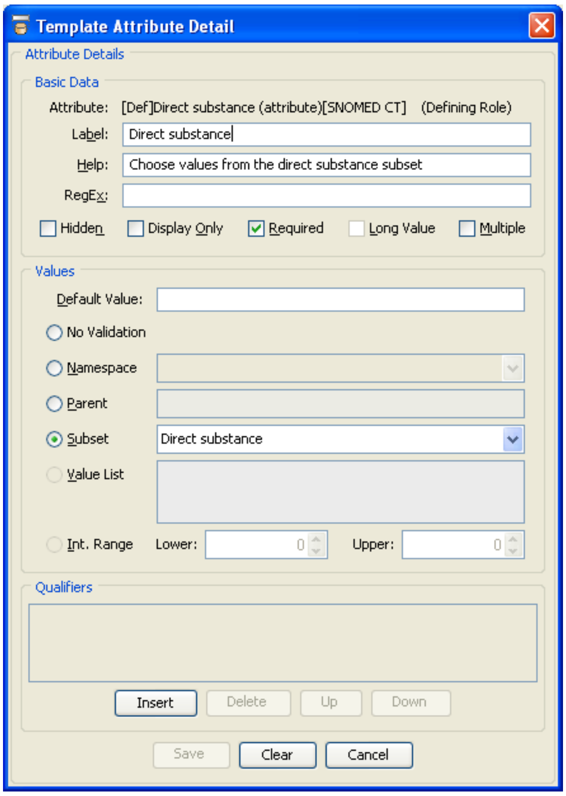
Suppose we need to extend SNOMED with more procedures. The *SNOMED CT Users Guide* specifies that the value of a *Direct substance* relationship (when present) on a *Procedure* concept should be a *Substance* or a *Pharmaceutical/biological product*. Thus, we create a *Direct substance* subset:



Name: Direct substance

- DTS Concepts
  - FROM SNOMED CT
    - ALL Substance (substance)
    - ALL Pharmaceutical / biologic product (product)

As we create a template for our procedures, we can require a value for the *Direct substance* relationship and require that it be restricted to members of our *Direct substance* subset:



Attribute: [Def]Direct substance (attribute)[SNOMED CT] (Defining Role)

Label: Direct substance

Help: Choose values from the direct substance subset

RegEx:

□ Hidden □ Display Only ☑ Required □ Long Value □ Multiple

Default Value:

○ No Validation

○ Namespace

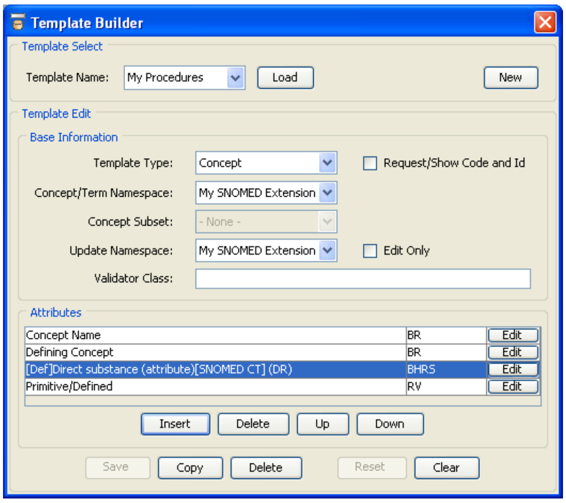
○ Parent

● Subset: Direct substance

○ Value List

○ Int. Range Lower: 0 Upper: 0

The *Direct substance* relationship is one attribute of an overall template for *My Procedures* (which are concepts in the *My SNOMED Extension* namespace):



Template Name: My Procedures

Template Type: Concept □ Request/Show Code and Id

Concept/Term Namespace: My SNOMED Extension

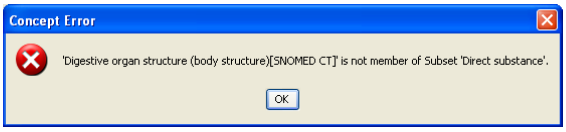
Concept Subset: - None -

Update Namespace: My SNOMED Extension □ Edit Only

Validator Class:

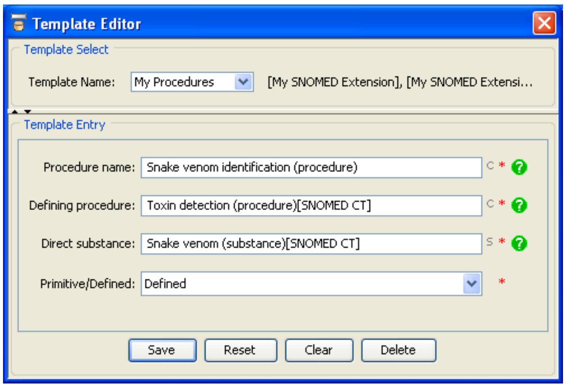
Concept Name	BR	Edit
Defining Concept	BR	Edit
[Def]Direct substance (attribute)[SNOMED CT] (DR)	BHRS	Edit
Primitive/Defined	RV	Edit

The DTS *Template Editor* enables creation and modification of concepts according to such templates. It reports an error if we attempt to use a concept that is not a member of the specified subset as the value for a *Direct substance* relationship:



'Digestive organ structure (body structure)[SNOMED CT]' is not member of Subset 'Direct substance'.

The Template Editor accepts a member of the subset, as in this definition of *Snake venom identification*:



Template Name: My Procedures [My SNOMED Extension], [My SNOMED Extensi...

Procedure name: Snake venom identification (procedure) C *

Defining procedure: Toxin detection (procedure)[SNOMED CT] C *

Direct substance: Snake venom (substance)[SNOMED CT] S *

Primitive/Defined: Defined *

Notice the template-specific labels: *Procedure name, Defining procedure* and *Direct substance*. Absent any intervening extension concepts, the modular classifier will place this *Snake venom identification* extension concept directly under the *Toxin detection (procedure)* concept from the *SNOMED CT* base namespace.

DISTRIBUTION AND SUBMISSION

There are several ways to transfer terminology content into, out of and between DTS instances. Apelon distributes full and incremental versions of many standard (and custom) terminologies using a compact data format which closely corresponds to the DTS database schema and can therefore be loaded very efficiently. DTS enables users to distribute their own DTS terminology content in the same format. In addition, DTS includes graphical tools – the *import wizard* and the *export wizard* – to easily move ad hoc terminology content in and out of DTS using delimited text and XML formats. However, SNOMED CT has its own release format, consisting of a set of related files, tailored to the SNOMED data model, which specifically support SNOMED extensions. A SNOMED CT Identifier (SCTID) uniquely identifies all concepts, descriptions and relationships in SNOMED CT. Those who wish to extend SNOMED CT can request their own, exclusively assigned range of SNOMED CT identifiers. To facilitate creation and distribution of SNOMED extensions using DTS, we have implemented new DTS capabilities in collaboration with a national terminology authority and with a leading academic medical center. These capabilities include generation of SCTIDs for all elements of a SNOMED Extension namespace in DTS, as well as import and export of extension namespaces in SNOMED release format. Thus, SNOMED extensions can be readily shared with collaborators, and as appropriate, could be submitted for possible inclusion in the SNOMED core. The fact that these extensions have already been successfully classified together with the SNOMED core should expedite review and possible acceptance.

CONCLUSION

Apelon DTS, now available via open source licensing, has proven to be a popular tool for enterprise terminology asset management, featuring comprehensive capabilities for working with multiple standard and local terminologies, both individually and in concert (e.g., via mappings) using a unified suite of software components. Recognizing the importance of SNOMED CT, we have added significant functionality to meet SNOMED's unique requirements and benefit from its unique capabilities.

Acknowledgments

We gratefully acknowledge the contributions of past and present Apelon colleagues to the ideas described in this paper and the DTS system. We deeply appreciate review and feedback on new SNOMED capabilities in DTS from Dr. James Campbell at the University of Nebraska Medical Center and from Australia's National E-Health Transition Authority.

References

1. Rector AL, Solomon WD, Nowlan WA, Rush TW, Zanstra PE, Claassen WM. A terminology server for medical language and medical information systems. Meth Inform Med. 1995, 34(1-2) p. 147-57.
2. Mays E, Weida R, Dionne R, Laker M, White B, Liang C, and Oles, FJ. Scalable and expressive medical terminologies. AMIA Annual Fall Symposium, 1996. p. 259-263.
3. Burgun A, Patrick D, Bodenreider O, Botti G, Delamarre D, Poulinquen B, Oberlin P, Leveque JM, Lukacs B, Kohler F, Fieschi M, LeBeux P. A web terminology server using UMLS for the description of medical procedures. J Am Med Inform Assoc. 1997; 4:356–363.
4. Chute CG, Elkin PL, Sherertz DD, Tuttle MS. Desiderata for a clinical terminology server. Proc AMIA Symp1999: 42-6.
5. Spackman KA, Campbell KE, and Cote RA. SNOMED RT: A reference terminology for health care. Proceedings of the AMIA Annual Fall Symposium, 1997. p. 640–644.
6. IHTSDO: SNOMED CT®. [Online]. 2007 [cited 2008 Jan 15]; Available from: http://www.ihtsdo.org/our-standards/snomed-ct/
7. Distributed Terminology System. [Online]. 2006 [cited 2008 Jan 15]; Available from: URL: http://www.apelon.com/products/white papers/DTS White Paper V34.pdf
8. Baader F, Calvanese D, McGuinness DL, Nardi D, and Patel-Schneider PF, editors. The description logic handbook: theory, implementation, and applications. Cambridge (U.K.): Cambridge University Press; 2003.
9. Spackman KA, Dionne R, Mays E, Weis J. Role grouping as an extension to the description logic of Ontylog motivated by concept modeling in SNOMED. AMIA Annual Symposium, 2002. p. 712-716.
10. Presidential Initiatives. [Online]. 2007 [cited 2008 Jan 15]; Available from: URL: http://www.hhs.gov/healthit/chiinitiative.html
11. Hartel FW, de Coronado S, Dionne R, Fragoso G, Golbeck J. Modeling a description logic vocabulary for cancer research. J Biomed Inform. 2005 Apr. p. 114-29.
12. CAP SNOMED Terminology Solutions. Pharmacy. [Online]. 2007 [cited 2008 Jan 15]; Available from: URL:http://www.cap.org/apps/docs/snomed/documents/pharmacy_nov07.pdf

Representing and sharing knowledge using SNOMED
Proceedings of the 3rd international conference on Knowledge Representation in Medicine (KR-MED 2008)
R. Cornet, K.A. Spackman (Eds)

LinKBase® and SNOMED: some distinct features and impact on NLP

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

ABSTRACT

In this paper a description is presented in which the architectural, lexical and mapping differences are foregrounded between two compositional systems, both operating in the health care domain: LinkBase® and SNOMED. Based on these distinctive features, repercussions on NLP applications are exemplified and briefly discussed.

1. INTRODUCTION

The use of an ontology as a resource to access and aggregate several different types of medical data for a range of purposes inside healthcare information systems has demonstrated significant advantages. Nevertheless, this very variability of the healthcare information to be reconciled within and across different healthcare organizations, as well as the diversity of information systems accessing this information, imposes a challenge on the identification of the ontology of choice. This document intends to describe a direct comparison between two medical domain ontologies, The Systemized Nomenclature of Medicine (SNOMED)[1,2,3] and LinKBase®[4], from the perspective of their applicability to healthcare information systems and their embedded requirements. This comparison focuses on structural and lexical aspects, as well as differences in mapping methodology.

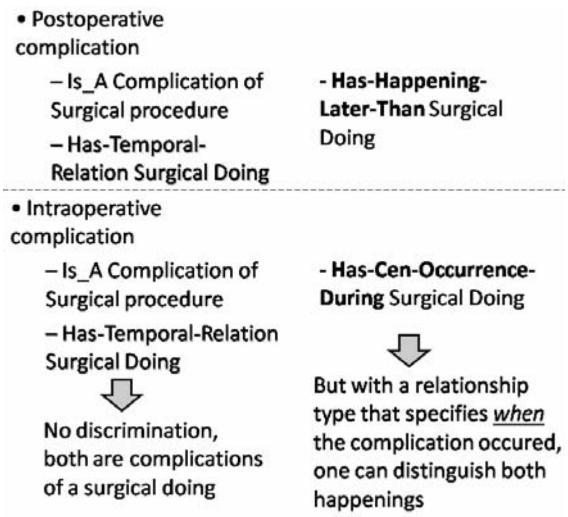
2. STRUCTURAL DIFFERENCES

2.1 Relationships and the principles behind them

SNOMED and LinKBase® are compositional systems: ontologies in which concepts can be specialized through combinations with other concepts. Both are based on Description Logics and contain binary relationships that interconnect the concepts. To enable semantic reasoning, a consistent meaning of the relationships is indispensable.

2.1.1 Relationships in LinKBase®

Concepts in LinKBase® are interrelated by a set of 383 relationship types that are structured in a multi-parented hierarchy, in which both the formal realistic ontological implications and the linguistic aspects of the relationships are taken into account. Most relationships are based on theories[5], that deal with topics such as mereology and topology[6,7], time and causality[8] and models for semantics driven natural language understanding[9, 10]. The large set of relationship types allows LinKBase® to define the sometimes subtle semantic differences between concepts (figure 1).



- Postoperative complication
  - Is_A Complication of Surgical procedure
  - Has-Temporal-Relation Surgical Doing

- **Has-Happening-Later-Than** Surgical Doing

- Intraoperative complication
  - Is_A Complication of Surgical procedure
  - Has-Temporal-Relation Surgical Doing

No discrimination, both are complications of a surgical doing

- **Has-Cen-Occurrence-During** Surgical Doing

But with a relationship type that specifies *when* the complication occured, one can distinguish both happenings

Figure 1- Detailed relationships types are needed to define subtle semantic differences
A large set of relationship types allows definition of subtle semantic differences as for example the relationship to time, which is needed to discriminate an intraoperative with a postoperative complication

In LinKBase®, consistency is maintained throughout the entire system for all types of relations by enforcement of an ontological principle named the "Principle of 100 % true relationships"[4]. According to this principle concepts can only receive a specific relationship if that relationship is true for *all* subclasses and instances of that concept. For example, in LinKBase®, "meningitis" will never be a subclass of "infectious disease" since it is not always the result of an infection. The concept "infectious meningitis”, on the other hand, is always caused by an infection and is a subclass of both “infectious disease” and “meningitis”.

2.1.2 Relationships in SNOMED

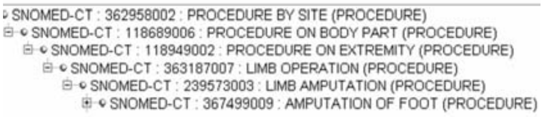
The set of relationship types in SNOMED is much smaller as compared to LinKBase®. SNOMED contains 50 relationship types and although a more restrictive, smaller set of relationship types might yield an easier to manage and utilize ontology for a user, a more granular set of relationship types allows the introduction of unique formal definitions to a much larger set of concepts in the ontology. For example in SNOMED the concepts ‘intraoperative care’ and ‘postoperative care’ cannot be defined, because there are no relationships which are specific enough to relate two procedures in different time aspects. In LinKBase® on the other hand these concepts can be defined by adding a temporal relation to the concept ‘surgical procedure’, the former relation being ‘occurs-during’ and the latter ‘occurs-after’.

The SNOMED relationship types are divided into three types: defining, refining and additional relationship types. Only the former type ‘defining’ are used to insert truly ontological information used, as its name states, to logically define concepts. ‘Refining’ relationships on the other hand can be seen more as application supporting artifacts which allow concepts to be connected to ‘qualifiers’ in specific instances[11]. For example, the concept ‘pneumonia’ contains the defining link *is-a* to ‘lung disorder’ and the refining link *clinical course* to ‘courses’. This latter relation allows for the relating of specific instances of pneumonia to one of the qualifier values under ‘courses’ as for example ‘acute’ or ‘chronic’. As in the example just described, for most concepts, the refining link inside the ontology/terminology itself, is an empty one. To the concept ‘pneumonia’, *clinical course* ‘courses’ does not add any useful information except that ‘pneumonia’ can have a ‘course’.

In order to cope with this distinction on the essence and use of relationship types SNOMED divides them in the three categories mentioned above. Also, when creating logical formal definitions for its concepts, only the ontologically based relations (‘defining’) are allowed to be used (see section 2.3). Although this restriction allows for the co-existence of both ontological and non-ontological information in the same syntax (i.e. binary relations), it does introduce problems when a refining characteristic becomes definitional for a given concept. For example the concept “acute inflammation” should ontologically be defined by being an ‘inflammation’ which has an ‘acute’ course. This becomes impossible given the ‘refining’ characteristic of the relationship type ‘*clinical course*’. The compositional model of SNOMED as such becomes limited to the specific boundaries of its relationship types and their characteristics. This impacts reasoning capabilities (see section 2.3) and applications which would rely on this resource such as decision support, run-time semantic interoperability (e.g. in messaging) and Natural Language Processing (NLP) and Natural Language Understanding (NLU) or more sophisticated Information Extraction.

SNOMED, just as LinKBase, aims at creating ‘defining’ relationships that are 100 % true. However, due to its structure and automated methods of hierarchy creation, inconsistencies are created (i.e. relationships that do not fulfil the “principle of 100 % true relationships”). A clear example is shown in figure 2 where ‘amputation of foot’ is incorrectly subsumed by ‘limb amputation’. These errors are the result of the SNOMED strategy to use Description Logic to create hierarchies based on other hierarchies. Although this is a valid method, the lack of specificity in relationship types creates a problem: by basing the procedure-hierarchy on the body part-hierarchy the ‘amputation of foot’ error was created.

The example in figure 2 is directly related to the fact that SNOMED has only one relationship type to relate a procedure with the procedure site: the relationship type ‘*procedure-site*’. According to formal ontological theories[12] there are several ways through which a given procedure may be related to a body part. The body part might be removed or placed during the procedure (e.g. excision or transplant), it might be altered structurally (e.g. incision), it might receive another structure (e.g. implant) etc. For each of these cases the participation of the body part is different, and only for some of these cases ontological reasoning properties, like transitivity, (see section 2.2) apply. As such, applying transitivity over all instances of the ‘*procedure-site*’ relationship creates hierarchical inconsistencies like the one described above. This occurs due to the fact that not all instances of this relationship type within SNOMED are in fact transitive. In order to make relationship generation via reasoning, either at production time to expand the ontology, or at run-time to connect information to the ontology, it is necessary to have



```
SNOMED-CT : 362958002 : PROCEDURE BY SITE (PROCEDURE)
⊟ SNOMED-CT : 118689006 : PROCEDURE ON BODY PART (PROCEDURE)
   ⊟ SNOMED-CT : 118949002 : PROCEDURE ON EXTREMITY (PROCEDURE)
      ⊟ SNOMED-CT : 363187007 : LIMB OPERATION (PROCEDURE)
         ⊟ SNOMED-CT : 239573003 : LIMB AMPUTATION (PROCEDURE)
            ⊞ SNOMED-CT : 367499009 : AMPUTATION OF FOOT (PROCEDURE)
```

Figure 2 - Incorrect subsumption in SNOMED
‘amputation of foot’ is incorrectly subsumed by ‘limb amputation’ since a foot is part of the limb, not the limb in total

relationship types logically defined and founded solely on sound ontological theory.

2.2 Top structure: Presence of Upper- and Mid-layer

Another structural aspect which determines an ontology's reasoning power and usability within specific information systems is the nature of its 'system of classification'. Here we refer to 'system of classification' when discoursing about the perspective into reality through which the concepts are organized or classified within the ontology. The 'system of classification' refers to the ontology's top structure and top classes which subsume the more granular ontological content. Ontological systems of classification may reside at three different layers: 1. The upper-layer, comprising of a small set of the most general classes formalized to be sufficient to described all that exists (examples of these top classes are 'process' (subsuming for example a 'hand-shake' or a 'surgical procedure') or 'property' (subsuming for example 'temperature')) , 2. The mid-layer, comprising of more specific classes, which are shared by different domains (for example 'temperature' (subsumed for example by 'fever' (medical domain) or 'atmosphere air temperature' (aviation domain), 3. The domain specific layer, comprising those classes specific to a given domain and organized according the perspective this domain takes at the world.

A domain ontology in information sciences is commonly a data model, holding a set of concepts and relationships between those concepts for a particular domain of interest, and represents or reflects 'reality' through that model of domain. It is used to reason about the objects within that domain. Both LinKBase® and SNOMED are medical domain ontologies, dealing with those aspects which make up the world of medicine. Formal ontology implies that the model is governed by strict logical (formal) axioms; in the case of LinKBase®, the mereological, axiomatic scheme is applied, which results in a structure characterized by: reflexivity (a concept A is part of itself), anti-symmetry (two distinct concepts cannot be part of each other) and, transitivity: if concept A has a transitive relation to concept B, and B has the same transitive relation to concept C, then A has also this same transitive relation to concept C.

The difference between the two resources in respect to this structural aspect pertains to the type of system of classification or top layer structure that each of them applies. Structurally SNOMED is a shared, health care classification system generated and applied by the medical domain and its actors. Its main branches (18 totally) and embedded top nodes or concepts are derived from a strictly medical classification perspective and reflect those entities through which the domain of medicine views and divides its realm (examples are Clinical Finding (main branch), Procedures (main branch), Body Structure (supporting branch), Substance (supporting branch), Organism (supporting branch), Context-Dependent Category (bridging branch), and Qualifier Value (bridging branch). LinKBase® applies an upper-layer and a mid-layer ontology at the top of its medical classification. The upper-layer is comprised of classes derived from the Basic Formal Ontology (BFO)[5], while the mid-layer is comprised of those generic domain unspecific concepts which connect the domain specific layer to the upper-layer concepts[4] (Examples of upper-layer concepts are (1) 'perdurants' (processes) or concepts with a time component, and (2) 'endurants' or concepts without a time component. Examples of mid-layer concepts are 'temperature' or 'movement'). From a usability perspective, the presence of an upper-and mid-layer influences on the degree of intelligence and specificity of logical reasoning and/or other process automation algorithms which can be applied over this ontology. Data integration and warehousing, semantic operability and NLP applications have the frequent need of mapping terminology (either in databases and terminological systems, data entry systems or free text documents) to the concepts in the ontology. Given the large degree of ambiguity in medical terminology, it is difficult to decide upon mappings between different concepts when performing terminological matching. The presence of an upper-layer ontology can strongly assist in this process. Figure 3 shows the example of an ambiguity which can be solved by combining language syntactical information with upper level ontological information: the term 'transposition' can either mean transposition as a surgical procedure (transposing a given part of the body to another part) (figure 3, panel A), or a transposition as a pathological structure (a part of the body which is constituently displaced) (figure 3 panel B). The fact that the term in the sentence in figure 3A functions as the object of the verb 'repair' which in its turn instantiates the concept 'surgical repair'; and that this concept is subsumed by the upper level concept 'process' which can only have 'substances'[13] as its participants, allows the deduction that the concept in question is the 'transposition' as a pathological structure due to the fact that this concept is subsumed by the upper-layer concept 'substance'. Opposed to the scenario above, in figure 3B the term transposition is related to the term 'technique' as if this latter was one of its ways through or one of its parts; this combined with the fact that a 'technique' is instantiated by processes and that processes can only have other processes as its parts, allows us to deduce that 'transposition' in this case must be a process and therefore disambiguate to the surgical procedure[14].

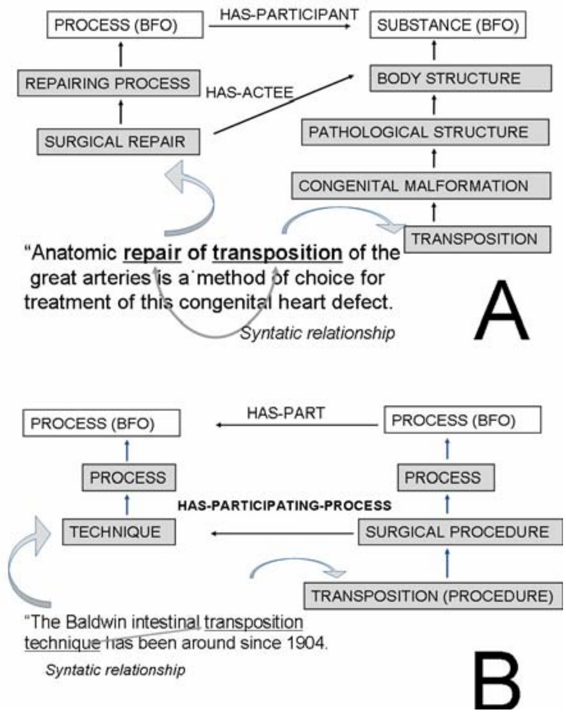


Figure 3 - Example of Ontology-based Terminological Disambiguation: ‘transposition’ as a congenital abnormality *See text for details*

Another consequence of the lack of an upper-and mid-layer structure is the need to create so called ‘ad hoc hierarchies’ to place those elements which do not pertain to any of the main domain classes, but which are nevertheless still important to represent the domain. An example of these is the main branch ‘Qualifier Value’ in the SNOMED system. The concepts under this hierarchy are only represented as values of an attribute or context used to represent other concepts or nodes inside SNOMED. As such nothing else can be understood or deduced about the essence and meaning of these concepts besides the fact that they can be used to ‘qualify’ other concepts. For example the concepts ‘entrance’ or ‘exit’ in SNOMED are classified only as ‘any hazardous entity’ without the ontological information of them actually being ‘openings’ (i.e. structures with a hole). Given the heterogeneity of the elements classified in such ad hoc top nodes, little to no formal reasoning can be applied safely over this content since no ontological property can be generalized and implied for it.

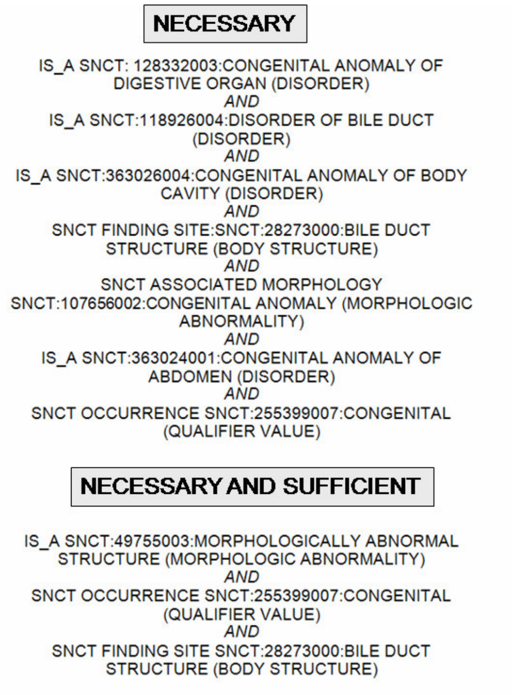
2.3 Methods and principles behind full definitions

Ontologies written in description logics[15], such as the case of both ontologies being compared in this document, rely on an artifact called a ‘formal definition’ in order to apply reasoning and as such explore the ontology’s intelligence inside information systems. Formal definitions in description logics are elements composed of a (sub)set of a concept’s relationships towards other concepts (both hierarchical and horizontal), which are supposed to uniquely define this given concept. For this particular ontological element, the distinction between the two ontologies here in question is given by the principles which govern the assignment of a formal definition to a concept, as well as by the methods used to insert and/or generate these formal definition assignments.

According to SNOMED a formal definition of a concept comprises *“the set of relationships which together define that concept plus an indication of whether this definition fully-defines the concept (i.e. whether the concept is primitive or fully-defined)”*[11], where **all** present *defining* relationships are used in the computation of this definition. For a *defining* relationship SNOMED understands those relationships which *“are always known to be true”*[16] (see section 2.1 in this document). When a concept is marked as ‘fully defined’ in SNOMED it implies that all ‘*necessary’* relationships required to uniquely define the given concept in SNOMED’s terms have been asserted to that concept.

In counterpart to the computational method of asserting formal definitions of SNOMED described above, the formal definitions in LinKBase® are asserted manually by human domain experts in 100% of the cases. This allow for the introduction of an extra notion in the principle which govern this formal definition assignment. Besides the notion of the ‘*necessary’* relationships, the notion of the ‘*sufficient’* relationships is also taken into account. For LinKBase® the smallest set possible of relationships which is ‘*sufficient’* to define a concept comprises the concept’s full definition, instead of the complete set of defining relationships such as it is with SNOMED. Figure 4 shows an example of the distinction between a formal definition of a concept as assigned by SNOMED (top panel) compared to what the definition would be if the notions of necessary and sufficient would be applied (lower panel). By applying the notion of ‘*sufficient’*, the set of relationships which comprise a formal definition becomes smaller, resulting in a reduction of computing time when reasoning over the ontology, since fewer conditions must be cross-checked. In addition, data retrieval is enhanced since the set of conditions to fulfill in order to result into correct subsumption, is smaller.

Another distinction in the nature of formal definitions between these two ontologies is given by the degree of specificity of their relationship types. LinKBase® is very granular with respect to the creation of relationship types, allowing the creation of a new relationship type for each unique



NECESSARY

IS_A SNCT: 128332003:CONGENITAL ANOMALY OF DIGESTIVE ORGAN (DISORDER)
AND
IS_A SNCT:118926004:DISORDER OF BILE DUCT (DISORDER)
AND
IS_A SNCT:363026004:CONGENITAL ANOMALY OF BODY CAVITY (DISORDER)
AND
SNCT FINDING SITE:SNCT:28273000:BILE DUCT STRUCTURE (BODY STRUCTURE)
AND
SNCT ASSOCIATED MORPHOLOGY SNCT:107656002:CONGENITAL ANOMALY (MORPHOLOGIC ABNORMALITY)
AND
IS_A SNCT:363024001:CONGENITAL ANOMALY OF ABDOMEN (DISORDER)
AND
SNCT OCCURRENCE SNCT:255399007:CONGENITAL (QUALIFIER VALUE)

NECESSARY AND SUFFICIENT

IS_A SNCT:49755003:MORPHOLOGICALLY ABNORMAL STRUCTURE (MORPHOLOGIC ABNORMALITY)
AND
SNCT OCCURRENCE SNCT:255399007:CONGENITAL (QUALIFIER VALUE)
AND
SNCT FINDING SITE SNCT:28273000:BILE DUCT STRUCTURE (BODY STRUCTURE)

***Figure 4 - Necessary versus Necessary AND Sufficient conditions for assigning formal definitions to concepts** See text for details*

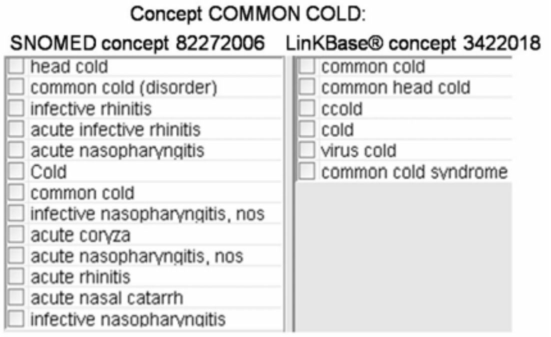
notion of how two concepts can be related in the real world[17]. SNOMED in counterpart is more restrictive, allowing only for a specific set of attributes (i.e. relationship types in SNOMED's terminology) to be used for particular hierarchies[16] (see section 2.1.2.) In summary, an ontology with a vast set of formal definitions has an exceptional capacity to reason both inside itself and over any external data that is connected, mapped or pointed to the ontology.

3. LEXICAL DIFFERENCES

The connection between an ontology and its lexicon is given by the assignment of terms (natural language descriptions) in the lexicon to the concepts and relationships within the ontology, where the ontological content become the symbolized elements while their lexical counterpart are the symbols. Although the content, in terms of the nature or types of descriptions, of both LinKBase's® and SNOMED's lexicons is quite similar, the distinction between these two resources, regarding this particular aspect, is given by the assignments of language descriptions to the ontological content. In addition, the lexicon of SNOMED and LinKBase® differ in their grammatical content.

3.1 Principles behind and nature of synonym assignments

LinKBase® follows a defined set of principles[17] for term assignment which intends to assert that every term assigned to an ontological element is strictly a natural language representation (synonym) for this specific element. SNOMED in counterpart allows for the association of not only terms that represent strictly the given ontological element in question, but also of other terms which are somehow related to this given element either from a taxonomic standpoint (i.e. more generic terms assigned to more specific concepts) or from a clinical perspective (e.g. terms representative of a symptom assigned to concepts representative of a disorder which can manifest with this given symptom). An example of this distinction showed in figure 5, which demonstrates the collection of terms for the concept which represents a "common cold" in both ontologies. While in LinKBase® all terms assigned to this concept are strict synonyms in natural language for this notion, in SNOMED other related terms like "infective rhinitis" or "acute coryza" are also assigned. The former term "infective rhinitis" is actually related to the concept which represents a "common cold" from a taxonomic perspective (i.e. common cold is one of the types of infective rhinitis), while the latter is related from a clinical perspective (i.e. acute coryza is one of the symptoms which may be present in the presentation of a common cold).



Concept COMMON COLD:

SNOMED concept 82272006	LinKBase® concept 3422018
□ head cold	□ common cold
□ common cold (disorder)	□ common head cold
□ infective rhinitis	□ ccold
□ acute infective rhinitis	□ cold
□ acute nasopharyngitis	□ virus cold
□ Cold	□ common cold syndrome
□ common cold	
□ infective nasopharyngitis, nos	
□ acute coryza	
□ acute nasopharyngitis, nos	
□ acute rhinitis	
□ acute nasal catarrh	
□ infective nasopharyngitis	

***Figure 5 - Term assignments in LinKBase® versus SNOMED** This figure shows the collection of terms assigned to the concept 'Common Cold' both in LinKBase® (right) and in SNOMED (left) and exemplifies the difference in principles behind these assignments.*

But how does this distinction affect the usability of each of these ontologies? There are mainly three use cases where we can see a direct influence of one choice of term assignment vs. the other described above. From a search and retrieval perspective one can see a benefit from the broader methods of term assignment of SNOMED, since the user has a higher probability of finding a related concept for the information he/she intends to encode given the larger spectrum of related terms to search for. On the other hand this larger spectrum of term assignment introduces more ambiguity as well as an 'indirect hierarchical inconsistency' (when the ontology is viewed from the lexical standpoint), which complicates processes for automated data integration, (semi)automated creation of mappings between other external resources to be reconciled via the reference ontology of choice, as well as for semantic interoperability between information systems relying on this ontology for their connectivity.

The usability of the ontology as terminological and intelligence resource inside NLP and NLU applications is also highly influenced by the principles applied behind terminological assignment. While broader (non strict) synonym assignment allows for simple uses of the ontology in NLP, such as indexing, it will not yield satisfactory results in more complex NLP/NLU information extraction applications. This is due to the fact that information extraction applications need to make use of the ontology to understand the very concept to which the text refers to, rather than other related content, and then place this identified concept into the context which surrounds it in the given text. Clinical associations in this case for example are highly disruptive, as it might lead the application to conclusions not necessarily true for the situation described by the given text being processed. For example if a mention of "acute coryza" is found in text when using SNOMED as terminological resource, it would be directly associated with the concept of "common cold". Nevertheless, in the particular text or instance in question, the "acute coryza" could as well be due to an allergy or to another disorder, which makes the association to "common cold" incorrect yielding erroneous results.

3.2 Specialized versus Non-Specialized lexicon

A computerized lexicon connected to a given ontology can contain different degrees of information about the terms or vocabulary it contains. SNOMED's lexicon is what we would call 'non-specialized' lexicon, constituting mainly from a collection of terms but without any extra grammatical information about these terms or the way they connect together. In counterpart LinKBase® is connected to what we would call a 'specialized' lexicon, containing for each of its terms (or lexemes in this case) information such as their part of speech or their inflections.

This distinction is vital when considering the ontology for use within NLP/NLU applications. A non-specialized lexicon will allow for the access of the ontology only from a basic indexing or keyword tagging capability. More sophisticated NLP/NLU applications frequently make use of a syntactic parser, which makes the availability of grammatical information about the terms or in other words of a 'specialized' lexicon mandatory.

4. MAPPINGS AND MAPPING METHODOLOGY

Terminologies and classifications are used for different purposes and have different structures and content. To allow exchange of information between different data sources, they need to be connected. For this reason, LinKBase® and SNOMED are linked to 3rd party terminologies such as the International Classification of Diseases (ICD)[18] or the Logical Observation Identifiers Names and Codes (LOINC®)[19]. Although both are linked to the original style and structure of these terminologies, they have a different strategy for building a bridge between them. Their solutions to deal with differences in granularity, an important but complex step[20], especially differ.

The mapping relationship between a LinKBase® concept and a concept or term in an external system is always of complete 'identity'. If needed, additional concepts are created to solve differences in granularity. In contrast, SNOMED solves differences in granularity with the creation of 'narrow to broad' or 'broad to narrow' relationships or no relationships at all, e.g. the concept SNCT2 : 276792008 : PULMONARY HYPERTENSION WITH EXTREME OBESITY (DISORDER), has a 'narrow to broad' relationship to the concepts: ICD-9-CM : 416.8 : OTHER CHRONIC PULMONARY HEART DISEASES and ICD-9-CM : 278.00/ : OBESITY, UNSPECIFIED. Needless to say, that the SNOMED approach results in an incorrect representation of the other terminologies. For example, the LOINC® codes for the two most common tests to diagnose pertussis in humans are LOINC® 548-8 and 549-6[21]. Although these LOINC® codes represent an assay involving a culture to test for the presence of Bordetella pertussis, both are linked to the SNOMED codes for the *organism* Bordetella pertussis and the *condition* pertussis since no such 'Bordetella test culture' exists in SNOMED. In LinKBase® however, this difference in granularity is solved by the creation of concepts that represent identical tests and/or cultures as in LOINC® and to link these to the condition and organisms involved. This method not only allows the correct representation of LOINC® and other terminologies in LinKBase®, but also allows for the reusability of existing mappings, the ability to cross map several data sources simultaneously and the ability to transpose divergent levels of granularity between external information sources[20]. The LinKBase® system allows physicians to enter a diagnostic term, to find its relationships based on its exact meaning and to select the terminology system they wish to use during a patient encounter. There is no need to search multiple terminologies, LinKBase® suffices, since the 3rd party terminologies are fully mapped, based on a consistent meaning without ambiguities.

CONCLUSION

In this paper, we have outlined some distinct features between SNOMED and LinkBase®. Currently, based on their differences in architectural and lexical approach, the principles behind these, together with their different mapping methodology, LinKBase® seems to be more suitable for use in NLP/NLU engineering. However, the comparison also provides a possibility to accommodate SNOMED to a level that it can be integrated in NLP technology and be used for the analysis of free text, as is the case for LinKBase®.

References

1. Côté, R.A., Robboy, S. (1980). Progress in medical information management: Systematized Nomenclature of Medicine (SNOMED), JAMA, 243; 756-762.
2. Spackman, K.A., Campbell, K.E. & Cote, R.A. (1997), SNOMED RT: a reference terminology for healthcare, proc AMIA Annu Fall Symp, 640-644.
3. SNOMED; http://www.snomed.org/
4. M. van Gurp, M. Decoene, M. Holvoet, M. Casella dos Santos. Proceedings of KRMED2006, LinKBase, a Philosophically-inspired Ontology for NLP/NLU Applications http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-222/krmed2006-p08.pdf
5. Smith, Basic formal Ontology (BFO) http://ontology.buffalo.edu/bfo/
6. B. Smith, A. C. Varzi,, Fiat and Bona Fide Boundaries, Proc COSIT-97 Springer-Verlag, 1997: 103-119.
7. Smith, Data and Knowledge Engineering 20, http://ontology.buffalo.edu/smith /articles/mereotopology.htm, 1996.
8. 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.
9. 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.
10. Bateman, Ontology construction and natural language. Proceedings of International Workshop on Formal Ontology, Padua (Italy), 1993: 83-93.
11. SNOMED CT Abstract Logical Model and Representational Forms – November 2006 Revision to External Draft Guide http://www.ihtsdo.org/fileadmin/user_upload/Docs_01/Technical_Docs/abstract_models_and_representational_forms.pdf
12. Degen, W., Herre, H., What is an Upper Level Ontology?, Workshop on Ontologies 2001, Vienna
13. B. Smith. On Substance, Accidents and Universals: In Defense of Constituent Ontology. Philosophical Papers 26, 105-127, 1997.
14. P. Grenon, B. Smith. SNAP and SPAN: Towards Dynamic Spatial Ontology. Spatial Cognition and Computation 4(1), 69-103, 2004.
15. F. Baader, D. Calvanese, D. McGuinness, D. Nardi, P. Patel-Schneider, The description logic handbook: theory, implementation, and applications. Cambridge University Press, New York, NY, 2003
16. SNOMED CT User Guide – January 2007 release http://www.ihtsdo.org/fileadmin/user_upload/Docs_01/Technical_Docs/snomed_ct_user_guide.pdf
17. Flett, M. Casella dos Santos, W. Ceusters Some Ontology Engineering Processes and their Supporting Technologies, in: Gomez-Perez A, Benjamins VR (eds.) Ontologies and the Semantic Web, EKAW2002, Springer 2002, 154-165.
18. International Classification of Diseases, Ninth Revision (ICD-9) http://www.cdc.gov/nchs/about/major/dvs/icd9des.htm
19. Logical Observation Identifiers Names and Codes (LOINC®) http://www.regenstrief.org/loinc/
20. 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, 2005.http://www.landcglobal.com/images/TSB_Methodology.pdf
21. R. Wurtz, ELR, LOINC, SNOMED, and Limitations in public Health, WHP 0042-A, 2005

SNOMED CT: Browsing the Browsers

J Rogers[1], MD, O Bodenreider[2], MD,
[1]Technology Office, NHS Connecting for Health, Leeds UK
[2]National Library of Medicine, NIH, Bethesda USA
jeremy.rogers@nhs.net, olivier@nlm.nih.gov

SNOMED CT is a complex ontology; sophisticated browsers are required to make it understandable and useful. We identified 23 SNOMED CT browsers that have been developed, and inspected 17. We enumerate and provide test criteria for a 'master list' of 143 browsing features supported by at least one inspected browser; future work will determine which of these features are implemented by individual browsers. Only 5 features were common to all 17 browsers; 89 were found in less than one third of browsers. We recommend that a core set of browsing features be defined and harmonized across browsers, particularly for text-to-concept search operations.

INTRODUCTION

SNOMED CT is a biomedical ontology and an associated terminology[1.] Formerly owned by the College of American Pathologists, it has been managed since April 2007 by the International Health Terminology Standards Development Organisation (IHTSDO), a not-for-profit international standards body. As distributed, it is a large, complex and evolving knowledge artifact. Sophisticated browsers must make that complexity accessible and understandable, and suppress distracting or unwanted detail[2-3]. A number of different SNOMED CT browsers have been constructed since it was first published. Some have been evaluated for a variety of use cases, including coding of clinical data[4-8] and terminology evaluation and management[9].

In this paper, we report interim results of a systematic inspection of some of these browsers. We enumerate a superset of browsing features, outline the variability with which these features are implemented in individual browsers, and consider the possible consequences of non-standardized browsing of a standardized terminology.

MATERIALS

SNOMED CT

The core of a SNOMED CT release comprises three tables (sct_concepts, sct_descriptions and sct_relationships) collectively defining a compositional description logic ontology of the medical domain, and a lexicon of associated preferred or synonymous descriptions. The most recent international release (January 2008) contains 311,313 active concepts, 1,357,719 relationships between those concepts and 794,061 active descriptions.

Working deployments of SNOMED CT require additional or ancillary information linked to that core, usually provided by either the IHTSDO or a National Release Centre. Examples of such data include crossmaps to other clinical classifications (e.g. ICD-10), definitions of subsets of concepts and/or their descriptions for navigational or localization purposes, and a history of changes between successive releases.

The January 2008 IHTSDO release therefore comprised 21 discrete table components in addition to the 3 defining the core ontology. The April 2008 UK National Release, which builds on the January 2008 IHTSDO release, comprised 122 separate tables.

In addition to this centrally provided additional content, it is also possible to link external data to the core or ancillary data sources. For example, crossmap target codes can be linked to their corresponding native rubrics or hierarchies.

SNOMED CT Browsers

The authors and their colleagues identified 23 different implementations of software[10-28] offering SNOMED CT browsing capability – either embedded in larger application environments or available as standalone browsers. 16 of these[10-23] were inspected as working software: CaTTS, CliniClue, CLIVE, EdBrowse, FDB Sphinx, HealthTerm, LexPlorer, Mycroft, NCI Terminology Browser, OntoBrowser, OpenKnoME, Protégé-OWL, SNOB, SnoFlake, the UMLS Rich Release Format (RRF) Browser and the Virginia Tech Browser. One additional feature was identified on a screen capture of the AxSys browser.

AxSys, CLIVE, FDB Sphinx, HealthTerm and LexPlorer require user privileges to access; OntoBrowser and EdBrowse are unsupported in-house prototypes. The remaining ten browsers are publicly available at zero cost. Both CliniClue and OpenKnoME require proprietary additional tooling to load SNOMED CT distribution files, although prebuilt CliniClue data is widely available. OpenKnoMe and OntoBrowser also require a proprietary terminology server.

The remaining 6 browsers not inspected[24-28] were: proprietary software from Informatics inc, Ocean Informatics and Visual Read; a demonstrator browser/encoder developed within the NHS Common User Interface Project; Kermanog's CLAW product[17] based on SNOMED in ClaML (EN 14463) format; and Linköping University's browser. These were excluded for reasons of time or lack of access.

METHODS

Each browser was inspected by one author against an emerging catalog of all features exhibited so far by at least one previously inspected browser. Whenever the choice was given to us, browsers were inspected using content based on the July 31, 2007 international release of SNOMED CT. A subset of SNOMED CT content converted into OWL DL was used for Protégé-OWL inspection.

The goal of each successive inspection was primarily to identify novel features implemented in the inspected browsers, for inclusion in a cumulative master catalog. The feature catalogue was iteratively organized by an emerging set of themes, and this resulted in a progressive systematization of the inspection process itself, with each theme considered in detail by turn. This iterative systematisation aided the process of new feature identification.

Where possible, operational definitions of new features were specified (reproduced in Tables 1-3). Subsequent inspections progressed by browsing or searching the Test Case column entry, and comparing the displayed result with the Expected Result column.

Although previously inspected browsers were subsequently re-inspected for newly discovered features, work is underway to confirm the validity and reproducibility of inspecting individual browsers against the feature catalog. Individual browser scores are therefore not presented here.

RESULTS

143 different browsing features were identified across 17 inspected browsers. 6 further features occurred to the authors during the inspection process as being potentially useful, but were not found in any inspected browser. The combined set of 149 features are presented in the accompanying tables, organised under the 8 major themes outlined below.

Our preliminary summary results, based on partially validated individual browser inspections, suggest most browser featuresets are an arbitrarily selected and small subset of all 149 features available. On average, individual browsers implement only 40 features (Range 21-107, StDev=13), but only 22 of the 149 features were found in more than two thirds of all browsers inspected, of which only 5 were implemented in **all** inspected browsers (Search by ConceptID or by Exact string, display of a ConceptID, its linkage to a Description, and the text of that Description). 89 features were found in less than a third of all browsers, but 70 of these are found in at least two browsers. Overall, these results suggests that most possible browsing features have been implemented independently by several SNOMED browser developers, but they have yet to become ‘standard’ across most browsers.

Core Data

A minimal requirement for a SNOMED CT browser is to give access to the data in the three core tables (concepts, relationships, descriptions). Table 1 lists the 22 fields from each of the three core tables that might be displayed by a browser.

Most browsers implement a concept-centric view of this core content, comprising one concept, its description(s), classification with respect to other concepts, and definition in terms of other concepts. This represents the minimum set of features required for the coding of clinical data and basic navigation.

Some fields (e.g. ConceptStatus) appear in the source release data as coded numeric values whose interpretation is given only in SNOMED release documentation; most browser implementations display only the human readable interpretation of these codes and not also (or only) the numeric values as actually distributed.

Despite their ‘core’ nature, however, only three of the 22 related features were displayed by all browsers inspected: the Concept ID, a link to (at least one) description for a concept, and display of the text of linked descriptions. Description status and Initial Capital Status, Relationship ID and Refinability were each visible in only two or three browsers.

Non-Core: Ancillary, 3rd Party and Derived Data

Advanced navigation and terminology maintenance work may require either additional data outside the core tables, or ‘derived’ views of the core data itself such as ‘reverse’ historical relationships (showing which inactive concepts point at the current browser focus concept as their replacement). Table 1 lists the ‘derived’ views found across the inspected browsers.

A complete set of SNOMED core and ancillary linked data is large and complex. Further, it changes with each biannual release. To reflect this configuration and versioning complexity, some browsers report exactly which versions of which release components are loaded, alert users when they are browsing non-current data, and support concurrent browsing of multiple release versions for direct discovery or comparison of changed content.

We found display of non-core data, and data from more than one release, to be the exception rather than the rule. Pointers from inactive concepts to their active replacement, and the set of concepts using the browser focus concept in their definition, are accessible in less than half of all browsers; all other ancillary, 3rd party or derived data browsing functions are present in less than one third of all browsers and usually only in two or three.

Visualisation and Navigation

Following from consideration of what data a browser displays is *how* it displays it. Additionally, the navigability of this data must be considered. Table 2 lists the visualization and navigation features encountered in the inspected browsers.

Most browsers implement some form of graphical tree browser, displaying the browser focus concept in the context of SNOMED's multiaxial subsumption hierarchy. Some off-the-shelf tree controls, however, are unsuitable for displaying trees with very many levels and very many siblings at the same level, such as SNOMED CTs subsumption hierarchy. Those showing the hierarchy always exploded from the root node downward (e.g. the NCI Terminology Browser and Protégé) are particularly unwieldy; those that do not detect very large sibling sets before attempting to display them can lead to very long refresh times.

Other visualization features observed include: sorting and grouping of components within concept definition or synonym sets, diacritic and superscript rendering, and typographic or colour coding of text.

Most browsers employ web browsing paradigms for navigation, with use of hyperlinks to refocus the browser on arbitrary concepts, as well as back/forward navigation. Bookmarked 'favourites', or a 'home' concept, however, were rarely observed.

Usability and Interoperability

The overall experience of working with a browser is influenced by a range of more generic user interface features, listed inTable 2. These include: the ability to transiently or persistently configure a custom view on the wealth of SNOMED related information, e.g., to occupy less of the desktop real estate; copy-and-paste or drag-and-drop of selected information either within the browser environment or into external applications, and the availability of an API allowing browser interface components to be instantiated and controlled by 3rd party software (a functionality distinct from the notion of a terminology services API per se).

Searching

Table 3 lists the range of features observed by which SNOMED CT is searched against a user-entered text string in order to identify candidate SNOMED ConceptIDs as possible entry points for subsequent visualization and navigation. These different search features observed may be further analysed into:

- lexical expansion of the original user search string in order to increase recall
- semantic or metadata filtering of the set of candidate concepts returned by a query, in order to increase precision
- collation and sorting of filtered results, so that the user may find (or be certain of **not** finding) the required concept

In general, SNOMED CT searching functionality in most browsers is impoverished and idiosyncratic. Although 37 different query expansion, filtering and collation features were observed across all browsers, thirteen of the browsers implemented less than 10 of them - and rarely the same set. 27 searching features were implemented in less than a third of all browsers inspected, of which 5 were unique to one browser.

Browsers differ in which features are on by default, which must be explicitly specified, and which can be, or by default are, combined in Boolean combinations. Not all strip trailing spaces; some default to an exact string match whilst others assume wildcarding unless specifically overridden. Where a search expression contains multiple words or tokens, few browsers support complex query logics such as requiring some tokens to be present and others not.

To demonstrate the effect of these differences, all browsers were used in their default configuration to search against the same string: 'ear catheter'. Six browsers found no matches. A further six found only 72683003 Removal of catheter from middle ear, and its two descendants. SNOB returned eleven matches, including 72683003 but also 232199004 Inflation of Eustachian tube using balloon. The latter has no directly associated descriptions containing either 'ear' or 'catheter' but instead is returned because it has at least one ancestor with at least one description matching 'ear', and a separate ancestor with a description matching 'catheter'. The UMLS RRF Browser returned sixty-six matches.

Postcoordination and Miscellaneous

Unlike traditional clinical terminologies, SNOMED CT can be 'postcoordinated' - dynamically extended by anybody, subject to certain ontological rules. Most trivially, this manifests as the option to qualify anatomical sites by a *Laterality* attribute and *Sidedness* value. Exposing SNOMED CT only as a static corpus significantly diminishes its expressivity. Further, a large part of the content – e.g. all Qualifier, and Linkage Concepts - is easily misunderstood outside the context of postcoordination.

The rules governing postcoordination are complex but compliance with them is a prerequisite for dynamic classification of the expressions so built. A dedicated postcoordinated expression building and validating interface is therefore highly desirable, but we found only five browsers that implement one. Three of these additionally implement some limited part of the rules and conventions. However, although compliance with the rules has limited value outside the context of dynamic classification, no browser inspected currently provides that function.

SNOMED CT contains many content errors and omissions. Empowering end users to log and report content errors offers a 'social computing' route to expand SNOMED CT's quality assurance capacity. However, only one inspected browser directly integrates content bug logging and reporting.

DISCUSSION

Accessing data vs. browsing. In seeking to review 'browser' technologies, we excluded command line or other direct SQL interfaces on the data tables. Although most browsers hide the raw data tables from the user, at least one explicitly provides a route to it. Whether 'display' of data by this route should pass or fail our core data theme tests is debatable.

Configurability. A minority of the features identified are orthogonal or graded values of one property. For example, whether a given hierarchy browser sorts sibling concepts randomly, alphabetically by description, or numerically by ConceptID are orthogonal values of a 'sibling sort' function. Although in theory it is possible to imagine a browser configurable to any one of the three, individual hierarchy display instances can only implement one at a point in time. In practice, all inspected browsers implement only one of these options throughout.

Operational test criteria. Differences between the browsers, particularly their default treatment of search strings, confounded attempts to specify tests that would work equally across all of them. Many of the tests specified in Tables 1-3 must be interpreted to take account of issues such as whether exact or wildcard string matching is assumed.

Absence of standard search features. The observed differences in text-to-concept search implementations have a striking effect on browsing experience. Further work to characterize this phenomenon is required.

Future work. We are currently validating the testing of specific browsers against the catalog of features. The quantitative results reported here are preliminary but confirm the authors' original motivation for the experiment: currently available SNOMED CT browsers are very different and often suboptimal.

We do not propose that all SNOMED CT browsers must always implement all the features we identify; further research is required to determine which features are required for specific use cases, but the prior existence of a master feature catalog such as we present here is a prerequisite for that research. Many of the features seem likely to be common across use cases, particularly text-to-concept search operations. We recommend that a core set of searching and browsing features be defined and harmonized across tools, so that a standard terminology is not transformed into multiple different objects by virtue of idiosyncratic and limited browsing experiences.

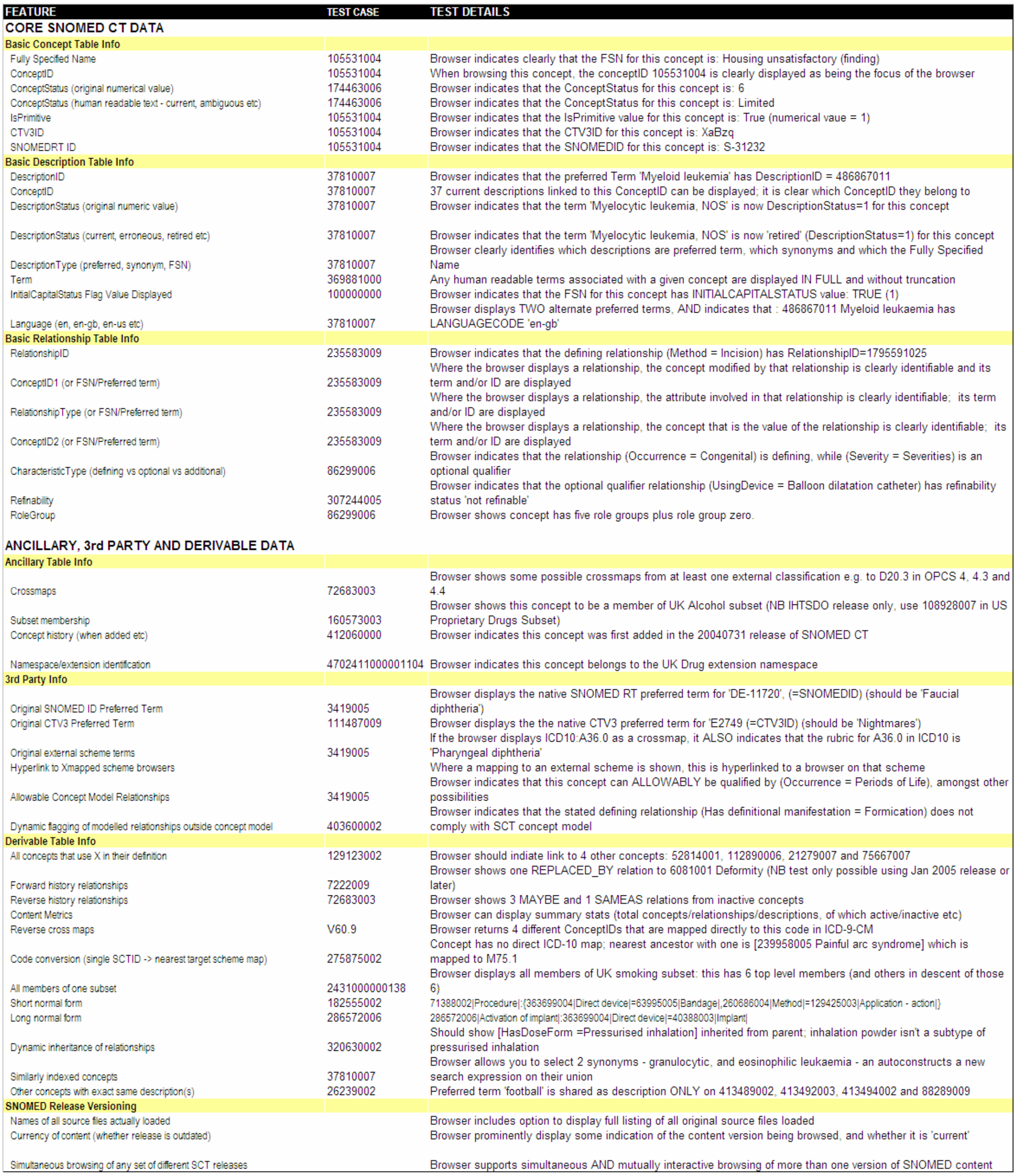
Acknowledgments

This research was supported by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM) and by NHS Connecting for Health. Thanks to Drs Malcolm Duncan and Christopher Wroe for their assistance.

References

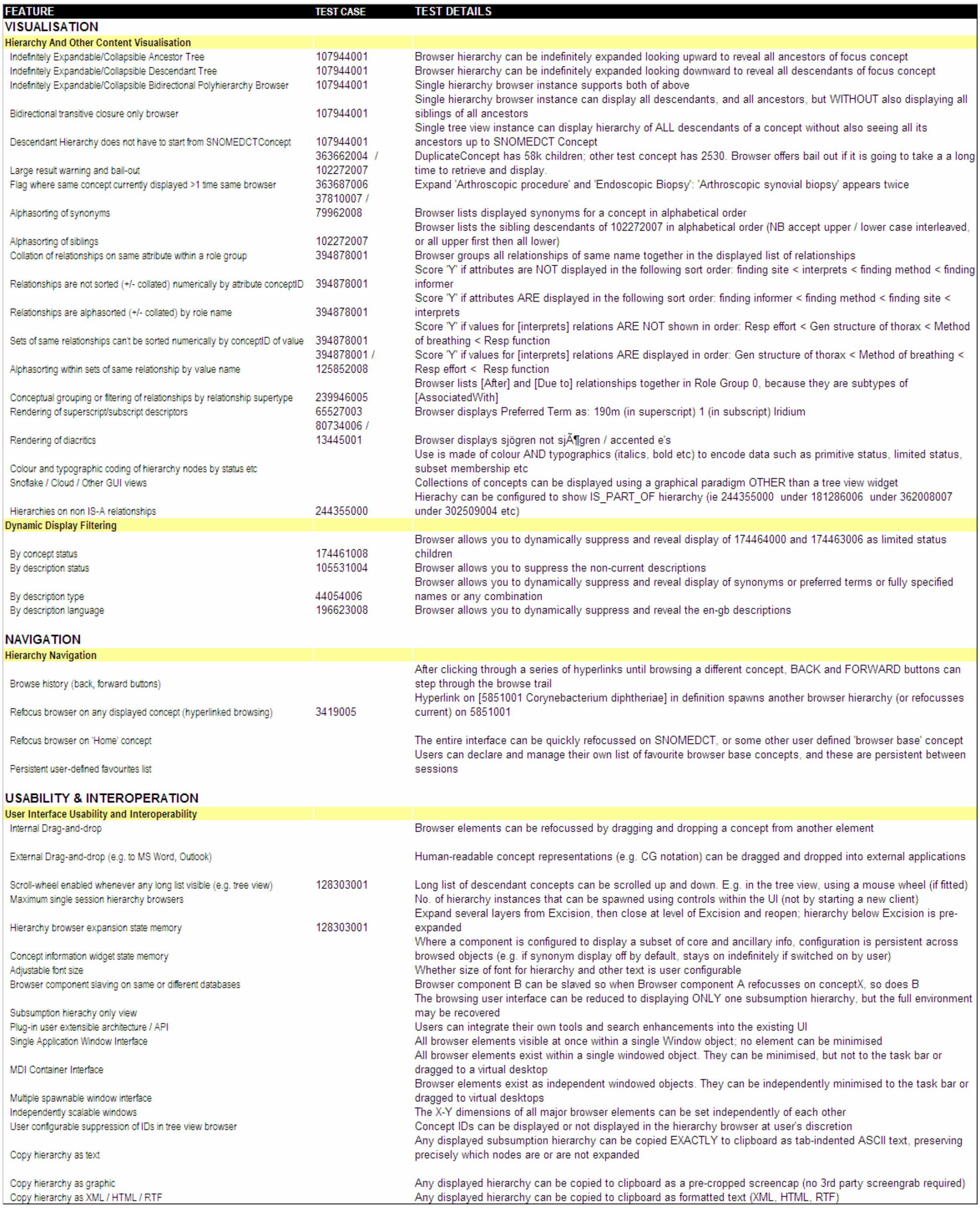
1. SNOMED CT. IHTSDO, Copenhagen 2007 www.ihtsdo.org
2. Tuttle MS, Cole WG, Sheretz DD, Nelson SJ. Navigating to knowledge. Methods Inf Med. 1995 Mar;34(1-2):214-31
3. Patel VL, Kushniruk AW. Understanding, navigating and communicating knowledge: issues and challenges. Methods Inf Med. 1998 Nov;37(4-5):460-70.
4. Windle J, Van-Milligan G, Duffy S et al. Web-based physician order entry: an open source solution with broad physician involvement. AMIA Annu Symp Proc. 2003;:724-7.
5. Elkin PL, Brown SH, Husser CS et al. Evaluation of the content coverage of SNOMED CT: ability of SNOMED clinical terms to represent clinical problem lists. Mayo Clin Proc. 2006 Jun;81(6):741-8.
6. Sundvall E, Nyström M. et al. Interactive visualization and navigation of complex terminology systems, exemplified by SNOMED CT. Std Health Technol Inform. 2006;124:851-6.
7. Chiang MF, Hwang JC, Yu AC et al. Reliability of SNOMED-CT coding by three physicians using two terminology browsers. AMIA Annu Symp Proc. 2006;:131-5.
8. Richesson R, Syed A, Guillette H et al. A web-based SNOMED CT browser: distributed and real-time use of SNOMED CT during the clinical research process. Medinfo. 2007;12(Pt 1):631-5.
9. Cornet R, de Keizer NF, Abu-Hanna A. A framework for characterizing terminological systems. Methods Inf Med. 2006;45(3):253-66.
10. CaTTS (browsed Dec 21st 2007) www.jdet.com/
11. CliniClue (build 2006.2.30) www.cliniclue.com
12. CLIVE (UK NHS in-house terminology authoring tool)
13. HealthTerm (v 4.3.2 browsed Dec 21st 2007)
14. HLi LExPlorer (v 4.4.1P build 48 browsed Dec 21st 2007 – Athens account required) www.snomed.cfh.nhs.uk/lexplorer/
15. Mycroft (v. 2.1.0.2) www.apelon.com/
16. NCI Terminology Browser (browsed Dec 21st 2007) nciterms.nci.nih.gov/NCIBrowser/
17. OpenKnoME 5.4d and ClaW Workbench www.opengalen.org/sources/software.html
18. Protégé (v4.0 build 59) protege.stanford.edu
19. SNOB (v1.64) snob.eggbird.eu
20. SnoFlake (v 2.0 browsed Dec 21st 2007) snomed.dataline.co.uk/
21. UMLS Rich Release Format Browser (2007AC) www.nlm.nih.gov/research/umls/
22. Virginia Tech Browser (browsed Dec 21st 2007) terminology.vetmed.vt.edu/SCT/menu.cfm
23. AxSys Browser (browsed Jan 5th 2008) www.axsys.co.uk/excelicare/eprclinicalcoding.htm

24. First DataBank www.firstdatabank.com/
25. Informatics inc www.informatics.com/
26. Ocean Informatics oceaninformatics.biz
27. Visual Read www.visualread.com/
28. NHS Common User Interface nww.cui.nhs.uk/



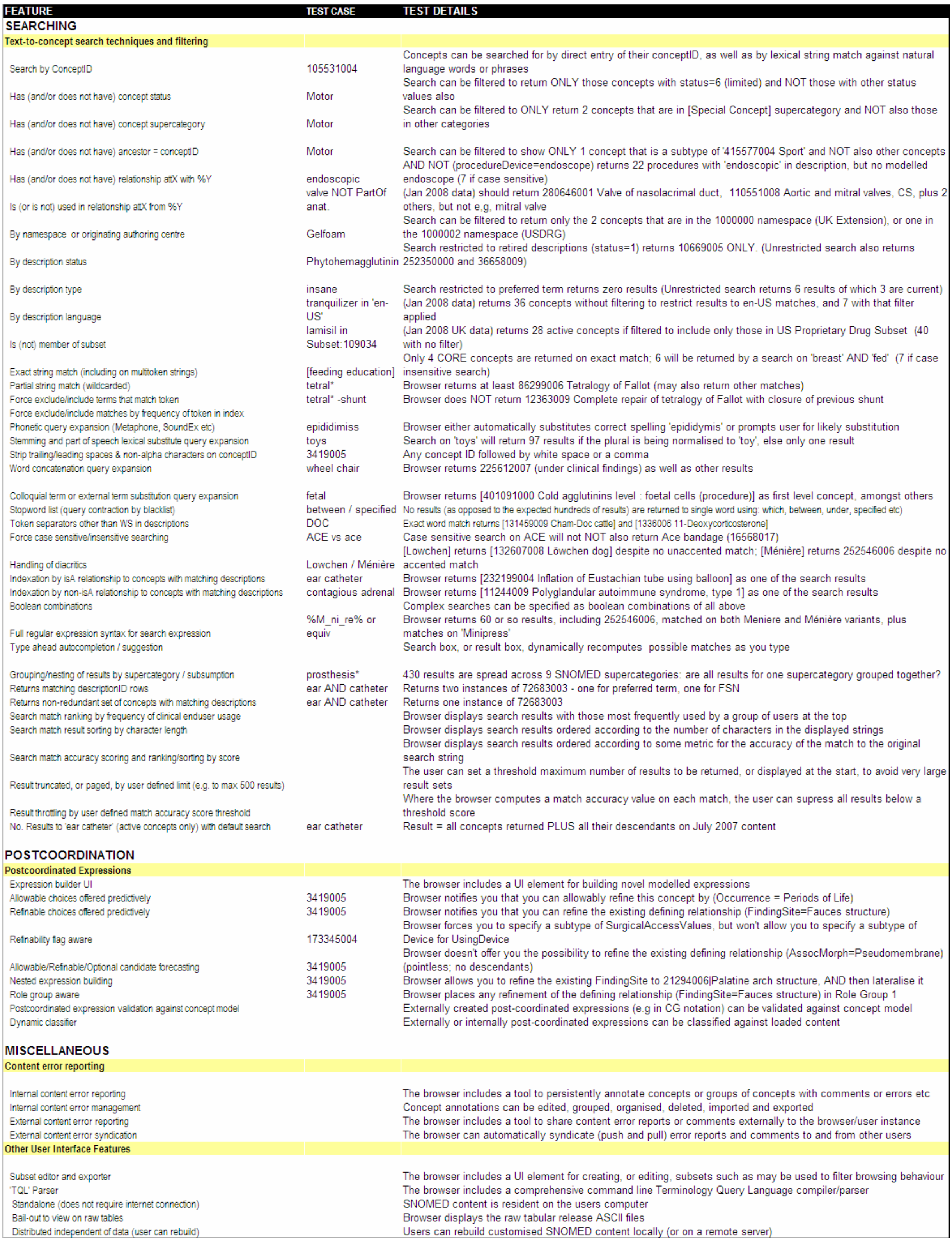
FEATURE	TEST CASE	TEST DETAILS
CORE SNOMED CT DATA		
Basic Concept Table Info		
Fully Specified Name	105531004	Browser indicates clearly that the FSN for this concept is: Housing unsatisfactory (finding)
ConceptID	105531004	When browsing this concept, the conceptID 105531004 is clearly displayed as being the focus of the browser
ConceptStatus (original numerical value)	174463006	Browser indicates that the ConceptStatus for this concept is: 6
ConceptStatus (human readable text - current, ambiguous etc)	174463006	Browser indicates that the ConceptStatus for this concept is: Limited
IsPrimitive	105531004	Browser indicates that the IsPrimitive value for this concept is: True (numerical vaue = 1)
CTV3ID	105531004	Browser indicates that the CTV3ID for this concept is: XaBzq
SNOMEDRT ID	105531004	Browser indicates that the SNOMEDID for this concept is: S-31232
Basic Description Table Info		
DescriptionID	37810007	Browser indicates that the preferred Term 'Myeloid leukemia' has DescriptionID = 486867011
ConceptID	37810007	37 current descriptions linked to this ConceptID can be displayed; it is clear which ConceptID they belong to
DescriptionStatus (original numeric value)	37810007	Browser indicates that the term 'Myelocytic leukemia, NOS' is now DescriptionStatus=1 for this concept
DescriptionStatus (current, erroneous, retired etc)	37810007	Browser indicates that the term 'Myelocytic leukemia, NOS' is now 'retired' (DescriptionStatus=1) for this concept
DescriptionType (preferred, synonym, FSN)	37810007	Browser clearly identifies which descriptions are preferred term, which synonyms and which the Fully Specified Name
Term	369881000	Any human readable terms associated with a given concept are displayed IN FULL and without truncation
InitialCapitalStatus Flag Value Displayed	100000000	Browser indicates that the FSN for this concept has INITIALCAPITALSTATUS value: TRUE (1)
Language (en, en-gb, en-us etc)	37810007	Browser displays TWO alternate preferred terms, AND indicates that : 486867011 Myeloid leukaemia has LANGUAGECODE 'en-gb'
Basic Relationship Table Info		
RelationshipID	235583009	Browser indicates that the defining relationship (Method = Incision) has RelationshipID=1795591025
ConceptID1 (or FSN/Preferred term)	235583009	Where the browser displays a relationship, the concept modified by that relationship is clearly identifiable and its term and/or ID are displayed
RelationshipType (or FSN/Preferred term)	235583009	Where the browser displays a relationship, the attribute involved in that relationship is clearly identifiable; its term and/or ID are displayed
ConceptID2 (or FSN/Preferred term)	235583009	Where the browser displays a relationship, the concept that is the value of the relationship is clearly identifiable; its term and/or ID are displayed
CharacteristicType (defining vs optional vs additional)	86299006	Browser indicates that the relationship (Occurrence = Congenital) is defining, while (Severity = Severities) is an optional qualifier
Refinability	307244005	Browser indicates that the optional qualifier relationship (UsingDevice = Balloon dilatation catheter) has refinability status 'not refinable'
RoleGroup	86299006	Browser shows concept has five role groups plus role group zero.
ANCILLARY, 3rd PARTY AND DERIVABLE DATA		
Ancillary Table Info		
Crossmaps	72683003	Browser shows some possible crossmaps from at least one external classification e.g. to D20.3 in OPCS 4, 4.3 and 4.4
Subset membership	160573003	Browser shows this concept to be a member of UK Alcohol subset (NB IHTSDO release only, use 108928007 in US Proprietary Drugs Subset)
Concept history (when added etc)	412060000	Browser indicates this concept was first added in the 20040731 release of SNOMED CT
Namespace/extension identification	4702411000001104	Browser indicates this concept belongs to the UK Drug extension namespace
3rd Party Info		
Original SNOMED ID Preferred Term	3419005	Browser displays the native SNOMED RT preferred term for 'DE-11720', (=SNOMEDID) (should be 'Faucial diphtheria')
Original CTV3 Preferred Term	111487009	Browser displays the the native CTV3 preferred term for 'E2749 (=CTV3ID) (should be 'Nightmares')
Original external scheme terms	3419005	If the browser displays ICD10:A36.0 as a crossmap, it ALSO indicates that the rubric for A36.0 in ICD10 is 'Pharyngeal diphtheria'
Hyperlink to Xmapped scheme browsers		Where a mapping to an external scheme is shown, this is hyperlinked to a browser on that scheme
Allowable Concept Model Relationships	3419005	Browser indicates that this concept can ALLOWABLY be qualified by (Occurrence = Periods of Life), amongst other possibilities
Dynamic flagging of modelled relationships outside concept model	403600002	Browser indicates that the stated defining relationship (Has definitional manifestation = Formication) does not comply with SCT concept model
Derivable Table Info		
All concepts that use X in their definition	129123002	Browser should indiate link to 4 other concepts: 52814001, 112890006, 21279007 and 75667007
Forward history relationships	7222009	Browser shows one REPLACED_BY relation to 6081001 Deformity (NB test only possible using Jan 2005 release or later)
Reverse history relationships	72683003	Browser shows 3 MAYBE and 1 SAMEAS relations from inactive concepts
Content Metrics		Browser can display summary stats (total concepts/relationships/descriptions, of which active/inactive etc)
Reverse cross maps	V60.9	Browser returns 4 different ConceptIDs that are mapped directly to this code in ICD-9-CM
Code conversion (single SCTID -> nearest target scheme map)	275875002	Concept has no direct ICD-10 map; nearest ancestor with one is [239958005 Painful arc syndrome] which is mapped to M75.1
All members of one subset	2431000000138	Browser displays all members of UK smoking subset: this has 6 top level members (and others in descent of those 6)
Short normal form	182555002	71388002|Procedure|:{363699004|Direct device|=63995005|Bandage|,260686004|Method|=129425003|Application - action|}
Long normal form	286572006	286572006|Activation of implant|:363699004|Direct device|=40388003|Implant|
Dynamic inheritance of relationships	320630002	Should show [HasDoseForm =Pressurised inhalation] inherited from parent; inhalation powder isn't a subtype of pressurised inhalation
Similarly indexed concepts	37810007	Browser allows you to select 2 synonyms - granulocytic, and eosinophilic leukaemia - an autoconstructs a new search expression on their union
Other concepts with exact same description(s)	26239002	Preferred term 'football' is shared as description ONLY on 413489002, 413492003, 413494002 and 88289009
SNOMED Release Versioning		
Names of all source files actually loaded		Browser includes option to display full listing of all original source files loaded
Currency of content (whether release is outdated)		Browser prominently display some indication of the content version being browsed, and whether it is 'current'
Simultaneous browsing of any set of different SCT releases		Browser supports simultaneous AND mutually interactive browsing of more than one version of SNOMED content

Table 1 Core and Additional SNOMED CT table browsing features



FEATURE	TEST CASE	TEST DETAILS
VISUALISATION		
Hierarchy And Other Content Visualisation		
Indefinitely Expandable/Collapsible Ancestor Tree	107944001	Browser hierarchy can be indefinitely expanded looking upward to reveal all ancestors of focus concept
Indefinitely Expandable/Collapsible Descendant Tree	107944001	Browser hierarchy can be indefinitely expanded looking downward to reveal all descendants of focus concept
Indefinitely Expandable/Collapsible Bidirectional Polyhierarchy Browser	107944001	Single hierarchy browser instance supports both of above
Bidirectional transitive closure only browser	107944001	Single hierarchy browser instance can display all descendants, and all ancestors, but WITHOUT also displaying all siblings of all ancestors
Descendant Hierarchy does not have to start from SNOMEDCTConcept	107944001	Single tree view instance can display hierarchy of ALL descendants of a concept without also seeing all its ancestors up to SNOMEDCT Concept
Large result warning and bail-out	363662004 / 102272007	DuplicateConcept has 58k children; other test concept has 2530. Browser offers bail out if it is going to take a a long time to retrieve and display.
Flag where same concept currently displayed >1 time same browser	363687006	Expand 'Arthroscopic procedure' and 'Endoscopic Biopsy': 'Arthroscopic synovial biopsy' appears twice
Alphasorting of synonyms	37810007 / 79962008	Browser lists displayed synonyms for a concept in alphabetical order
Alphasorting of siblings	102272007	Browser lists the sibling descendants of 102272007 in alphabetical order (NB accept upper / lower case interleaved, or all upper first then all lower)
Collation of relationships on same attribute within a role group	394878001	Browser groups all relationships of same name together in the displayed list of relationships
Relationships are not sorted (+/- collated) numerically by attribute conceptID	394878001	Score 'Y' if attributes are NOT displayed in the following sort order: finding site < interprets < finding method < finding informer
Relationships are alphasorted (+/- collated) by role name	394878001	Score 'Y' if attributes ARE displayed in the following sort order: finding informer < finding method < finding site < interprets
Sets of same relationships can't be sorted numerically by conceptID of value	394878001	Score 'Y' if values for [interprets] relations ARE NOT shown in order: Resp effort < Gen structure of thorax < Method of breathing < Resp function
Alphasorting within sets of same relationship by value name	394878001 / 125852008	Score 'Y' if values for [interprets] relations ARE displayed in order: Gen structure of thorax < Method of breathing < Resp effort < Resp function
Conceptual grouping or filtering of relationships by relationship supertype	239946005	Browser lists [After] and [Due to] relationships together in Role Group 0, because they are subtypes of [AssociatedWith]
Rendering of superscript/subscript descriptors	65527003	Browser displays Preferred Term as: 190m (in superscript) 1 (in subscript) Iridium
Rendering of diacritics	80734006 / 13445001	Browser displays sjögren not sjÃ¶gren / accented e's
Colour and typographic coding of hierarchy nodes by status etc		Use is made of colour AND typographics (italics, bold etc) to encode data such as primitive status, limited status, subset membership etc
Snoflake / Cloud / Other GUI views		Collections of concepts can be displayed using a graphical paradigm OTHER than a tree view widget
Hierarchies on non IS-A relationships	244355000	Hierachy can be configured to show IS_PART_OF hierarchy (ie 244355000 under 181286006 under 362008007 under 302509004 etc)
Dynamic Display Filtering		
By concept status	174461008	Browser allows you to dynamically suppress and reveal display of 174464000 and 174463006 as limited status children
By description status	105531004	Browser allows you to suppress the non-current descriptions
By description type	44054006	Browser allows you to dynamically suppress and reveal display of synonyms or preferred terms or fully specified names or any combination
By description language	196623008	Browser allows you to dynamically suppress and reveal the en-gb descriptions
NAVIGATION		
Hierarchy Navigation		
Browse history (back, forward buttons)		After clicking through a series of hyperlinks until browsing a different concept, BACK and FORWARD buttons can step through the browse trail
Refocus browser on any displayed concept (hyperlinked browsing)	3419005	Hyperlink on [5851001 Corynebacterium diphtheriae] in definition spawns another browser hierarchy (or refocusses current) on 5851001
Refocus browser on 'Home' concept		The entire interface can be quickly refocussed on SNOMEDCT, or some other user defined 'browser base' concept
Persistent user-defined favourites list		Users can declare and manage their own list of favourite browser base concepts, and these are persistent between sessions
USABILITY & INTEROPERATION		
User Interface Usability and Interoperability		
Internal Drag-and-drop		Browser elements can be refocussed by dragging and dropping a concept from another element
External Drag-and-drop (e.g. to MS Word, Outlook)		Human-readable concept representations (e.g. CG notation) can be dragged and dropped into external applications
Scroll-wheel enabled whenever any long list visible (e.g. tree view)	128303001	Long list of descendant concepts can be scrolled up and down. E.g. in the tree view, using a mouse wheel (if fitted)
Maximum single session hierarchy browsers		No. of hierarchy instances that can be spawned using controls within the UI (not by starting a new client)
Hierarchy browser expansion state memory	128303001	Expand several layers from Excision, then close at level of Excision and reopen; hierarchy below Excision is pre-expanded
Concept information widget state memory		Where a component is configured to display a subset of core and ancillary info, configuration is persistent across browsed objects (e.g. if synonym display off by default, stays on indefinitely if switched on by user)
Adjustable font size		Whether size of font for hierarchy and other text is user configurable
Browser component slaving on same or different databases		Browser component B can be slaved so when Browser component A refocusses on conceptX, so does B
Subsumption hierachy only view		The browsing user interface can be reduced to displaying ONLY one subsumption hierarchy, but the full environment may be recovered
Plug-in user extensible architecture / API		Users can integrate their own tools and search enhancements into the existing UI
Single Application Window Interface		All browser elements visible at once within a single Window object; no element can be minimised
MDI Container Interface		All browser elements exist within a single windowed object. They can be minimised, but not to the task bar or dragged to a virtual desktop
Multiple spawnable window interface		Browser elements exist as independent windowed objects. They can be independently minimised to the task bar or dragged to virtual desktops
Independently scalable windows		The X-Y dimensions of all major browser elements can be set independently of each other
User configurable suppression of IDs in tree view browser		Concept IDs can be displayed or not displayed in the hierarchy browser at user's discretion
Copy hierarchy as text		Any displayed subsumption hierarchy can be copied EXACTLY to clipboard as tab-indented ASCII text, preserving precisely which nodes are or are not expanded
Copy hierarchy as graphic		Any displayed hierarchy can be copied to clipboard as a pre-cropped screencap (no 3rd party screengrab required)
Copy hierarchy as XML / HTML / RTF		Any displayed hierarchy can be copied to clipboard as formatted text (XML, HTML, RTF)

Table 2: Visualisation, Navigation and Interoperation browsing features



FEATURE	TEST CASE	TEST DETAILS
SEARCHING		
Text-to-concept search techniques and filtering		
Search by ConceptID	105531004	Concepts can be searched for by direct entry of their conceptID, as well as by lexical string match against natural language words or phrases
Has (and/or does not have) concept status	Motor	Search can be filtered to return ONLY those concepts with status=6 (limited) and NOT those with other status values also
Has (and/or does not have) concept supercategory	Motor	Search can be filtered to ONLY return 2 concepts that are in [Special Concept] supercategory and NOT also those in other categories
Has (and/or does not have) ancestor = conceptID	Motor	Search can be filtered to show ONLY 1 concept that is a subtype of '415577004 Sport' and NOT also other concepts
Has (and/or does not have) relationship attX with %Y	endoscopic	AND NOT (procedureDevice=endoscope) returns 22 procedures with 'endoscopic' in description, but no modelled endoscope (7 if case sensitive)
Is (or is not) used in relationship attX from %Y	valve NOT PartOf anat.	(Jan 2008 data) should return 280646001 Valve of nasolacrimal duct, 110551008 Aortic and mitral valves, CS, plus 2 others, but not e,g, mitral valve
By namespace or originating authoring centre	Gelfoam	Search can be filtered to return only the 2 concepts that are in the 1000000 namespace (UK Extension), or one in the 1000002 namespace (USDRG)
By description status	Phytohemagglutinin	Search restricted to retired descriptions (status=1) returns 10669005 ONLY. (Unrestricted search also returns 252350000 and 36658009)
By description type	insane	Search restricted to preferred term returns zero results (Unrestricted search returns 6 results of which 3 are current)
By description language	tranquilizer in 'en-US'	(Jan 2008 data) returns 36 concepts without filtering to restrict results to en-US matches, and 7 with that filter applied
Is (not) member of subset	lamisil in Subset:109034	(Jan 2008 UK data) returns 28 active concepts if filtered to include only those in US Proprietary Drug Subset (40 with no filter)
Exact string match (including on multitoken strings)	[feeding education]	Only 4 CORE concepts are returned on exact match; 6 will be returned by a search on 'breast' AND 'fed' (7 if case insensitive search)
Partial string match (wildcarded)	tetral*	Browser returns at least 86299006 Tetralogy of Fallot (may also return other matches)
Force exclude/include terms that match token	tetral* -shunt	Browser does NOT return 12363009 Complete repair of tetralogy of Fallot with closure of previous shunt
Force exclude/include matches by frequency of token in index		
Phonetic query expansion (Metaphone, SoundEx etc)	epididimiss	Browser either automatically substitutes correct spelling 'epididymis' or prompts user for likely substitution
Stemming and part of speech lexical substitute query expansion	toys	Search on 'toys' will return 97 results if the plural is being normalised to 'toy', else only one result
Strip trailing/leading spaces & non-alpha characters on conceptID	3419005	Any concept ID followed by white space or a comma
Word concatenation query expansion	wheel chair	Browser returns 225612007 (under clinical findings) as well as other results
Colloquial term or external term substitution query expansion	fetal	Browser returns [401091000 Cold agglutinins level : foetal cells (procedure)] as first level concept, amongst others
Stopword list (query contraction by blacklist)	between / specified	No results (as opposed to the expected hundreds of results) are returned to single word using: which, between, under, specified etc)
Token separators other than WS in descriptions	DOC	Exact word match returns [131459009 Cham-Doc cattle] and [1336006 11-Deoxycorticosterone]
Force case sensitive/insensitive searching	ACE vs ace	Case sensitive search on ACE will not NOT also return Ace bandage (16568017)
Handling of diacritics	Lowchen / Ménière	[Lowchen] returns [132607008 Löwchen dog] despite no unaccented match; [Ménière] returns 252546006 despite no accented match
Indexation by isA relationship to concepts with matching descriptions	ear catheter	Browser returns [232199004 Inflation of Eustachian tube using balloon] as one of the search results
Indexation by non-isA relationship to concepts with matching descriptions	contagious adrenal	Browser returns [11244009 Polyglandular autoimmune syndrome, type 1] as one of the search results
Boolean combinations		Complex searches can be specified as boolean combinations of all above
Full regular expression syntax for search expression	%M_ni_re% or equiv	Browser returns 60 or so results, including 252546006, matched on both Meniere and Ménière variants, plus matches on 'Minipress'
Type ahead autocompletion / suggestion		Search box, or result box, dynamically recomputes possible matches as you type
Grouping/nesting of results by supercategory / subsumption	prosthesis*	430 results are spread across 9 SNOMED supercategories: are all results for one supercategory grouped together?
Returns matching descriptionID rows	ear AND catheter	Returns two instances of 72683003 - one for preferred term, one for FSN
Returns non-redundant set of concepts with matching descriptions	ear AND catheter	Returns one instance of 72683003
Search match ranking by frequency of clinical enduser usage		Browser displays search results with those most frequently used by a group of users at the top
Search match result sorting by character length		Browser displays search results ordered according to the number of characters in the displayed strings
Search match accuracy scoring and ranking/sorting by score		Browser displays search results ordered according to some metric for the accuracy of the match to the original search string
Result truncated, or paged, by user defined limit (e.g. to max 500 results)		The user can set a threshold maximum number of results to be returned, or displayed at the start, to avoid very large result sets
Result throttling by user defined match accuracy score threshold		Where the browser computes a match accuracy value on each match, the user can supress all results below a threshold score
No. Results to 'ear catheter' (active concepts only) with default search	ear catheter	Result = all concepts returned PLUS all their descendants on July 2007 content
POSTCOORDINATION		
Postcoordinated Expressions		
Expression builder UI		The browser includes a UI element for building novel modelled expressions
Allowable choices offered predictively	3419005	Browser notifies you that you can allowably refine this concept by (Occurrence = Periods of Life)
Refinable choices offered predictively	3419005	Browser notifies you that you can refine the existing defining relationship (FindingSite=Fauces structure)
Refinability flag aware	173345004	Browser forces you to specify a subtype of SurgicalAccessValues, but won't allow you to specify a subtype of Device for UsingDevice
Allowable/Refinable/Optional candidate forecasting	3419005	Browser doesn't offer you the possibility to refine the existing defining relationship (AssocMorph=Pseudomembrane) (pointless; no descendants)
Nested expression building	3419005	Browser allows you to refine the existing FindingSite to 21294006|Palatine arch structure, AND then lateralise it
Role group aware	3419005	Browser places any refinement of the defining relationship (FindingSite=Fauces structure) in Role Group 1
Postcoordinated expression validation against concept model		Externally created post-coordinated expressions (e.g in CG notation) can be validated against concept model
Dynamic classifier		Externally or internally post-coordinated expressions can be classified against loaded content
MISCELLANEOUS		
Content error reporting		
Internal content error reporting		The browser includes a tool to persistently annotate concepts or groups of concepts with comments or errors etc
Internal content error management		Concept annotations can be edited, grouped, organised, deleted, imported and exported
External content error reporting		The browser includes a tool to share content error reports or comments externally to the browser/user instance
External content error syndication		The browser can automatically syndicate (push and pull) error reports and comments to and from other users
Other User Interface Features		
Subset editor and exporter		The browser includes a UI element for creating, or editing, subsets such as may be used to filter browsing behaviour
'TQL' Parser		The browser includes a comprehensive command line Terminology Query Language compiler/parser
Standalone (does not require internet connection)		SNOMED content is resident on the users computer
Bail-out to view on raw tables		Browser displays the raw tabular release ASCII files
Distributed independent of data (user can rebuild)		Users can rebuild customised SNOMED content locally (or on a remote server)

Table 3: Searching, Postcoordination and Miscellaneous browsing features

Comparing SNOMED CT and the NCI Thesaurus through Semantic Web Technologies

Olivier Bodenreider
U.S. National Library of Medicine, NIH, Bethesda, Maryland, USA
olivier@nlm.nih.gov

Objective: *The objective of this study is to compare two large biomedical terminologies, SNOMED CT and the National Cancer Institute (NCI) Thesaurus, through Semantic Web technologies.* ***Methods:*** *The two terminologies are converted into the Resource Description Framework (RDF) and loaded into a common triple store. The Unified Medical Language System (UMLS) is used to identify correspondences between concepts across terminologies. Concepts common to both terminologies are compared based on shared relations to other concepts.* ***Results:*** *A total of 20,369 pairs of equivalent SNOMED CT and NCI Thesaurus concepts were identified through the UMLS. The highest proportion of shared relata is for the superclasses traversed recursively (75% of the concepts share at least one superclass). Slightly more than half of the concepts studied share at least one associative relation (direct relation or inherited from some ancestor).* ***Conclusions:*** *Overall, SNOMED CT and NCI Thesaurus concepts exhibit a relatively small proportion of shared relata. Semantic Web technologies, including RDF and triple stores, are suitable for comparing large biomedical ontologies, at least from a quantitative perspective.*

INTRODUCTION

In the era of translational medicine, i.e., the application of the discoveries of basic research (made at the bench) to clinical medicine (the patient's bedside) and the refinement of research hypotheses based on clinical findings, basic researchers and healthcare practitioners need to exchange information back and forth. In order to be processed efficiently, both research data and clinical data must be annotated to some reference terminology or ontology. Although some research ontologies and clinical ontologies have a significant degree of overlap, there has typically been little coordination between the groups developing them. As a consequence, the definitions – textual or formal – provided in research ontologies and clinical ontologies for the same biomedical entity may vary significantly, which constitutes a hindrance to the effective integration of data from basic research and clinical practice.

The evaluation of biomedical terminologies for completeness and accuracy remains largely an open research question. In this paper, we propose to compare two large biomedical ontologies developed for different purposes: the NCI Thesaurus (NCIt), used for the annotation of cancer research data, and SNOMED CT, the largest clinical terminology used in electronic patient records. We take advantage of the fact that both ontologies were developed using Description Logic-based systems. Although most classes are not defined with a set of necessary and sufficient conditions, the set of relations in which a given concept is involved still provides a formal definition for this concept, which can be used to compare it to other concepts. We also take advantage of the fact that both ontologies are represented in the Unified Medical Language System (UMLS), which asserts the equivalence between concepts across biomedical ontologies. Finally, we exploit Semantic Web technologies, such as the Resource Description Framework (RDF) to carry out the comparison between these two ontologies.

The objective of this study is to compare the formal definitions of SNOMED CT and NCIt concepts, using Semantic Web technologies. The assumption underlying this study is that two concepts, one from SNOMED CT and one from NCIt, when identified as equivalent in the UMLS, should have similar formal definitions. In other words, our hypothesis is that equivalent concepts from SNOMED CT and NCIt should have related concepts that are also equivalent. To our knowledge, this is the first study to compare biomedical ontologies on a large scale using RDF.

BACKGROUND

The general framework of this study is that of quality assurance in biomedical terminologies and ontologies, which is known to be is a difficult task [1]. Several approaches to auditing terminologies have been proposed, including semantic methods [2], structural methods [3] and linguistic and formal ontological approaches [4]. Methods based on description logics have also been proposed, but have generally been restricted to subsets of large medical ontologies [5].

Various methods have been applied to SNOMED CT [3, 4] and to the NCIt [6]. In contrast to these approaches, we propose to evaluate SNOMED CT and the NCIt simultaneously and against each other. In other words, we want to cross-validate the definitions or assertions provided in one ontology for a given entity with the definitions or assertions provided in the other ontology for the same entity.

The Semantic Web provides a common framework that enables the integration, sharing and reuse of data from multiple sources. Recent research in Semantic Web technologies has delivered promising results to enable information integration across heterogeneous knowledge sources, particularly in the biomedical domain [7]. Semantic Web technologies are a collection of formalisms, languages and tools created to support the Semantic Web. Among them, the Resource Description Framework (RDF) is a W3C-recommended framework for representing data in a common format that captures the logical structure of the data [8]. The RDF representational model uses a single schema in contrast to multiple heterogeneous schemas or Data Type Definitions (DTD) used to represent data in XML by different sources. In conjunction with a single Uniform Resource Identifier (URI), all data represented in RDF form a single knowledge repository that may be queried as one knowledge resource. An RDF repository consists of a set of assertions or triples. Each triple comprises three entities namely, subject, predicate and object. A collection of triples forms a graph and can be stored in a specialized database called a triple store.

MATERIALS

SNOMED CT

SNOMED CT is a concept system and an associated terminology for healthcare [9].· It is managed by the International Health Terminology Standards Development Organisation (IHTSDO), a not-for-profit international standards body with nine member countries. Although its development is based on the Description Logic system KRSS, SNOMED CT is provided as a set of relational tables corresponding to an "inferred view", i.e., the set of non-redundant defining relations for each concept. The July 2007 international release contains 310,311 active elements (309,175 concepts and 1,136 relationships, of which only 61 are actually used to relate concepts) and 1,218,983 relations (pairs of semantically-related concepts). The source files for SNOMED CT (sct_concepts and sct_relationships) were downloaded from the UMLS Knowledge Source Server (http://umlsks.nlm.nih.gov/).

NCI Thesaurus

The National Cancer Institute Thesaurus (NCIt) is a "terminology based on current science that helps individuals and software applications connect and organize the results of cancer research" [10]. The NCIt is produced by the National Cancer Institute, and is a key element of the cancer common ontologic representation environment (caCORE) [11]. The NCIt uses the description logic flavor of the Web Ontology Language (OWL-DL) for its representation [12]. Version 07.05e of the NCIt contains 58,869 active classes, 123 associative relationships and 124,775 relations (subsumption and equivalence relations, as well as restrictions in the OWL file). The OWL file for the NCIt was downloaded from the caCORE FTP site (ftp://ftp1.nci.nih.gov/pub/cacore/), under EVS.

Unified Medical Language System

The Unified Medical Language System (UMLS) is a terminology integration system developed at the U.S. National Library of Medicine [13]. The UMLS Metathesaurus is a repository of integrated biomedical terms drawn from 143 biomedical vocabularies and ontologies. Terms referring to the same entity in several vocabularies are clustered together and given the same concept unique identifier (CUI). Both SNOMED CT (July 31, 2007) and NCIt (07.05e) are integrated in version 2007AC of the Metathesaurus, which provides a convenient way of identifying equivalences between terms from these two ontologies. The UMLS is available for download from the UMLS Knowledge Source Server (http://umlsks.nlm.nih.gov/). (A free license is required).

METHODS

The method developed for comparing concepts from SNOMED CT and NCIt can be summarized as follows. The formal definition of concepts is extracted from SNOMED CT and NCIt and converted to RDF triples. Equivalence relations between SNOMED CT and NCIt concepts are extracted from the UMLS . All triples are loaded into a triple store. Additional triples are generated from inference rules applied to the original knowledge base. The triple store is then queried to compare the representation of concepts in SNOMED CT and NCIt.

Acquiring RDF triples

For each concept and relationship from SNOMED CT and NCIt, we extract the following information: original identifier, preferred name, source (SNOMED CT or NCIt), type (concept or relationship). RDF triples are created to represent this information, in which the subject is the concept itself. The predicates corresponding to the properties listed above are *hasID*, *hasName*, *hasSource* and *hasType*, respectively. The object of these triples is a literal corresponding to, for example, the concept name for the predicate *hasName*. Triples are also created for representing the relations of each concept to other concepts from the same source. The relationship indicated in the source is used as predicate for these triples, whose objects are concepts. Similarly, triples are created for representing relations among relationships (e.g., *sub-*

PropertyOf). Finally, we create triples to represent the mapping of concepts to the UMLS Metathesaurus. For each concept from SNOMED CT and NCIt, we create one triple with the predicate *hasCUI* and the corresponding UMLS CUI as object literal.

SNOMED CT. The fields 'CONCEPTID' and 'FULLYSPECIFIEDNAME' from the table stc_concept were used to instantiate the properties *hasID* and *hasName*, respectively. All nodes were assigned the value 'concept' for the property *hasType*, except for the elements of the table stc_concept actually corresponding to relationships, namely, *Linkage concept (linkage concept)* and its descendants, to which the value 'relationship' was assigned. All nodes were assigned the value 'SNOMEDCT' for the property *hasSource*.

NCI Thesaurus. The elements 'code' and 'Preferred_Name' from the '<owl:Class>' sections of the OWL file were used to instantiate the properties *hasID* and *hasName*, respectively. All nodes were assigned the value 'concept' for the property *hasType*. Analogously, information extracted from the '<owl:ObjectProperty>' sections of the OWL file was used to create the corresponding triples for properties (i.e., predicates). These nodes were assigned the value 'relationship' for the property *hasType*. All nodes were assigned the value 'NCI' for the property *hasSource*.

UMLS Metathesaurus. The table MRCONSO.RRF from the UMLS distribution was used for acquiring the mapping between terms from SNOMED CT and the UMLS concepts, as well as between terms from the NCIt and the UMLS concepts. We used the source abbreviation (SAB) to identify strings contributed by SNOMED CT (SAB = 'SNOMEDCT') or NCTt (SAB = NCI). We extracted the concept identifier in the source (SCUI) and UMLS concept unique identifier (CUI) and created triples of the form (concept, *hasCUI*, CUI) for each pair (SCUI, CUI).

Creating the triple store

These triples generated from SNOMED CT, NCIt and the UMLS were represented in N-triple format and loaded into the open source triple store *Mulgara*™ (http://mulgara.org/) in a linux environment. *Mulgara* automatically indexes the triples, as well as the subject, predicate and object elements of each triple.

Inference rules

Inference rules are typically added to a triple store in order to infer new RDF statements (i.e., triples) from existing RDF statements. *Mulgara* provides a series of rules, which implement RDF Schema (RDFS) entailment, including rules for the transitivity of the relationships *rdfs:subClassOf* and *rdfs:subPropertyOf*. We found the set of rules for RDFS impractical to use on this triple store and ended up not using it. (The lack of generalized transitive closure in the triple store was compensated for by graph traversal functions in the queries.)

In practice, the only rule we created and applied to the store makes a concept from SNOMED CT equivalent to a concept from NCIt when both concepts are mapped to the same UMLS concept (i.e., share the same UMLS CUI). This relation was implemented by creating an *owl:sameAs* relationship between the two concepts, bidirectionally.

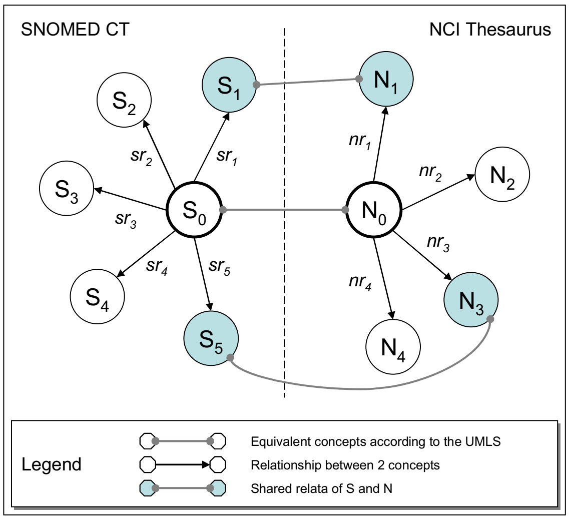


Figure 1. Graph formed by the related concept of one pair of equivalent concepts (S_0, N_0)

Querying the triple store

A set of queries was developed to explore the relata of those concepts that are equivalent between SNOMED CT and NCIt according to the UMLS. More specifically, these queries explore the set of relata of the SNOMED CT concept and that of the NCIt concept, and select from the two sets the relata identified as equivalent in the UMLS. For example, as illustrated in Figure 1, the concepts S_0 from SNOMED CT and N_0 from NCIt are equivalent according to the UMLS. Among the relata of S_0 (S_1 to S_5) and N_0 (N_1 to N_4), the pairs {S_1, N_1} and {S_5, N_3} denote equivalent concepts and constitute the set of shared relata of {S_0, N_0}.

Each relation between two concepts (e.g., (S_0, sr_4, S_4)) is represented as a triple in the RDF store and the set of all relations forms a graph. Comparing the set of relata of two concepts can thus be expressed as a set of constraints on the graph. For example, {S_1, N_1} are shared relata of {S_0, N_0}, because there is a path between S_0 and N_0, constituted of any link from S_0 to S_1, any link from N_0 to N_1, and a "UMLS equivalence" link between S_1 and N_1.

The set of relata is not necessarily limited to direct relata. Some relations can be traversed recursively in order to explore, for example, the set of common ancestors (as opposed to common direct subclasses).

Depending on the constraints put on the graph, various kinds of relationships can be explored, together or independently.

One of the major query languages for RDF stores is SPARQL. *Mulgara* currently provides no support for SPARQL. Instead, it provides iTQL^{TM} (Interactive Tucana Query Language^{TM}), which is functionally equivalent to SPARQL for most purposes.

```
select $n_sub $n_rel $n_obj $s_sub $s_rel $s_obj
from <rmi://localhost/server1#nci_snomed_full>
where
(
   # ---------- NCIT side ----------
   walk(<ncit:C2986> <rdfs:subClassOf> $n_obj
             and $n_sub_tmp <rdfs:subClassOf> $n_obj)
   and $n_rel <mulgara:is> <rdfs:subClassOf>
   and $n_sub <mulgara:is> <ncit:C2986>
)
and
(
   # ---------- SNCT side ----------
   walk(<snct:46635009> <snct:116680003> $s_obj
        and $s_sub_tmp <snct:116680003> $s_obj)
   and $s_rel <mulgara:is> <snct:116680003>
   and $s_sub <mulgara:is> <snct:46635009>
)
and $n_obj <owl:sameAs> $s_obj
in <rmi://localhost/server1#nci_snomed_full_ent_sameAs>
;
```

Figure 2. iTQLquery used to explore the common superclasses of the concepts C2986 from NCIt and 46635009 from SNOMED CT

```
[ ncit:C2986, rdfs:subClassOf, ncit:C2991, snct:46635009, snct:116680003, snct:64572001 ]
[ ncit:C2986, rdfs:subClassOf, ncit:C3009, snct:46635009, snct:116680003, snct:362969004 ]
[ ncit:C2986, rdfs:subClassOf, ncit:C2985, snct:46635009, snct:116680003, snct:73211009 ]
[ ncit:C2986, rdfs:subClassOf, ncit:C27067, snct:46635009, snct:116680003, snct:17346000 ]
[ ncit:C2986, rdfs:subClassOf, ncit:C53655, snct:46635009, snct:116680003, snct:126877002 ]
[ ncit:C2986, rdfs:subClassOf, ncit:C2990, snct:46635009, snct:116680003, snct:53619000 ]
[ ncit:C2986, rdfs:subClassOf, ncit:C26842, snct:46635009, snct:116680003, snct:3855007 ]
```

Figure 3. Results of the query in Figure 2 (aliases are used in lieu of the full URIs)

Comparing the shared relata of concepts

In order to compare the formal definitions of a concept S_0 from SNOMED CT and N_0 from NCIt, we prepared queries to explore the following sets of shared relata: all shared relata (including through associative relations), shared superclasses, shared wholes (of which the entity is a part of), shared subclasses and shared parts. More precisely, these kinds of relations were first explored directly to extract the set of relata in direct relation to the original concepts, and indirectly, allowing the recursive traversal of *isa* and *part_of* relationships. Finally, in order to account for the inheritance of properties from a superclass to its subclasses, we also explored the concepts in associative relation to any of the superclasses of the original concepts.

In practice, starting from the list of pairs of equivalent concepts, we generated one query per pair for each type of relationship to be explored. The relata in common were recorded for each pair of equivalent concepts for each type of relationship explored. Figure 2 shows a typical query used to explore (recursively) the common superclasses of two concepts. Figure 3 displays the output of this query, showing the 7 ancestors in common.

Data analysis

We analyzed the lists of shared relata resulting from the queries from a quantitative perspective, in order to examine the distribution of the number of common relata for the various kinds of relationships under investigation.

RESULTS

Triple store

A total of 3,194,215 triples were created, 2,770,477 for SNOMED CT and 423,738 for NCIt. It took about 20 minutes to load these N-triples into *Mulgara*, including the creation of indexes.

The rule asserting the equivalence of SNOMED CT and NCIt concepts when they share the same UMLS CUI generated 40,738 additional triples (representing the *owl:sameAs* relations bidirectionally). It took about 5 minutes to apply this rule to the triple store.

Queries were executed in batches, one batch for each set of equivalent concepts for a given kind of relationship. Executing a batch of queries took anywhere between several minutes (for direct relations) to several hours (when relations are allowed to be traversed recursively).

Overlap between SNOMED CT and NCIt concepts

Of the 309,175 SNOMED CT concepts, 19,506 (6.3%) mapped to the same UMLS concept as some NCIt concept. Analogously, 14,054 (23.9%) of the 58,869 NCIT concepts mapped to the same UMLS concept as some SNOMED CT concept. A total of 20,369 pairs of SNOMED CT and NCIt concepts were identified in which the two concepts are deemed equivalent based on their mapping to the UMLS.

Quantitative results

The distribution of the number of relata for several types of relationships investigated is summarized in Table 1. The first column (N) shows the total number of pairs of concepts for which both concepts have at least one related concept for this relation. This number is used as the denominator for computing the percentage of pairs of equivalent concepts having a given number of related concepts in common. The minimum, maximum and median number of shared relata are presented in the last three columns. For example, the row "Dir. Superclass" corresponds to the shared direct parent classes (traversing *isa* in SNOMED CT and *subClassOf* in NCIt). N = 20,360 indicates that almost all concepts have at least one ancestor. 18.4% of the pairs of equivalent concepts studied share a parent class and only 1.3% share two. Over 80% of the pairs do not share any direct parents. The row "Ind. Superclass" corresponds to the shared ancestors (traversing *isa* or *subClassOf* recursively). Only 25% of the pairs of equivalent concepts studied do not have any ancestors in common. The largest number of ancestors in common is 22.

Details about shared relata for other kinds of relationships are provided in the other rows of Table 1, including direct parent and child classes for the taxonomic relation (super/subclass) and for the meronomic relation (whole/part). The identification of indirect relata involves the recursive traversal of taxonomic and meronomic relations and combination of *sucblassOf* and associative relations.

EXTENDED EXAMPLE

In order to illustrate our approach to comparing ontologies, we explore how *Type 1 diabetes mellitus* is represented in SNOMED CT and NCIt. As shown in Figure 4, this concept has many relata both in SNOMED CT and in NCIt, of which a large number are shared, including 7 shared ancestors (e.g., *Disorder of pancreas*) and 4 shared concepts in associative relation (e.g., *Gastrointestinal System*). Dotted lines represent indirect *isa* relations through concepts that are not shown. The equivalence between concepts in SNOMED CT and NCIt assessed through the UMLS is shown with grey links. Of note, two distinct concepts in one ontology can be equivalent to one concept in the other (e.g., *Endocrine Pancreas* and *Islet of Langerhans* in NCIt vs. *Endocrine pancreatic structure* in SNOMED CT).

DISCUSSION

SNOMED CT and NCIt

Overall, the two ontologies under investigation in this study were found to have a relatively small proportion of relata in common, including when the properties (e.g., associative relations) are explored in the ancestors to simulate the inheritance of properties along *isa* hierarchies. The highest proportion of shared relata is for the superclasses traversed recursively (75% of the concepts share at least one superclass). Slightly more than half of the concepts studied share at least one associative relation (direct relation or inherited from some ancestor).

Further research is needed to distinguish among primitive concepts in both ontologies (e.g., *Aneurismal bone cyst*), concepts for which a relatively rich description is provided, but only in one ontology (e.g., the description provided for many cancers in NCIt is typically richer than in SNOMED CT), and concepts defined in both ontologies, but with minimal overlap in their relata. We did not complete the comparison of shared descendants, but, even in the absence of a rich description, a large proportion of shared descendants can be a good indicator of consistency between ontologies (e.g., *Sulfonamide agents* share 18 descendants).

Semantic Web technologies

We found RDF to be suitable for comparing terminological ontologies, especially when the two ontologies are large and are not both available in OWL. While OWL classifiers are useful for consistency checking purposes, they tend to be limited in the number of classes they can handle. Moreover, the queries presented in this study arguably allow more flexibility than OWL DL classifiers.

The triple store approach also offers clear advantages over relational databases, as SQL provides no support for performing transitive closures (i.e., for performing joint operations recursively). While *ad hoc* programs (or stored procedures) embedding SQL queries can be written against the database, we showed that simple queries against the RDF store were sufficient to carry out this study. Because it supports the seamless traversal of complex graphs (recursive traversal of one relationship and traversal of selected combinations of relationships), RDF is an effective approach to comparing terminologies.

The comparison of large ontologies remains nonetheless difficult. The inference engine of *Mulgara* could not apply the set of rules defined for RDFS, including the transitivity of *subClassOf* to large, heavily hierarchical structures. However, the graph traversal functions supported by the query language partially compensated for the absence of precomputed transitive closures.

Limitations and future work

This approach essentially provides a quantitative comparison between two ontologies and is insufficient for fine-grained comparisons. Although we did not study whether pairs of related concepts in both ontologies were linked by similar relations, the information could be easily extracted from the triple store. We also would like to test the structural consistency of the combined ontologies (e.g., by testing the presence of cycles in *isa* relations in the RDF store containing both SNOMED CT and NCIt). The advantage of using the UMLS perspective on concept equivalence outweighs the potential bias it introduces with its “concept view”.

Acknowledgements

This research was supported by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM). Our thanks go to Ramez Ghazzaoui who helped create the triple store and Lee Peters who processed SNOMED CT.

References

1. Rogers JE. Quality assurance of medical ontologies. Methods Inf Med 2006;45(3):267-74
2. Cimino JJ. Auditing the Unified Medical Language System with semantic methods. J Am Med Inform Assoc 1998;5(1):41-51
3. Wang Y, Halper M, Min H, Perl Y, Chen Y, Spackman KA. Structural methodologies for auditing SNOMED. J Biomed Inform 2007;40(5):561-81
4. Ceusters W, Smith B, Kumar A, Dhaen C. Ontology-based error detection in SNOMED-CT. Medinfo 2004;11(Pt 1):482-6
5. Cornet R, Abu-Hanna A. Auditing description-logic-based medical terminological systems by detecting equivalent concept definitions. Int J Med Inform 2007
6. Ceusters W, Smith B, Goldberg L. A terminological and ontological analysis of the NCI Thesaurus. Methods Inf Med 2005;44(4):498-507
7. Ruttenberg A, Clark T, Bug W, Samwald M, Bodenreider O, Chen H, et al. Advancing translational research with the Semantic Web. BMC Bioinformatics 2007;8 Suppl 3:S2
8. RDF: http://www.w3.org/RDF/
9. SNOMED CT: http://www.ihtsdo.org/
10. de Coronado S, Haber MW, Sioutos N, Tuttle MS, Wright LW. NCI Thesaurus: using science-based terminology to integrate cancer research results. Medinfo 2004;11(Pt 1):33-7
11. Phillips J, Chilukuri R, Fragoso G, Warzel D, Covitz PA. The caCORE Software Development Kit: streamlining construction of interoperable biomedical information services. BMC Med Inform Decis Mak 2006;6:2
12. Golbeck J, Fragoso G, Hartel F, Hendler J, Oberthaler J, Parsia B. The National Cancer Institute's Thesaurus and Ontology. Web Semantics: Science, Services and Agents on the World Wide Web 2003;1(1):75-80
13. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004;32(Database issue):D267-70

Table 1. Distribution of the number of related concepts shared by pairs of equivalent concepts (N) for various kinds of relationships (top: direct relations, bottom: indirect relations, including recursive traversal and combination of sucblassOf and associative relations)

	Relationship	N	Number of related concepts							min	max	med ian
			0	1	2	3	4	5	> 5			
Dir.	Any	20,363	66.8%	21.1%	5.9%	2.9%	1.3%	0.7%	1.3%	0	47	0
	Superclass	20,360	80.3%	18.4%	1.3%	0.0%	0.0%	0.0%	0.0%	0	4	0
	Whole	1,004	96.2%	3.8%	0.0%	0.0%	0.0%	0.0%	0.0%	0	1	0
	Subclass	3,699	48.9%	21.9%	15.2%	6.4%	2.8%	1.8%	2.9%	0	19	1
	Part	76	57.9%	34.2%	7.9%	0.0%	0.0%	0.0%	0.0%	0	2	0
Ind.	Superclass	20,360	25.0%	28.5%	18.7%	11.1%	5.5%	3.6%	7.7%	0	22	1
	Whole	1,004	93.3%	6.1%	0.6%	0.0%	0.0%	0.0%	0.0%	0	2	0
	Associative	6,548	46.3%	18.6%	11.3%	10.6%	6.8%	2.4%	4.1%	0	11	1

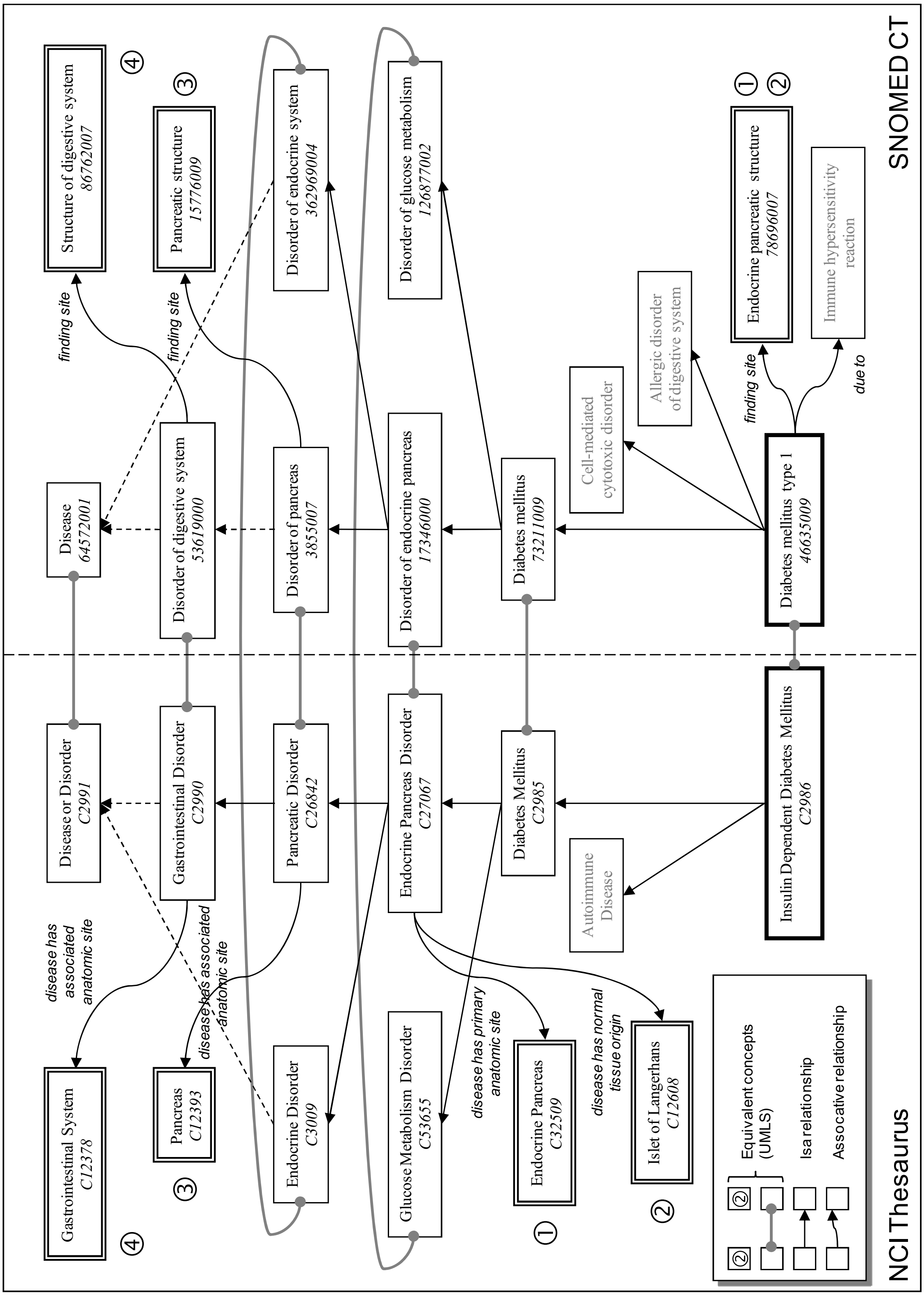

SNOMED CT
Diabetes mellitus type 1 46635009
Diabetes mellitus 73211009
Disorder of endocrine pancreas 17346000
Disorder of pancreas 3855007
Disorder of digestive system 53619000
Disease 64572001
Cell-mediated cytotoxic disorder
Allergic disorder of digestive system
Endocrine pancreatic structure 78696007
Immune hypersensitivity reaction
finding site
due to
Disorder of glucose metabolism 126877002
Disorder of endocrine system 362969004
Pancreatic structure 15776009
Structure of digestive system 86762007
finding site
finding site
①
②
③
④
NCI Thesaurus
Insulin Dependent Diabetes Mellitus C2986
Diabetes Mellitus C2985
Endocrine Pancreas Disorder C27067
Pancreatic Disorder C26842
Gastrointestinal Disorder C2990
Disease or Disorder C2991
Autoimmune Disease
Islet of Langerhans C12608
Endocrine Pancreas C32509
Glucose Metabolism Disorder C53655
Endocrine Disorder C3009
Pancreas C12393
Gastrointestinal System C12378
disease has normal tissue origin
disease has primary anatomic site
disease has associated anatomic site
disease has associated anatomic site
②
①
③
④
②
②
Equivalent concepts (UMLS)
Isa relationship
Assocative relationship

Figure 4. Representation of *Type 1 diabetes mellitus* in SNOMED CT and NCIt, showing shared relata for ancestors and associative relationships

Exploratory Reverse Mapping of ICD-10-CA to SNOMED CT

Dennis Lee, M.Sc., Francis Lau, Ph.D.
School of Health Information Science, University of Victoria, Victoria, B.C., Canada
dlkh@uvic.ca, fylau@uvic.ca

ABSTRACT

This paper describes the findings of an exploratory study on reverse mapping of ICD-10-CA, the Canadian Adaptation, to SNOMED CT. For this study a set of 5,000 most frequent ICD-10-CA codes from the health ministry of a Canadian province was used. The methods included applying six mapping algorithms to each ICD-10-CA description to find the matching SNOMED CT concepts, and comparing the output against the UK SCT-ICD10 cross map for accuracy. Overall, we found successful SNOMED CT matches for ~63% of the ICD-10-CA codes. Issues requiring further attention include ways to increase successful matches and independent validation of mapping output. This study provides a glimpse of the methods that could lead to a SNOMED CT to ICD-10-CA cross map. It should be of interest to those responsible for secondary use of discharge abstracts in epidemiological and statistical reporting.

INTRODUCTION

The Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) is a terminology system used to capture information relating to a patient's condition and care in a consistent manner. Currently, there are ~376000 concepts in SNOMED CT, organized into 19 hierarchies such as clinical finding, observations, body structure and social context. There are another ~1 million commonly used terms to describe these concepts, and ~1.4 million semantic relationships to define the logical connections between concepts [1].

While SNOMED CT is the terminology of choice for capturing details of a clinical encounter, it is considered too fine grained for non-clinical purposes such as the reporting of resource use and billing. Many have advocated the need to link SNOMED CT to established classification systems, such as the International Statistical Classification of Diseases and Related Health Problems Version 10 (ICD-10), that are already used extensively in statistical reporting [2,3]. Currently there is a cross map from SNOMED CT to ICD-10 in the UK, and one to ICD-9-CM (Clinical Modification) in the United States. Neither of these maps have been validated externally, and no map exists for ICD-10-CA, the Canadian Adaptation. There are other cross maps that have been created for specific domains including the SNOMED-to-ICD-O map for oncology, the SNOMED-to-LOINC map for laboratory test results, and those for nursing terminologies. Otherwise there is limited experience in cross mapping from SNOMED CT to existing classification systems to facilitate secondary uses.

In this paper, we describe the initial findings of an exploratory study to create a reverse map from ICD-10-CA to SNOMED CT. It originated as part of a Master of Science project by the lead author. We contend that reverse mapping could be one way to produce the SNOMED CT to ICD-10-CA cross map. This paper describes the mapping algorithms and process used, the key results on matches found, and the lessons and implications from the study.

METHODS

Overview of ICD-10-CA

The ICD-10-CA is an enhanced version of the ICD-10 published by the World Health Organization (WHO). The ICD-10-CA has 23 chapters and is used for classifying morbidity, diseases, injuries and causes of death in Canada. It also covers non-disease situations and conditions that pose a risk to health including occupational and environmental factors, lifestyle and psycho-social circumstances. The ICD-10-CA has an alphanumeric coding format of 3-6 characters. The major difference between ICD-10 and ICD-10-CA is that the latter has two additional chapters: XXII on morphology of neoplasms and XXIII on provisional codes for research and temporary assignment. There are also minor changes in some chapters in the form of addition, subdivision, deletion and revision of selected ICD codes [4].

Source Mapping Terms

For this study, we obtained a set of 5,000 most frequently reported ICD-10-CA codes and their long descriptions for the fiscal year of 2005/06 from the health ministry of a Canadian province. These source mapping terms were from inpatient separations in acute care settings including designated sub-acute care facilities for patients that require more care and time before returning home. The profile of the discharge abstracts for the 5,000 ICD-10-CA codes selected for the study is in Table 1.

Description	Count
Total separations 2005/06 in province	364,977
Total diagnosis codes reported	1,481,285
Average no. of codes reported per separation	4.1
Total discrete diagnosis codes (all)	10,529
Frequency of top 5,000 diagnosis codes	1,460,730
% of total diagnosis in top 5000 codes	98.6%
% of total discrete diagnosis in top 5000 codes	47.5%
Total discrete most responsible diagnosis codes	6,651

Table 1. Profile of the Discharge Abstracts

Mapping Algorithms

After conducting a detailed review of the literature on cross mapping of terminology systems, we adopted five related mapping algorithms and created Web-based versions of these algorithms in to find matching SNOMED concepts for each of the ICD-10-CA descriptions in the data set [5]. Four of the algorithms are lexical techniques for exact-match, match-all-words-only, match-all-words and partial-match. The fifth is semantic matching that involves retrieving the current concepts based on entries in the SNOMED historical relationship table if the initial concepts found are inactive. These mapping algorithms are summarized in Table 2.

Algorithm	Explanation
1. Exact match	Exact string match where all words are same and in same sequence for both source and target terms, including punctuation
2. Match all only	String match where all words are same but not necessary in same order; additional words not allowed in target term
3. Match all	String match where all words are same but not necessary in same order; additional words allowed in target term
4. Partial match	String match where one or more words in source term is found in target term
5. Semantic match	For inactive concepts found use historical relationships of Was-A Same-As, May-Be-A, Replaced-By to find current concepts
6. Unmappable	Assigned when no match is found

Table 2. Mapping algorithms used in this study

Normalization Steps

In addition to using the original SNOMED CT terms and the ICD-10-CA long descriptions in mapping, we normalized all of these original terms to remove "noise" such as genitives and spelling errors using the Unified Medical Language System (UMLS) normalization steps, as shown in Table 3a [6]. To improve successful mapping, we expanded step-2 to remove both "stop words" and "exclude words," as well as SNOMED prefixes, shown in Table 3b. For step-5 we included both the lookup and stemming methods to uninflect the phrase. The lookup method uses the UMLS SPECIALIST Lexicon's inflection table with ~1 million entries, whereas the stemming method uses the computational technique first published by Porter Stemming that reduces word variants to a single canonical form [7,8].

Steps 1 to 6	Example
Remove genitive	Hodgkin***'s*** disease, NOS → Hodgkin diseases, NOS
Remove stop words	Hodgkin diseases, ***NOS*** → Hodgkin diseases,
Convert to lowercase	***H***odgkin diseases, → hodgkin diseases,
Strip punctuation	hodgkin diseases***,*** → hodgkin diseases
Uninflect phrase	hodgkin disease***s*** → hodgkin disease
Sort words	***hodgkin*** ***disease*** → disease hodgkin

Table 3a. UMLS six normalization steps[7, slide 20]

Step-2	Explanation
Stop words	Frequent short words that do not affect the phrase: and, by, for, in, of, on, the, to, with, no, and (nos)
Exclude words	Words that may change meaning of the word but if ignored help to locate a term otherwise missed: about, alongside, an, anything, around, as, at, because, before, being, both, cannot, chronically, consists, covered, does, during, every, find, from, instead, into, more, must, no, not, only, or, properly, side, sided, some, something, specific, than, that, things, this, throughout, up, using, usually, when, while, without
SNOMED Prefixes	[X] – concepts with ICD-10 codes not in ICD-9 [D] – concepts in ICD-9 XVI and ICD-10 SVII [M] – morphology of neoplasm concepts in ICD-O [SO] – concepts in OPCS-4 chapter Z in CTV3 [Q] – temporary qualifying terms from CTV3 [V] – concepts in ICD-9 and ICD-10 on factors influencing health status and contact with health services (V-codes and Z-codes)

Table 3b. Expanded UMLS normalization step-2

Reverse Mapping Process

The reverse mapping of ICD-10-CA terms to SNOMED CT concepts involved cycling through the mapping algorithms one at a time to find the best candidate SNOMED CT concepts as the target terms. For each algorithm we always started with the original terms, then the UMLS normalized terms, followed by the stemmed terms. In each cycle, we would review the candidate concepts found to see if it was a match, and if so, what type of match it was based on the algorithm applied. When no matching concepts were found, we would label the term as unmappable. Our experience with the matching techniques was that, the sooner we could find a match in the cycle, i.e. first-match, the greater confidence we would have that the candidate concept is appropriate. The preferred order of matched terms was always exact-match first, match-all-only, then match-all, with partial-match last. Whenever inactive concepts were found a semantic-match was done to find the current concepts through their historical relationships. During mapping we tallied frequency statistics on the different types of matches with summary/detailed outputs. Only the first-matches were counted to determine the effectiveness of each mapping algorithm.

Comparison with UK SCT-ICD10 Map

To determine the accuracy of the mapping results from this study, we compared our output with the UK SNOMED CT to ICD-10 (SCT-ICD10) cross map. To do so, the 5,000 ICD-10-CA codes were matched with the *TargetCodes* of the *SCT_CrossMapTargets* table from the July 2007 version of the IHTSDO distribution set [1]. While the UK cross map is from SNOMED CT to ICD-10 and not ICD-10-CA, the two ICD versions share many similar codes. Thus, if the ICD-10-CA code was found among the *TargetCodes* of the UK map, we would look up the *SCT_CrossMaps* table to find the corresponding SNOMED concepts. If multiple similar SNOMED concepts were found, they would be filtered to include only the unique SNOMED concepts. Each of the concepts found were then compared with our mapping output from matches found by the exact-match, match-all-only and match-all algorithms.

RESULTS

Summary of Mapping Output

Of the 5,000 ICD-10-CA descriptions used in this study, we were able to match 1,619 source ICD terms (32.38%) to 2,625 target SNOMED concepts by the exact-match technique. Next, we matched 63 ICD terms (1.26%) to 87 SNOMED concepts by match-all-only; another 1,478 ICD terms to 4,829 concepts by match-all; and 1,839 ICD terms to ~25 million concepts by partial-match. One ICD term *C8800 Waldenstr* was umappable. A summary of the mapping output by match-type is shown in Table 4.

Match Type	Source	Target	Percentage
Exact match	1,619	2,625	32.38%
Match all only	63	87	1.26%
Match all	1,478	4,829	29.56%
Partial match	1,839	24,950,238	36.78%
Unmappable	1	0	0.02%
Total	**5,000**	**24,957,779**	**100.00%**

Table 4. Summary of Mapping Output

Detailed Analysis of Mapping Output

Each ICD term was cycled through all the matching techniques to determine the number of candidate target SNOMED concepts found for each match type. The first-match reported for each match type excluded the target concepts already identified in previous iterations to avoid duplicate counting. We tracked not only the total matches but also which technique found the first match. The output produced suggested exact-match, match-all-only and match-all could be considered as successful matches, since they returned one or more identical or similar SNOMED concepts based on the ICD term provided. The number of first-matches found for these match types by ICD Chapter are shown in the Appendix. One can see that the percentages of matches were very low for Chapters *IV Endocrine, nutritional and metabolic diseases* at 36%; *XIII Diseases of the musculoskeletal system and connective tissue* at ~36%; and *XV Pregnancy, childbirth and the puerperium* at ~4%. Of the overall 3,160 ICD terms or ~63% that were mapped to one or more SNOMED concepts, most were found by exact-match and match-all during the first-match. The profiles of first-matches found by each match type are briefly described below.

Exact Match – Table 5 shows 1,237 original ICD terms had exact-matches with 2,064 candidate concepts. Another 364 ICD terms had exact-matches with 527 concepts using the UMLS normalized version, and 18 ICD with 34 concepts using the stemmed version. In all, 2,625 candidate SNOMED concepts were found, which means that there were multiple exact matches for some of the ICD terms.

Exact Match	First Match	Target
Original Term	1,237	2,064
UMLS Version	364	527
Stemmed Version	18	34
Total	**1,619**	**2,625**

Table 5. Exact match output

Match All Only – Table 6 shows 33 original ICD terms had match-all-only with 48 candidate concepts; 29 UMLS normalized terms had 37 concepts, and 1 stemmed term had 2 only. In all, 87 candidate SNOMED concepts were found, which means that there were multiple match-all-only for some terms.

Match All Words Only	First Match	Target
Original Term	33	48
UMLS Version	29	37
Stemmed Version	1	2
Total	**63**	**87**

Table 6. Match all only output

Match All Words – Table 7 shows 1,343 original ICD terms had match-all with 4,558 candidate concepts; 114 UMLS normalized terms had 217 concepts, and 21 stemmed terms had 54. In all, 4,829 SNOMED concepts were found, which means that there were multiple match-all for some terms.

Match All Words	First Match	Target
Original Term	1,343	4,558
UMLS Version	114	217
Stemmed Version	21	54
Total	**1,478**	**4,829**

Table 7. Match all words output

Partial Match – Table 8 shows 1,839 ICD terms had partial-matches with 25 million SNOMED concepts. We found the results of partial matches to be more unpredictable than the previous match types. If a source term was long and contains common words such as *disorder* or *procedure*, the results returned could be numerous as only one word from the source term needed to be present in the target term.

Partial Match	First Match	Target
Original Term	1,839	24,950,238
UMLS Version	0	0
Stemmed Version	0	0
Total	**1,839**	**24,950,238**

Table 8. Partial match output

Comparison with SCT-ICD10 Map

Six comparisons were made between our mapping output and the UK map to see if: (a) both contained the same results; (b) both contained similar results; (c) both contained dissimilar results; (d) only UK map contained the results; (e) only our mapping output contained the results; (f) both had unmappable results. The overall results are shown in Table 9. Only (b), (c) and (f) are illustrated in this paper.

Type of comparison	Frequency	Percentage
Contained exactly same results	11	0.22%
Contained similar results	2,401	48.02%
Contained dissimilar results	122	2.44%
UK map with results only	896	17.92%
Mapping outputs with results only	370	7.40%
Both had unmappable results	1,200	24.00%
Total	**5,000**	**100.00%**

Table 9. Comparing UK map and mapping outputs

Similar Results - Where both maps contained similar results, the UK map usually had more mapped terms than our output, as shown in Table 10. An example is with the ICD term *Q61.2 Polycystic kidney, autosomal dominant* where the UK map had six SNOMED concepts but only four in ours.

Description	Total
UK map had more results than mapping outputs	2,125
Mapping outputs had more results than UK map	224
UK and mapping outputs had same no. of results	63
Total	**2,401**

ConceptId	Fully Specified Name	UK	CA
66091009	Congenital disease (disorder)	√	
204955006	Polycystic kidney disease	√	
204962002	Multicystic kidney (disorder)	√	
28728008	Polycystic kidney disease, adult type (disorder)	√	√
253878003	Adult type polycystic kidney disease type I (disorder)	√	√
253879006	Adult type polycystic kidney disease type II (disorder)	√	√
274567009	[EDTA] Polycystic kidneys, adult type (dominant) associated with renal failure (disorder)		√

Table 10. Comparing both with similar results

Dissimilar Results – Where both had dissimilar results, our output were more specific as each concept must contain all the words in the source term. For 100 (82%) of these terms the UK map had more candidate concepts; for 9 terms (7.4%) both had same number of concepts; whereas for 13 (10.7%) our mapping output had more concepts. An example is the ICD term *S597 Multiple injuries of forearm*, shown in Table 11, where both maps had four concepts but none are similar.

ConceptId	Fully Specified Name	UK	CA
122549002	Injury (disorder)	√	
125596004	Injury of elbow (disorder)	√	
210557006	Severe multi tissue damage lower arm (disorder)	√	
210558001	Massive multi tissue damage lower arm (disorder)	√	
210860005	Injury of multiple blood vessels at forearm level (disorder)		√
211290004	Multiple superficial injuries of forearm (disorder)		√
212308001	Injury of multiple nerves at forearm level (disorder)		√
212464002	Injury of multiple muscles and tendons at forearm level (disorder)		√

Table 11. Comparing both with dissimilar results

Unmappable Results – These were in almost every ICD chapter but most notable in *XVII: Congenital malformations, deformations and chromosomal abnormalities; XIX: Injury, poisoning and certain other consequences of external causes; and XIII: Diseases of the musculoskeletal system and connective issue* (Table 12). It is possible these ICD terms have further refinement making it difficult to find concept and lexical matches. An example is the ICD-10-CA term *O2450 Pre-existing Type 1 diabetes mellitus arising in pregnancy,* which could be refined as: *delivered with or without antepartum condition (1), delivered with postpartum complication (2), or antepartum condition or complication (3).*

Chapter	Range	Freq	%
XVII: Congenital malformations, deformations, and chromosomal abnormalities	Q00-Q99	292	24.33%
XIX: Injury, poisoning and certain other consequences of external causes	S00-T98	278	23.17%
XIII: Disease of the musculoskeletal system and connective tissue	M00-M99	207	17.25%
IV: Endocrine, nutritional and metabolic diseases	E00-E90	119	9.92%
XX: External causes of morbidity and mortality	V01-Y98	60	5.00%
		956	79.67%

Table 12. Unmappable ICD-10-CA terms

DISCUSSION

Lessons and Issues

This study was our initial effort to apply a set of mapping algorithms on a set of ICD-10-CA terms to find the matching target SNOMED concepts. Our output showed most of the matches were found using the exact-match and match-all algorithms. The match-all-words-only algorithm did not add a great deal to the number of matches found, and the partial-match was considered too unpredictable with respect to the candidate target concepts returned. Due to space limitation, we did not report on additional matches found after normalization with UMLS and stemming techniques were applied to the original ICD terms, or those found by semantic matching.

A major issue is how one should define "successful match." In our output we had just over 60% of the matches found by exact-match and match-all, which we reviewed and deemed correct. However, more formal validation preferably by an independent source is needed. While our results showed successful matches in only ~63% of the 5,000 ICD-10-CA codes, we were surprised to find the UK cross map had similar successful matches of ~68% against the same 5,000 ICD-10-CA codes (see Table 9). Equally intriguing were the different matches found between the two maps. Almost 50% of the concepts found were similar but not identical, whereas ~20% were dissimilar or found only in the UK map. One possible explanation is the minor differences that exist between ICD-10 and ICD-10-CA with respect to the addition, subdivision, deletion and revision made in some ICD-10-CA chapters. Another is that a concept-based method was used to create the UK cross map, which seemed to outperform the lexical techniques in this study. One possible solution to improve mapping precision is to combine methods, such as the use of semantic and lexical mapping between SNOMED CT and ICD-9-CM by Fung.[9]

Another issue is the extent that our semi-automated matching algorithms can aide in the cross-mapping process by health records staff when encoding the inpatient discharge abstracts. The current abstracting process is mostly an intellectual and manual exercise. As such, explicit cross-mapping guidelines need to be established, including the use of any computer-based mapping tools, to improve this abstracting process. With our mapping algorithms, a consensus-based process is needed for the health record staff to verify the accuracy of the ~63% successful matches. Guidelines are also needed to reconcile the remaining ~37% partially-matched terms.[2,10] Still, we contend there is merit in exploring the use of reverse mapping with lexical algorithms to identify candidate SNOMED concepts for a given set of ICD-10-CA terms. Our next steps are to enhance the mapping algorithms to include contexts, incorporate these algorithms into the abstracting process, and conduct further field evaluation. Last, the idea of applying reverse mapping to identify candidate SNOMED CT concepts for a set of mapping terms can be a helpful approach when creating a cross map from SNOMED CT to another terminology system.

Implications

This study provides a glimpse of the feasible mapping methods that could eventually lead to a SNOMED CT to ICD-10-CA cross map for Canada. We believe the intent, methods and results of this current study should be of interest to those responsible for secondary use of patient discharge abstracts in epidemiological and statistical reporting. The notion of reverse mapping is also highly generalizable to include the encoding of local terms that already exist in legacy systems within many health organizations to a reference terminology such as SNOMED CT.

Acknowledgments

We wish to thank the Provincial Ministry that provided the 5,000 ICD-10-CA codes for the study. We also thank Ms. Robyn Kuropatwa in facilitating the process to obtain the ICD codes from the ministry. Funding support for this work was provided by the Canadian Institutes for Health Research through its Strategic Training Initiative. Note that the views presented in this paper are those of the authors only and do not represent the official position of any Canadian government agencies.

REFERENCES

1. IHTSDO, International Health Terminology Standards Development Organization. *SNOMED Clinical Terms Technical Reference Guide*. International Release, July 2007.
2. Bowman S. Coordination of SNOMED CT and ICD-10: Getting the Most out of Electronic Health Record Systems. *Perspectives in Health Information Management*, Spring 2005.
3. McBride S, Gilder R, et al. Data mapping. *Journal of American Health Information Management Association* 2006; 77(1): 44-48.
4. CIHI, Canadian Institute for Health Information. Canadian Coding *Standards for ICD-10-CA and CCI for 2006.* Ottawa, Canada. 2006.

5. Lee DHK. *Reverse Mapping ICD-10-CA to SNOMED CT*. UVic Master of Science research project report, Oct 2007. Unpublished.
6. National Library of Medicine. *The SPECIALIST Lexicon*. http://lexsr3.nlm.nih.gov/LexSysGroup/Projects/Summary/lexicon.html
7. Kleinsorge R, Willis J, et al. UMLS Overview – Tutorial T12. *AMIA Annual Symposium* 2006. http://165.112.6.70/research/umls/pdf/AMIA_T12_2006_UMLS.pdf. Jan15/2006.
8. Goldsmith JA, Higgins D, Soglasnova S. *Automatic Language-specific Stemming in Information Retrieval.* Springer-Verlag Berlin Heidelberg 2001.
9. Fung KW, Bodenreider O, Aronson AR, Hole WT, Srinivasan S. Combining lexical and semantic methods of inter-terminology mapping using the UMLS. In Kuhn K. et al. (Eds) *MedInfo 2007*, p605-610. IOS Press, 2007.
10. Vikstrom A, Skaner Y, et al. Mapping of the categories of the Swedish primary health care version of ICD-10 to SNOMED CT concepts: Rule development and intercoder reliability in a mapping trial. *BMC Medical Informatics and Decision Making* 2007;7:9.

Appendix. Mapping Output for top 5,000 ICD-10-CA codes by ICD Chapter

Chapter	Title	Range	Source	Exact	Only	All	Total	Percent
I	Certain infections and parasitic disease	A00-B99	136	47	2	57	106	77.94%
II	Neoplasms	C00-D48	343	174		58	232	67.64%
III	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism	D50-D89	80	35	1	20	56	70.00%
IV	Endocrine, nutritional and metabolic diseases	E00-E90	225	56	1	24	81	36.00%
V	Mental and behavioural disorders	F00-F99	218	66	3	141	210	96.33%
VI	Diseases of the nervous system	G00-G99	196	75	1	56	132	67.35%
VII	Diseases of the eye and adnexa	H00-H59	89	56	3	18	77	86.52%
VIII	Diseases of the ear and mastoid process	H60-H95	42	24		11	35	83.33%
IX	Diseases of the circulatory system	I00-I99	279	136	1	74	211	75.63%
X	Diseases of the respiratory system	J00-J99	165	67	4	41	112	67.88%
XI	Diseases of the digestive system	K00-K93	276	136	9	56	201	72.83%
XII	Diseases of the skin and subcutaneous tissue	L00-L99	105	42		20	62	59.05%
XIII	Diseases of the musculoskeletal system and connective tissue	M00-M99	383	78	1	61	140	36.55%
XIV	Diseases of the genitourinary system	N00-N99	226	120	3	48	171	75.66%
XV	Pregnancy, childbirth and the puerperium	O00-O99	313	5	1	6	12	3.83%
XVI	Certain conditions originating in the perinatal period	P00-P99	169	57	17	47	121	71.60%
XVII	Congenital malformations, deformations, chromosomal abnormalities	Q00-Q99	205	105	2	57	164	80.00%
XVIII	Symptoms, signs and abnormal clinical and laboratory findings not elsewhere classified	R00-R99	181	99	2	52	153	84.53%
XIX	Injury, poisoning and certain other consequences of external causes	S00-T98	691	175	8	169	352	50.94%
XX	External causes of morbidity and mortality	V01-Y98	297	9	4	249	262	88.22%
XXI	Factors influencing health status and contact with health services	Z00-Z99	333	29		199	228	68.47%
XXII	Morphology of neoplasms	8000/0-9989/1	28	28			28	100.00%
XXIII	Provisional codes for research and temporary assignment	U00-U99*	20			14	14	70.00%
	Total		**5,000**	**1,619**	**63**	**1,478**	**3,160**	**63.20%**

Interoperability of Data Models and Terminology Models: Issues with using the SNOMED CT terminology

Rahil Qamar S., Ph.D.[1], Jay (Subbarao) Kola, M.B.B.S.[1], Alan L. Rector, M.D.,Ph.D.[1]
1 Bio-health Informatics Group, University of Manchester, Manchester, U.K.

Abstract

Work in the field of recording standard, coded data in electronic health records and messages is important to support interoperability of clinical systems. It is also important for reducing medical errors caused by misinterpretation and misrepresentation of data. Standardisation of structured and unstructured data to one or more terminologies such as SNOMED-CT, or ICD requires the help of various integration procedures. We have previously highlighted issues in data models (*open*EHR Archetypes) when mapping to a terminology model (SNOMED CT) [1]. In this paper, we describe issues with terminology models (SNOMED CT) when aligning the concepts to a data model (*open*EHR Archetypes).

Terminologies and data models play an important role in building structured EHRs and achieving semantic interoperability. Semantic interoperability requires that all recorded data conforms to some reference terminology in order to interpret and reuse it uniformly in all partaking information systems. In the medical domain, standardising data is of great significance, as controlling the vocabulary used to record patient data is critical to making EHRs safe for exchange and reuse.

The paper recognises the value of SNOMED CT but demonstrates the difficulties of working with it at an integration level. The difficulties in integration arise primarily due to the semantic gaps in the content of the structured data models and terminology models. The same issues might also arise with data obtained from unstructured sources. Despite the broad coverage that SNOMED offers, there are several concepts that are missing. An efficient process for submission of concepts for inclusion is in need, along with formal rules for post coordination.

We believe that in order to achieve the overall objective of semantic interoperability, it is imperative that both data and terminology models are developed with the aim of being able to integrate their clinical content. It is important that both modeling communities are not only cognizant of each others existence but also work closely with each other to ensure that conformance is built into the systems from conception stage. These conformance or compatibility rules should be extended to all other stages of the modeling process i.e. at design time, data integration time, as well as at run-time. It is only then that true interoperability will be achieved, making it possible to build safer health care systems. Reliable and high quality data in these systems will improve the functioning of all health care units heavily dependent on data, reducing medical errors and ultimately providing safer and better patient care.

Reference

[1] Rahil Qamar, Jay Kola, and Alan Rector. Unambiguous data modeling to ensure higher accuracy term binding to clinical terminologies. AMIA 2007 Annual Symposium, November 2007. Chicago, U.S.A.

Essential SNOMED: Simplifying SNOMED CT and Supporting Integration with Health Information Models

Peter MacIsaac, MB.BS, FRACGP, MPH[1], Don Walker, MB.BS [2] , Rachel Richesson PhD, MPH[3], Heather Grain FACHI[4],Peter Elkin, MD[5], Jon Patrick PhD[6]

[1]Terminology Central, Canberra, Australia, peter@macisaacinformatics.org;
[2]University of Adelaide, Adelaide, South Australia, donald.walker@adelaide.edu.au;
[3]University of South Florida College of Medicine, Tampa, FL, USA, richesrl@epi.usf.edu;
[4]LaTrobe University, Melbourne, Victoria, Australia, h.grain@latrobe.edu.au;
[5]Mayo Clinic, Rochester, MN, USA, elkin.peter@mayo.edu
[6] University of Sydney, NSW, Australia, jonpat@cs.usyd.edu.au

ABSTRACT

SNOMED CT (SCT) has been designed and implemented in an era when health computer systems generally required terminology representations in the form of singular pre-coordinated concepts. Consequently, much of SCT content represents pre-coordinated concepts and their relationships. In this conceptual paper the role of pre- and post-coordinated terminology expressions are considered in the context of the current development direction of Electronic Health Records and the use of communications and knowledge repositories. The move from current SCT structures to an implementation form of SCT that focuses on "atomic concepts" will support post-coordination and terminology binding to information models. This core or "essential" SNOMED CT - called SNOMED Essential Terminology (S-ET) - would be smaller in terms of core concept numbers, simpler, easier to maintain and more intuitive for implementers. Our proposed implementation form of SNOMED CT would contain only "atomic concepts" with their attendant hierarchies and relationship data. These would be supported by a strict model for representing current and future pre-coordinated concepts based on the use of an existing specific post- coordination expression, grammar, or representation. The resulting concept expressions would be post-coordinated from a smaller core of atomic components. Using definitional relationships, the proposed implementation form could equate existing pre-coordinated terms with post-coordinated representations, allowing SCT to maintain links with legacy data. A strategy for testing and implementing this approach is discussed and empirical research and feasibility testing is recommended.

INTRODUCTION

SNOMED CT (SCT) is becoming the international standard clinical terminology with a new international licensing and governance process which makes it widely accessible. The adoption of SCT by multiple countries was influenced by many published studies demonstrating its comprehensive coverage [1-4] and advanced structural features. SNOMED CT has antecedents in the College of American Pathologists family of terminologies, the UK National Health Service Read Codes. As with any living language, it has absorbed content from a number of other terminologies and classifications. SCT contains concepts and terms that describe the "language of use" as well as concepts which define the "language of meaning"[5-7]. Consequently, SCT contains many pre-coordinated concepts that have varying levels of semantic complexity alongside the component or essential concepts which are themselves the building blocks of these complex clinical expressions. While there are sound historical and ongoing pragmatic reasons for this evolutionary development, the resulting mix of concept structures makes implementation within various information models complex and prone to variation. Currently, SCT is "cluttered" with pre-coordinated terms that are incompletely defined by the internal information model that exists within SCT, making transformations between existing pre-coordinated terms and post- coordinated representations difficult to achieve. This result limits opportunities for interoperability across systems, [8] which is one of the key objectives of a controlled terminology.

This conceptual paper brings to notice issues that are well known within the health informatics community and proposes what may, at first glance, seem to be radical surgery. This proposal is in reality an extension and combination of existing features of SCT to create a more tractable solution to support both SCT development and the art and science of terminology development. This paper is not a report of a quantitative analysis of SCT structures or experimental results of the types of change proposed. These should come later, if the fundamental proposition is believed to be sound and a potential contribution to terminology development and maintenance methods.

The computational representation of data is a combination of the use of information models and terminology. We propose a variation, restructure and extension of the current SNOMED-CT terminology to support implementation in various information models.

Using a pragmatic approach the SCT would be altered in that existing pre-coordinated concepts would be identified, flagged and then defined through linkage to their atomic concepts and relationship types. The atomic or essential concepts would continue to be placed in logical and definitional hierarchies and relationship structures and subject to the use of description logic for definition, inference, and classification purposes. Existing pre-coordinated SCT concepts would retain their identifiers and be linked to the modified terminology as “pre-defined-post-coordinated concepts”, and would be logically equivalent to any post-coordination representing the same meaning. The retention of pre-coordinated concepts and the specification of their computational definitions would allow pre-coordinated terms to be used in interface applications, as pre-coordinated terms can be useful in helping data entry to be more consistent: supporting the language of use. If users have a retrieval list of pre-coordinated concepts that have post-coordinated equivalents, application developers can encourage users to use a more consistent post-coordinated form or to use entry terms that have relationships to post-coordinated expressions using fully-defined atomic concepts.

This approach to the re-organization of SCT with the formal expression of the canonical form for pre-coordination is described as “SNOMED Essential Terminology” (S-ET), or simply “Essential SNOMED”; the name coined by Dr. Walker when first describing this approach. This paper describes the case for change in SCT representation and advantages of moving to this representation, the background to the development of this approach, a representation model for pre-coordinated concepts, and an implementation perspective. Simple examples have been selected, not to prove the feasibility of this approach, but to illustrate the principles. The need for a more technically challenging and quantitative approach to evaluation of this proposal is recognized and discussed.

TERMINOLOGIES AND INFORMATION MODELS

It is now widely accepted that health information storage is achieved through a combination of the use of controlled terminologies and standardized data models or architecture, yet the boundaries between the models used for terminology construction and health record construction are blurred. [9-12] The HL7 TermInfo project attempted to resolve this by providing guidance on how SNOMED CT could be used in HL7 version 3 messages and data structures. [12-14]

An example of this terminology model - information model interface is the question of whether concept negation should be managed within the terminology or within the data model. Should the negation be expressed as part of the terminological unit: “no history of breast cancer”, or as different components within an information model: “history of breast cancer” + “negative”?. [15] The semantics can be represented in the terminology as a pre or post coordinated concept or in a combination of the data model and terminology. The machinery to support this latter approach is contained in standard information models such as the HL7 Reference Information Model (RIM).[12] HL7’s TermInfo working group has recommended that when SNOMED is being used in HL7 V3 models, negation be managed in the terminology and that the model based approach to attaching a negation indicator be deprecated. This issue points to the need for sufficient flexibility in the management of post-coordination to allow for the transformation of concept structures and modifiers between the various options. The existence of other data and information models (e.g., CDISC) – which might develop and endorse their own guidance for use of complex terminologies such as SCT - suggest that standardization of SCT terminology use in HL7 (RIM-based) applications might not guarantee interoperability with applications using other information models. [16]

While the issue of terminology and information model interaction is somewhat independent of the way that coordination of complex concepts occurs, there is a need for both pre and post-coordinated approaches to co-exist to fully support the spectrum of information representation. It is also recognized that equivalence between pre-coordinated and post-coordinated concepts has to be established to maintain consistency in interpretation of terminology and between concept representation using different combinations of terminology and information model binding. In the current SCT infrastructure this is achieved using computation and testing the equivalence of the canonical form of the two terminological variations. This requires that all of the atomic or component concepts and pre- coordinated concepts are fully defined - not the case in practice. Having a formal definition explicitly developed for current and future pre- coordinated concepts within SNOMED would support the recognition of equivalence. [8, 17, 18]

ISSUES WITH SNOMED-CT IMPLEMENTATION

Several studies have shown that inter-rater reliability of SCT coding is poor, at least in part due to the complexity of the SCT structure and the inconsistency of existing content. [3, 19-21] This paper proposes that a simpler, more consistent representation of SCT will reduce confusion and improve the quality of SCT implementation. This would need to be tested once working subsets of the S-ET have been developed and so examine the impact on coding consistency of the interaction between the information model and the use of differently coordinated terminology.

SCT size will certainly grow as new countries adopt it, especially when it becomes the terminology to support the many uses of coded clinical data, such as public heath. Trying to keep up with the need for language of use through definition of pre-coordinated concept phrases is a recipe for “combinatorial explosion” in the size of a terminology. This is bad enough in a terminology of simple structure, yet in one of SCTs complexity and richness of function, the impact is especially significant. A key technical challenge involves keeping the terminology to a manageable size and level of complexity so that it is both maintainable and supports end users’ applications. A second challenge for SCT maintenance is to allow compatibility with historical versions used by legacy applications while maintaining relevance as the core terminology resource for the current and future generations of health information systems. The model proposed in this paper will support both of these objectives.

BACKGROUND TO S-ET DEVELOPMENT

In 1999 a combined pre and post-coordinated model for a medicines terminology was proposed by two of the authors (DW and PM) for Australia, based on an architecture designed earlier by DW for a proprietary drug information service. Both pre-coordinated concepts and their contained atomic components and relationship types were accommodated. The Essential SNOMED notion, which was initially canvassed informally within the health terminology community in 2001, was further developed following a comparative technical analysis of several terminology options that were then being investigated for use in Australian General Practice [23] and which subsequently recommended use of SCT leading in time to Australia becoming an early adopter of a national SCT license. A review of candidate terminologies at that time for use in General Practice examined several options. One terminology, DOCLE, was constructed of atomic concepts, joined by operators using a Bachus Naur Form (BNF), a standard system for representation of computable expressions using syntax or rules. [24] What was notable was the extensive use of pre-coordinated terms that were constructed from atomic elements. For example “cancer@breast” was a pre-coordinated concept for “breast cancer”, yet it is constructed using a post-coordination model of atomic concepts and the location operator .“@”. The process of normalization of DOCLE for inclusion in a terminology service found that a number of atomic concepts needed to be created to support existing content. Considering this approach and drawing on prior experience with the development of a medicines terminology requiring a full set of atomic elements which were combined to create fully defined pre-coordinated medicines concepts, it was postulated that the SCT terminology could be significantly simplified by creating a separate data structure for the pre-coordinated concepts where these were parsed and then described in a post- coordination grammar. [25]

DESIGN OF ESSENTIAL SNOMED

Essential SNOMED would contain a complete set of “atomic concepts” from which all other concepts could be constructed by post-coordination. These atomic concepts would be carefully crafted into their hierarchies and defined by their relationships. SNOMED CT most likely contains many - if not most - of these atomic concepts. They would consist of both primitive and fully defined concepts. The large number of pre-coordinated concepts that are not in the above group should be “flagged” in the complete SNOMED CT data structure as “predefined-post-coordinated equivalent concepts”, and eventually associated with their post-coordinated defining atomic concepts using a formal post coordination or compositional syntax as is already described. This group of pre-coordinated concepts would not rely on their hierarchical position or SNOMED relationships for their definition - instead they would be defined by the compositional expression (or formalism) used to construct their atomic post-coordinated concepts (combination of existing atomic concepts), as described above. All the existing pre-coordinated concepts in SCT could remain, along with their identifier for use where the situation required this approach. Pre-coordinated concepts being added to SCT would also follow this pattern.

For example, the pre-coordinated concept "Cellulitis of the left foot with osteomyelitis of the third metatarsal without lymphangitis" can be expressed using atomic concepts and relationship types, is shown in Table 1. The concepts and relationship concepts that comprise the definition would all be considered core Essential SNOMED content.

Oper-ator	**Disorder**	**Has-FindingSite**	**Has-Laterality**
	Cellulitis	Foot	Left
and	Osteomyelitis	third metatarsal	
Without	Lymphangitis		

Table 1 – Definitional relationships of an existing pre-coordinated SCT concept.

The current SNOMED CT terminology model specifies relationships between concepts and terms, but does not make a distinction between post-coordinated concepts expressions and pre-coordinated concepts. We propose that this distinction be made explicitly, as a tool to assist in SNOMED-CT terminology maintenance and implementation. Figure 1 describes the way that the new architectural elements could be linked with existing S- CT structures which are represented by the three elements placed at the right hand side of the figure.

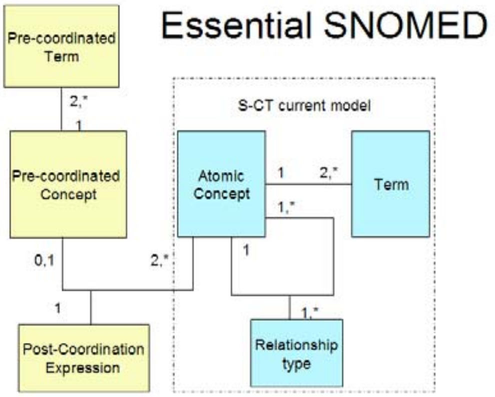


Figure 1 – Conceptual terminology model for Essential SNOMED.

DISCUSSION

At the outset it is acknowledged that this proposal is grounded in the excellent overall design and management features of SCT.

The advantages of this proposed structure for SCT are reduced size and complexity for ease of implementation and maintenance. An inevitable outcome would be a reduction in the combinatorial explosion that occurs when rampant pre-coordination of concepts and phrases occurs, yet this comes at the cost of introducing a new element in the post coordination expression that links the pre-coordinated concepts to their atomic elements. The core terminology concepts and hierarchies should be however much simplified.. The core of S-ET would grow some as new atomic concepts were added. The S-ET structure would be expressively intuitive as its approach to concept representation would support concept constructions. Hierarchical simplification would result as the definition of the many pre-coordinated-concepts would be independent of immediate hierarchies or relationships – S-ET would use the compositional expression to link with hierarchies and defining relationships of the atomic concepts. Existing approaches to canonical forms would continue and allow equivalence testing between different pre-coordinated concepts and post-coordinated expressions. Pre-coordinated concepts would still be able to be represented in a hierarchical arrangement to support inference and subsumption, however these could be calculated rather than explicit expressed as happens currently in SCT. In this model the hierarchical relationships would be inferred rather than the canonical form.

Equating pre- and post-coordination may be easier, as the computational form is actually specified for concepts within SCT. It is acknowledged that the current approach in SNOMED is not comprehensive due to incomplete set of canonical representations and possible lack of semantics to fully describe the meaning of existing semantically complex pre-coordinated concepts. Both the pre-coordinated form and the various representations of the post-coordinated concept are valid ways of describing the same concept. The first is more aligned with human interpretation and the second supports computer processing of the data. It is clear that both forms of concept representation are needed and both have to be supported by clinical terminologies such as SNOMED CT. The approach recommended in SNOMED Essential Terminology is believed to be consistent with the current SCT approach to canonical form definition.

This model is highly dependent upon an expressive and computable syntax for post-coordination. The process of moving to an S-ET distribution format will highlight any deficiencies in the current post-coordination methods and constraints as these will become explicit and subject to development. SNOMED –

CT authors can continue to develop pre-coordinated expressions if required. End users, particularly those who rely on the use of pre-coordinated concepts, will have the capacity to add locally relevant pre-coordination through a minor modification of the SCT way of managing local extensions, and in doing so would not require frequent change submissions to the core essential SNOMED terminology. As discussed earlier an S-ET model, coupled with an improved model for managing pre-coordination will support the terminology user interface.

One of the difficulties faced by SNOMED is the need to harmonize with widely used terminologies that are heavily structured on pre-coordination. LOINC and MEDCIN and most health classifications would be examples {28]. A SNOMED Essential Terminology would not need to include the pre- coordinated concepts imported from such terminologies, and could instead relate their concepts to SNOMED-CT by mapping which used the post-coordination syntax. Alternatively such pre-coordinated concepts could be placed along with existing SCT pre-coordinated content. The end result would provide the flexibility of incorporating or mapping to external terminologies, even though they may not share the same data models as SCT.

It is recognized that there are situations where pre-coordination is more efficient from a computational perspective, as in the recognition of commonly used text strings in natural language processing (NLP) applications. SNOMED Essential Terminology will allow the further development of such concepts without undue concern about the combinatorial explosion that might otherwise exist. NLP requires the consistent application of terminology and parsing of text. If a SNOMED Essential Terminology model is not adopted then it is likely that some equivalent derivative product will be created by necessity by these key application areas. Having a standard form will support consistency of output of different NLP applications.

One of the current strategies to simplify SNOMED is to restructure the relationship between terminologies and classifications. Removing or retiring classification concepts from SCT will allow them to reside in their respective classifications and have linkage to the clinical terminology by mapping or other formal constructs. SNOMED Essential Terminology proposes making a similar change to manage both the historical terminological clutter resulting from SNOMED's antecedents and use in legacy information systems. In addition it will meet the widely accepted need to continue to manage post-coordination in a modern terminology to support the computer-human interface.

Making the transformation to S-ET

The transformation to a SNOMED Essential Terminology would require a set of suitable "relationship-types" and an appropriate post-coordination representation form or "syntax" that catered for the "pre-defined-postcoordinated equivalent concepts". SNOMED has published a BNF for this syntax. This syntax describes the core SCT concepts, and their relationships. An XML equivalent (in addition or as an alternative) may be helpful for the current computer engineering environment. This paper is not exploring the relative merits of these approaches; however the process of defining the post-coordination equivalents of existing concepts will also provide a validity check on the completeness of the syntax or post coordination model, and as such is complementary to activities of the International Health Terminology Standards Development Organisation's (IHTSDO) Concept Model Special Interest Group.

As the "pre-defined-post-coordinated concepts" could be related back to their atomic components (which are themselves part of the SNOMED hierarchy and relationship structure) it would no longer be necessary to separately define the hierarchies or associations for the pre-coordinated concepts within the terminology. This does not preclude such constructs being employed, much like current indexing activity at run-time. These hierarchies could be machine classified. For example, if the phrase "fractured ankle" was compiled from two concepts as follows:

[problem, action or issue] = "fracture"
    [which has FindingSite] = "ankle"

Consequently , if it was necessary to locate "injuries of the lower limb", then the hierarchical ancestors of "fracture" would include "injuries" and those of "ankle" would include "lower limb".

The issue of what is and what isn't an atomic or pre-coordinated concept is subject to debate and the boundaries can be fuzzy. Is headache a single concept or a post-coordinated 'pain' with 'location' of 'head'? Technically, it should not be part of an S-ET based on atomic concepts but is it sufficiently common and semantically 'simple' enough to warrant inclusion? Under our proposal, "atoms refer to semantic units, not term labels or compound term labels. While it is clear that even single concepts can have compound names, it would be a conceptual error to consider a concept such as Hodgkin's lymphoma to be a pre-coordinated concept, whereas a "fractured right femur" is patently so.

Building essential SNOMED will be necessarily a pragmatic exercise which can cope with one or other of these forms or both. The consequences of the fuzziness in determining whether an existing concept should be managed as a pre or post coordinated concepts are not expected to be significant. All are still included in SCT.

New pre-coordinated concepts could be created (if required by information systems or user preference), although this temptation may best be resisted as it is expected that the requirements for pre-coordination would become less pressing with the introduction of standard information models (e.g., HL7 V3, archetypes or OpenEHR) and the advancement of Natural Language Processing (NLP) to support data entry.[26] Complex pre-coordinated concepts can sometimes be useful in encouraging consistency in representation where small nuances may be unintentionally instantiated where no difference in clinical meaning exists. For example: Colon cancer can be represented either as: Malignant Neoplasm - Has finding site – Colon; or as: Colon – HasSpecimen - Malignant Neoplasm. This ambiguity is undesirable, and the availability of pre-coordinated concept expressions at the interface level can prevent this type of variation, and set patterns for good practice in post- coordination within terminology services.

The atomic-concepts included in S-ET would be those that are necessary and appropriate to build pre-coordinated concepts that currently exist, or may be added subsequently, as well as the atomic concepts currently in use. The boundaries around atomic concept definition are often fuzzy as discussed above. Editorial rules would be required to consider inclusion of concepts that are not "semantically atomic" but are very common. A pragmatic approach would need to be developed, and the following may suggest one strategy:

1. The entries expected to be found as defined concepts in a large medical dictionary [27]; this would likely include items that have a distinct clinical meaning and are used frequently – e.g. Lung cancer; breast cancer; direct inguinal hernia; chest pain.
2. Those concepts that cannot be adequately defined by the composition of their post-coordinated concepts due perhaps to use of an uncommon or unsupported semantic type for the relationship between elements.

As a result many of the pre-coordinated concepts found in SCT diagnoses, findings and procedures would be excluded from the atomic- concept list and be placed in the pre-coordinated group.

It is clear that any change to the structure or representation forms of SCT may have an impact on reference set (subset) development, use within value-sets, and mapping to classifications and use in local extensions. These areas need to be further examined, however S-ET would not have a significant impact, as the current SCT and S-ET would contain the same concepts and relationships. There is a significant advantage for local extensions as local terminology experts could map new local concepts to atomic elements within SCT, hence gaining the benefits of classification and relationship modelling, without having to wait for formal inclusion in later releases. The development of reference sets based on concept and hierarchy selection would also include related pre-coordinated concepts.

The feasibility of remodelling large sections of SNOMED CT, particularly when there are competing priorities for terminology development, must be assessed. While a conversion strategy has not been covered in detail, the re-organization could consider using current SCT relationships - but with some care because of their known limitations. The size of the term string, the number of individual words, the presence of relationships, and a comparison with lists of terms extracted from medical dictionaries might help identify potential pre-coordinated concepts. It is possible that a functional result to create S-ET could result from flagging pre-coordinated concepts and terms, without substantially altering the publication structure. As with most terminology development, specific tools to manage the transition to S-ET would need to be developed, refined and the end result would need appropriate checking and quality control processes and upfront attention to ongoing maintenance.

While the first efforts at instantiation of the S-ET model may involve the restructure of arbitrary twigs and branches of the SNOMED hierarchical tree, an approach proposed would be to operate on concepts identified in large sub-setting exercises where terminology of use has been identified from analysis of actual clinical use in a specific domain such as intensive care [29] or general practice .[23]

CONCLUSION

This paper has proposed a modest alteration to the structure of SNOMED CT so that it supports the co-existence of pre and post coordination in a form that advances the basic structure of what might be regarded as good terminology practice [30].

The changes do not require any fundamental changes in SCT methods, but rather a structural extension and the incorporation of existing post-coordination methods of expression into the core terminology.

This paper outlines a number of issues with the current SCT architecture and proposes a solution which is consistent with its current design and which may have a number of advantages. If the proposed model creates resonance with the end users of SNOMED CT, it should be exposed to empirical testing and considered by the IHTSDO and their related organizations. The authors hope this paper stimulates discussion and feedback. We look forward to formal testing of these ideas for feasibility and acceptance.

ACKNOWLEDGMENTS

This paper was inspired by an original paper authored by Don Walker with contributions from Peter MacIsaac arising from the "GP Vocabulary Project" supported by the Australian Department of Health and Ageing. http://www.adelaide.edu.au/health/gp/research/current/informatics/#vocab).

The authors would like to acknowledge the contributions of Tad McKeon, anonymous KRMED reviewers, and numerous colleagues who have contributed in some way to this paper and our education and research on terminology and informatics.

This work has been supported in part by grants from the Centers for Disease Control and Prevention (PH00022 and HK00014; PLE).

REFERENCES

1. Wasserman, H. and J. Wang, *An applied evaluation of SNOMED CT as a clinical vocabulary for the computerized diagnosis and problem list.* AMIA Ann Symp Proc, 2003: p. 699-703.
2. Elkin, P.L., et al., *Evaluation of the Content Coverage of SNOMED CT: Ability of SNOMED Clinical Terms to Represent Clinical Problem Lists.* Mayo Clinic Proceedings, 2006. **81**(6): p. 741-748.
3. Warren, J.J. and R.P. Wilson. *Representing Cardiovascular Concepts in an Electronic Health Record Using SNOMED CT®.* in *American Medical Informatics Association Annual Symposium*. 2006. Washington, D.C.
4. Richesson, R.L., J.E. Andrews, and J.P. Krischer, *Use of SNOMED CT to Represent Clinical Research Data: A Semantic Characterization of Data Items on Case Report Forms in Vasculitis* Research Journal of the American Medical Informatics Association, 2006. **13**: p. 536-546.
5. Patrick, J. *Metonymic and Holonymic Roles and Emergent Properties in the SNOMED CT Ontology, Advances in Ontologies, M.H. Orgun & Tf, Meyer (Eds). Advances in Ontologies (AOW 2006)*, Tasmania, Conferences in Research & Practice in Information Technology, 72, pp61-68, 2006
6. Rector, AL, Qamar R, Marley T. *Binding Ontologies and Coding Systems to Electronic Health Records and Messages.* KR-MED 2006 – Biomedical Ontology in Action. November 8, 2006, Baltimore, Maryland, USA
7. Elkin PL, Brown SH, Lincoln MJ, Hogarth M, Rector A. *A formal representation for messages containing compositional expressions.* Int J Med Inform 2003 Sep;71(2-3):89-102.
8. Andrews,J.E.,et al,*Comparing Heterogeneous SNOMED CT Coding of Clinical Research Concepts by Examining Normalized Expressions.* Journal of Biomedical Informatics, 2008. **in press**.
9. Rector, A.L. *The Interface Between Information, Terminology, and Inference Models.* in *Tenth World Conference on Medical and Health Informatics: MedInfo-2001*. 2001. London.
10. Dampney, C.N.G., G. Pegler, and M. Johnson. *Harmonising Health Information Models - A Critical Analysis of Current Practice.* in *Ninth National Health Informatics Conference*. 2001. Canberra ACT, Australia.
11. Huff, S. and J. Carter. *A Characterization of Terminology Models, Clinical Templates, Message Models, and Other Kinds of Clinical Information Models.* in *AMIA Symposium*. 2000.
12. Chute, C.G., *Medical Concept Representation*, in *Medical Informatics. Knowledge Management and Data Mining in Biomedicine.*, Chen H., et al., Editors. 2005, Springer U.S. p. 163-182.
13. Markwell, D., *Meaning Well & Well meaning - HL7 TermInfo*. 2005, The Clinical Information Consultancy Ltd. www.clininfo.co.uk
14. HL7, *Health Level Seven*. 2005, Health Level Seven, Inc.
15. Elkin, P.L., et al., *A controlled trial of automated classification of negation from clinical notes.* BMC Med Inform Decis Mak, 2005. **5**(1): p. 13.
16. 16. Richesson, R.L. and J.P. Krischer, *Data Standards in Clinical Research: Gaps, Overlaps, Challenges and Future Directions.* Journal of the American Medical Informatics Association, 2007. **14**(6): p. 687-696.
17. Spackman, K.A. and K.E. Campbell, *Compositional concept representation using SNOMED: towards further convergence of clinical terminologies.* Proc AMIA Symp, 1998: p. 740-4.
18. Rector, A.L., et al., *The GRAIL Concept Modelling Language for Medical Terminology.* Artificial Intelligence in Medicine 1997. **9**: p. 139-171
19. Rothschild, A.S., H.P. Lehmann, and G. Hripcsak, *Inter-rater Agreement in Physician-coded Problem List*, in *American Medical Informatics Association*. 2005: Washington,

D.C.

20. Burkhart, L., et al., *Mapping parish nurse documentation into the nursing interventions classification: a research method.* Comput Inform Nurs, 2005. **23**(4): p. 220-9.
21. Chiang, M.F., et al. *Reliability of SNOMED-CT Coding by Three Physicians using Two Terminology Browsers*. in *American Medical Informatics Association Annual Symposium*. 2006. Washington, D.C.
22. Andrews, J.E., R.L. Richesson, and J.P. Krischer, *Variation of SNOMED CT Coding of Clinical Research Concepts among Coding Experts.* JAMIA, 2007. **14**: p. 497-506.
23. Walker, D. and e. al, *General Practice Vocabulary Project*. 2001, Commonwealth Department of Health and Ageing.
24. *Welcome to* DOCLE *Systems*. 2008, . http://www.docle.com.au/
25. Walker, D. and e. al., *Essential SNOMED, unpublished project report*, University of Adelaide.
26. Elkin, P.E., et al., *An Evaluation of the Content Coverage of SNOMED CT for Clinical Problem Lists.* Mayo Clin Proc, 2006. **81**(6): p. 741-8.
27. Elkin, P.L., et al., *Guideline and Quality Indicators for Development, Purchase and Use of Controlled Health Vocabularies.* International Journal of Medical Informatics, 2002. **68**(1-3): p. 175-186.
28. Rosenbloom ST et al. *Using SNOMED CT to Represent Two Interface Terminologies* In review
29. Patrick J, Herkes R, Ryan A. *Enhancement technologies for clinical information systems.* HISA NSW conference proceedings 2008.
30. Cimino JJ. *Desiderata for controlled medical vocabularies in the twenty-first century*. Methods Inf Med 1998;37(4–5):394–403

Comparing the Effects of Two Semantic Terminology Models on Classification of Clinical Notes: A Study of Heart Murmur Findings

Guoqian Jiang, Ph.D. and Christopher G. Chute, M.D., Dr. P.H.
Division of Biomedical Informatics, Mayo Clinic College of Medicine, Rochester, MN
(mailto:Jiang.Guoqian@mayo.edu)

Abstract

Objectives: We compared the effects of two semantic terminology models on classification of clinical notes through a study in the domain of heart murmur findings. *Methods:* One schema was established from the existing SNOMED CT model (S-Model) and the other was from a template model (T-Model) which uses base concepts and non-hierarchical relationships to characterize the murmurs. A corpus of clinical notes (n=309) was collected and annotated using the two schemas. The annotations were coded for a decision tree classifier for text classification task. The standard information retrieval measures of precision, recall, f-score and accuracy and the paired t-test were used for evaluation. *Results:* The performance of S-Model was better than the original T-Model ($p<0.05$ for recall and f-score). A revised T-Model by extending its structure and corresponding values performed better than S-Model ($p<0.05$ for recall and accuracy). *Conclusion:* We discovered that content coverage is a more important factor than terminology model for classification; however a templatestyle facilitates content gap discovery and completion.

Introduction

While modern terminologies have advanced well beyond simple one-dimensional subsumption relationships through the introduction of composite expressions, there is an emerging convergence of approaches toward the use of a concept-based clinical terminology with an underlying formal semantic terminology model (STM) [1]. SNOMED CT, the most comprehensive clinically oriented medical terminology system, currently adopts a foundation based on a description logic (DL) model and the underlying DL-based structure to formally represent the meanings of concepts and the interrelationships between concepts [2-3]. The existing SNOMED CT model is mainly pre-coordination oriented, i.e. containing many pre-coordinated terms, and also supports post-coordination. For example, a compositional expression "[*hypophysectomy (52699005)*] + [*transfrontal approach (65519007)*]" could be used to describe a more specific clinical statement than that only using the term "hypophysectomy (52699005)".

For a specific domain, a template model having a semantic structure with a coherent class of terms can be used as a formal representation [4]. This kind of model is mainly post-coordination oriented and a list of atomic terms is organized within a semantic structure. For example, the latest version of the International Classification of Nursing Practice (ICNP) uses a 7-Axis model to support the representation of nursing concepts and integrates the domain concepts of nursing in a manner suitable for computer processing [5].

One of the main goals of the semantic terminology models is to support capturing structured clinical information that is crucial for computer programs such as information retrieval systems and decision support tools [6]. Structured recording has the potential to improve information retrieval from a patient database in response to clinically relevant questions [1]. However, functional difference in retrieval performance has not been clearly demonstrated between these two different semantic terminology models.

In this study, we focus upon the specific domain of heart murmur findings. Two schemas were established from two different semantic terminology models for evaluation: one schema is extracted from the existing SNOMED CT model (S-Model) and the other is a template model (T-Model) extracted from a concept-dependent attributes model recently published by Green, et al [7]. The objectives of the study are to annotate the real clinical notes using the two schemas and to compare and evaluate the effects of two models on classification of the clinical notes.

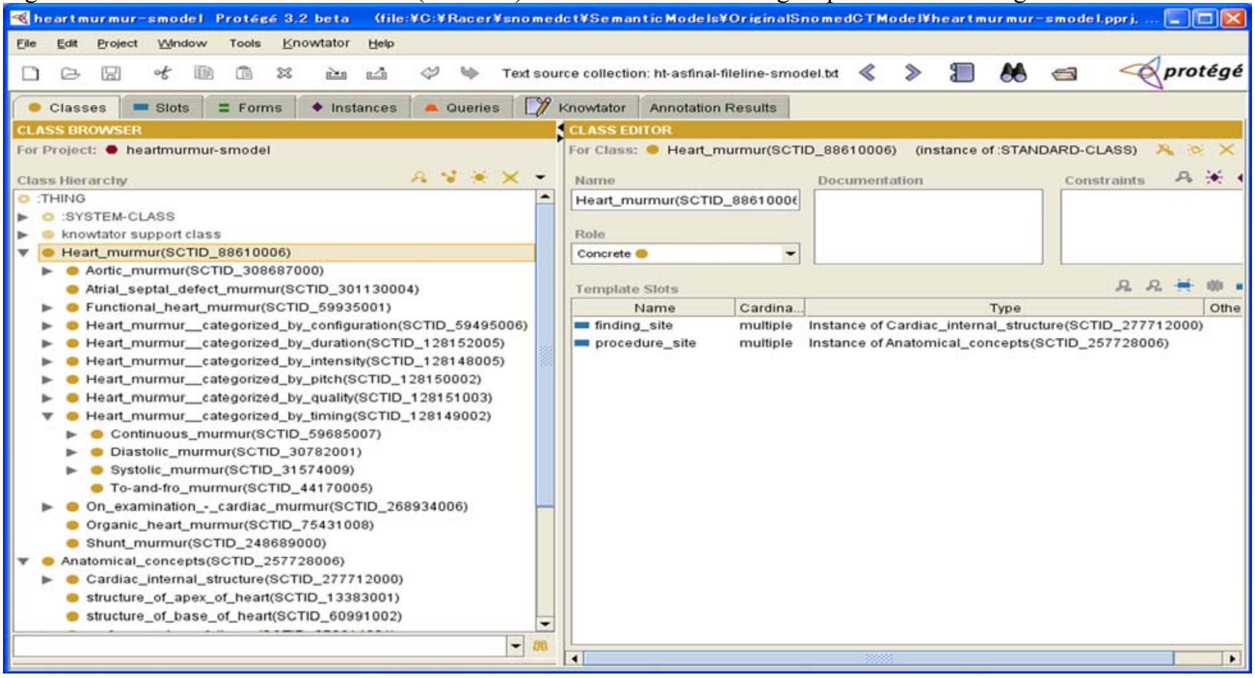
Methods and Materials

Defining the annotation schemas

We defined two schemas for both S-Model and T-Model and represented the two schemas in Protégé (version 3.2 beta), which is an ontology editing environment and was developed by Stanford Medical Informatics [8].

For the S-Model, we established a schema by extracting concept trees from the existing sub-hierarchy of heart murmur findings in January 2006 version of SNOMED CT (see Fig. 1). One root concept is "Heart murmur (SCTID_88610006)" which includes 86 sub-concepts of pre-coordinated terms of heart murmur findings. The other root concept is "Anatomical concepts (SCTID_257728006)" which includes two parts relevant to our schema. One part is the concept "Cardiac internal structure (SCTID_277712000)" and its sup-concepts. The other part contains only those anatomical concepts appearing in our clinical notes corpus on the basis of a manual review. For all heart murmur concepts, two semantic attributes derive from SNOMED CT context model for heart murmur findings that frame post-coordination. One is "procedure site" that represents the auscultation site of a heart murmur and the other is "finding site" that represents the potential etiological site of a heart murmur. The values of the former one were set as the instances of "anatomical concepts (SCTID_257728006)" and the values of the latter one were set as the instances of "Cardiac internal structure (SCTID_277712000)".

Fig. 1 Schema of SNOMED CT Model (S-Model) for heart murmur findings represented in Protégé



For Project: heartmurmur-smodel

- :THING
- :SYSTEM-CLASS
- knowtator support class
- Heart_murmur(SCTID_88610006)
  - Aortic_murmur(SCTID_308687000)
  - Atrial_septal_defect_murmur(SCTID_301130004)
  - Functional_heart_murmur(SCTID_59935001)
  - Heart_murmur__categorized_by_configuration(SCTID_59495006)
  - Heart_murmur__categorized_by_duration(SCTID_128152005)
  - Heart_murmur__categorized_by_intensity(SCTID_128148005)
  - Heart_murmur__categorized_by_pitch(SCTID_128150002)
  - Heart_murmur__categorized_by_quality(SCTID_128151003)
  - Heart_murmur__categorized_by_timing(SCTID_128149002)
    - Continuous_murmur(SCTID_59685007)
    - Diastolic_murmur(SCTID_30782001)
    - Systolic_murmur(SCTID_31574009)
    - To-and-fro_murmur(SCTID_44170005)
  - On_examination_-_cardiac_murmur(SCTID_268934006)
  - Organic_heart_murmur(SCTID_75431008)
  - Shunt_murmur(SCTID_248689000)
- Anatomical_concepts(SCTID_257728006)
  - Cardiac_internal_structure(SCTID_277712000)
  - structure_of_apex_of_heart(SCTID_13383001)
  - structure_of_base_of_heart(SCTID_60991002)

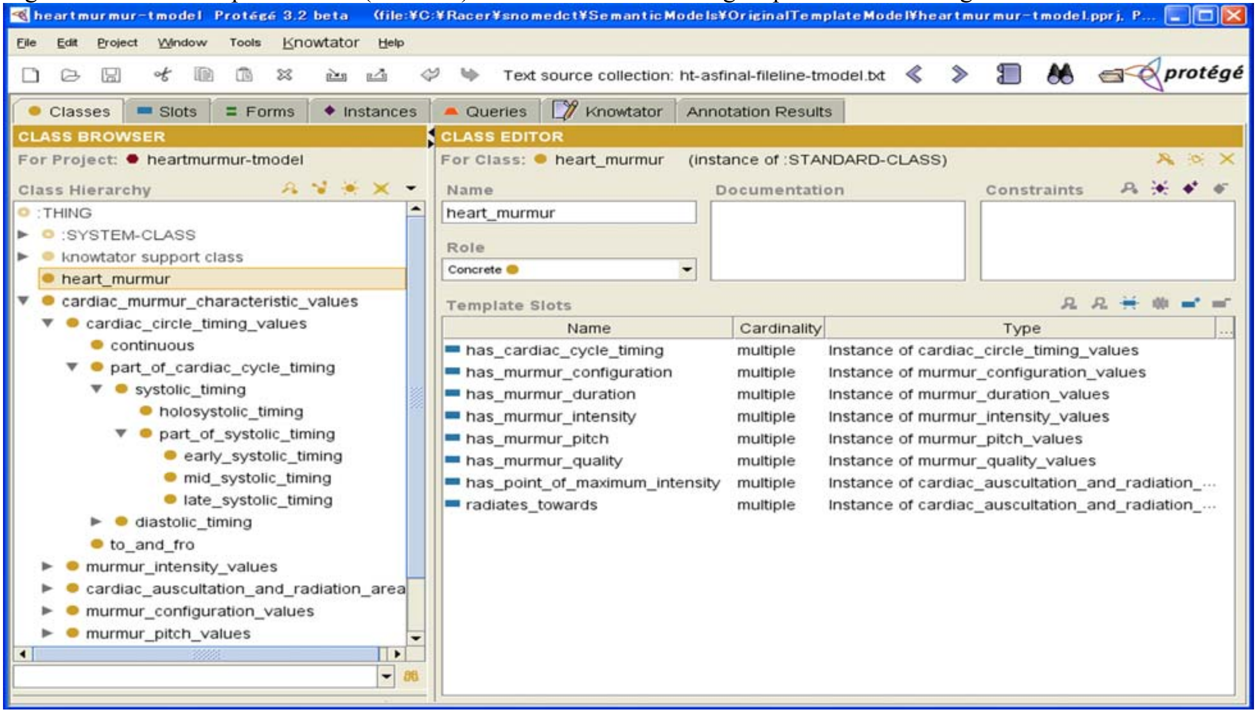
For Class: Heart_murmur(SCTID_88610006) (instance of :STANDARD-CLASS)

Name: Heart_murmur(SCTID_8861000€

Role: Concrete

Name	Cardina...	Type	Othe
finding_site	multiple	Instance of Cardiac_internal_structure(SCTID_277712000)	
procedure_site	multiple	Instance of Anatomical_concepts(SCTID_257728006)	

Fig. 2 Schema of Template Model (T-Model) for heart murmur findings represented in Protégé



For Project: heartmurmur-tmodel

- :THING
- :SYSTEM-CLASS
- knowtator support class
- heart_murmur
- cardiac_murmur_characteristic_values
  - cardiac_circle_timing_values
    - continuous
    - part_of_cardiac_cycle_timing
      - systolic_timing
        - holosystolic_timing
        - part_of_systolic_timing
          - early_systolic_timing
          - mid_systolic_timing
          - late_systolic_timing
      - diastolic_timing
    - to_and_fro
  - murmur_intensity_values
  - cardiac_auscultation_and_radiation_area
  - murmur_configuration_values
  - murmur_pitch_values

For Class: heart_murmur (instance of :STANDARD-CLASS)

Name: heart_murmur

Role: Concrete

Name	Cardinality	Type
has_cardiac_cycle_timing	multiple	Instance of cardiac_circle_timing_values
has_murmur_configuration	multiple	Instance of murmur_configuration_values
has_murmur_duration	multiple	Instance of murmur_duration_values
has_murmur_intensity	multiple	Instance of murmur_intensity_values
has_murmur_pitch	multiple	Instance of murmur_pitch_values
has_murmur_quality	multiple	Instance of murmur_quality_values
has_point_of_maximum_intensity	multiple	Instance of cardiac_auscultation_and_radiation_...
radiates_towards	multiple	Instance of cardiac_auscultation_and_radiation_...

For the T-Model, a schema was established from a concept-dependent attributes model published in a recent paper of Green, et al [7]. In this schema (see Fig. 2), one root concept is "heart murmur" which had eight semantic attributes, consisting of "has cardiac cycle timing", "has murmur configuration", "has murmur duration", "has murmur intensity", "has murmur pitch", "has murmur quality", "has point of maximum intensity", "radiates towards". The corresponding values of these eight attributes were set as the sub-concepts of the other root concept "cardiac murmur characteristic values". We adopted the model attributes are directly from Green's model, as well as their values (kindly provided by Green, interpersonal communication).

Preparing clinical notes corpus

The Mayo Clinic has a repository of approximately twenty million clinical notes that consist of documents dictated by physicians that are subsequently transcribed and filed as part of the patient's electronic medical record. The following criteria were made to sample those notes. Firstly, we extracted notes with these criteria from Mayo repository in an automatic way: 1) created between January 1, 2005 to January 31, 2005; 2) Having a heart murmur description in *Physical Examination* section; 3) age ≥ 21; 4) Having a Hospital International Classification of Disease Adaptation (HICDA) code of the Heart Valvular Disease, and 5) removing patients with a code for status prosthetic valve or complication of a prosthetic valve. Secondly, we flagged extracted documents containing a diagnosis of aortic stenosis (AS), yielding 103 documents. Thirdly, we randomly selected controls among the extracted documents having no diagnosis of AS by matching the following conditions: 1) no history of vavular surgeries; 2) matching gender and age within 1 year for each case (see Table 1). Two controls were retained for each case, totaling to 309 documents. Finally, we parsed out cardiac exam from the *Physical Examination* section of each document to create an annotation corpus.

Table 1. Control documents selection by matching with gender and age

Age	Male	Control	Female	Control	Total
21-30	1	2	0	0	3
31-40	0	0	0	0	0
41-50	0	0	2	4	6
51-60	4	8	0	0	12
61-70	7	14	5	10	36
71-80	26	52	7	14	99
81-90	24	48	21	42	135
91-	2	4	4	8	18
Total	64	128	39	78	309

Annotation software and Annotators

A general purpose text annotation tool, Knowtator [9], was used to map text contents to our schema. Knowtator is a Java plug-in for Protégé and mainly used for creating gold-standard training and evaluation corpora for natural language processing (NLP) systems. The annotation schemas described in section above were instantiated in Knowtator.

One author (GJ) performed the annotation task and then the other author (CGC) verified the annotations for 10% of all documents. Differences were mutually adjudicated and lessons generalized to the remaining 90% of cases.

Coding for machine learning classification

We coded the annotated corpora for classification using a machine learning classification algorithm. The target category of the classification is binary, i.e. aortic stenosis (AS) or non-AS. In other words, the goal of the classification is to predict whether a document with a heart murmur description belongs to AS category or not. The annotations of each document were used as the predictive features and coded as binary.

We used a Weka implementation of the decision tree (J4.8) [10], which is a well-known supervised approach to classification.

Outcome measures and statistical analysis

For the annotation task, we compared the description completeness between the two models. The annotators were asked to judge whether the heart murmur descriptions of each document could be described completely through using the schema of a model while they performed annotation task. If they judged a document as "incomplete", they indicated a reason for the judgment.

To evaluate the data retrieval task, we used the standard evaluation metrics of precision, recall, f-score and accuracy. Precision is defined as the ratio of correctly assigned AS category (true positive) to the total hit number (true positives and false positives). Recall is the ratio of correctly assigned AS category (true positive) to the number of target category in the test set (true positives and false negatives). The f-score represents the harmonic mean of precision and recall. Accuracy is the ratio of correctly assigned categories (true positives and true negatives) to total number of instances in test dataset.

For S-Model, one dataset (SM) that contains the annotations of both heart murmurs and anatomical concepts was prepared. For T-Model, three datasets were prepared. The first one (TM1) is that contains the annotations from Green's original model. The other two datasets are extension of TM1. We extended TM1 to create TM2 by completing the values for all eight semantic attributes whenever a description appearing in the clinical notes corpus did not have a corresponding value in TM1. For example, we added "upper sternal border", "mid sternal border" and "lower sternal border" into the schema because they appeared frequently in our corpus to describe the auscultation areas and the original model only contains "sternal border".

Building on TM2, we created our third model (TM3) by adding a new semantic attribute "has inferences to (specific murmurs or etiological mentions)" to the root concept "heart murmur" and also completing its corresponding values from those descriptions appearing in the corpus. We re-annotated all documents using the extended models respectively.

Ten-fold cross validation for retrieval was performed 10 separate times over all four datasets and the paired t-test was performed to test the statistical significance of performance measures between the dataset of S-Model and three datasets of T-Model.

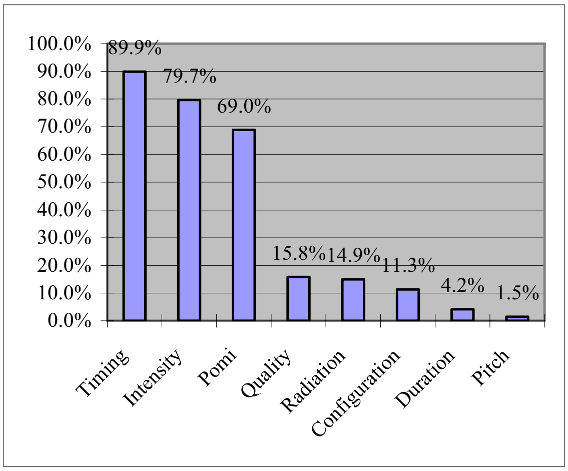
Results

For annotations

In S-Model, we made 995 annotations across all 309 documents. The average number of annotations per document is 3.2. Among the annotations, 728 belonged to 33 different sub-concepts of heart murmur (88610006). Of the heart murmur annotations, 509 (70.0%) had the values of the attribute "procedure site" filled and 6 (0.8%) had the values of the attribute "finding site" filled.

In T-Model, we made 1377 annotations against the original T-Model (TM1). The average number of annotations per documents is 4.5. Among 335 discrete heart murmur annotations, 89.9% include timing, 79.7% include intensity and 69.0% include points of maximum intensity (POMI). (see Fig.3)

Fig. 3 The annotation distribution of the eight attributes for all 335 heart murmurs annotated in original T-Model.



For comparison, the average number of annotations per document in S-Model was less than those in T-Model, indicating that S-Model supports more abstract way for description of heart murmur findings than T-Model.

Considering description completeness, 88 documents (28%) in S-Model were judged as "incomplete"; in the original T-Model, 201 documents (65%) were judged as "incomplete". Thus, S-Model exhibits more complete domain coverage than the original T-Model.

The reasons for the incompleteness of four datasets from two models were listed in Table 2. We found that S-Model (SM) could describe most of "auscultation area" and the original T-Model (TM1) could not. For "radiation", both SM and TM1 could not describe it well (we noticed that for SM, it is due to lacking of semantic attribute for "Radiation", whereas that in TM1 is due to lacking of appropriate values for "Radiation" attribute). In addition, SM could describe all "ejection murmur" mentions and part of "aortic valve related" etiological mentions; TM1 could not. The results indicated that the strict template model, per Green, assumes that observers are using strict descriptions, and not making inferences to specific murmurs and etiological mentions, whereas SNOMED CT model accommodates partly the variability in inferences and strict descriptions, by providing terms that covers both.

Table 2 Frequency of reasons for the incompleteness of four datasets from two models

	SM	TM1	TM2	TM3
Auscultation area	1	78	0	0
Radiation	47	47	0	0
Configuration	8	8	0	0
Quality	7	5	0	0
Specific murmurs				
Ejection murmur	0	107	107	0
Regurgitant murmur	3	3	3	0
Flow murmur	2	2	2	0
Etiological mentions				
Aortic valve related	19	25	25	0
Mitral valve related	4	4	4	0
Pulmonary valve related	1	1	1	0
Septal defect	1	1	1	0

For TM2 and TM3, zero values in Table 2 indicated our synthetic completion of the values of each corresponding attribute in T-Model. The description completeness of TM2 was corresponding up to 57.6%, and that of TM3 up to 100%. Table 3 provided the examples (a AS case vs. a Non-AS case) to show how annotations were taken for all four schemas from two models.

For classification

As described in above section, four datasets (SM, TM1, TM2 and TM3) from two models were formed for evaluation. The results of the evaluation metrics of the four datasets were shown in Table 4. We found that the classification performance of SM was better than TM1 (i.e. original Green's model), with statistical significance identified for recall and f-score ($p<0.05$, paired t-test). We consider that the reason was probably that the TM1 did not contain a complete list of murmur characteristic values for many of its semantic attributes.

The performance of TM2 was better than TM1, but still lesser than SM. The result indicates that the original T-Model using strict physical descriptions may not fully represent descriptions of heart murmur findings in clinical notes, negatively impacting functional performance.

The classification performance of TM3 was the significantly best among the datasets ($p<0.05$, paired t-test vs. SM). The result provided further evidence that inferences to specific murmurs and etiological mentions were important part of descriptions of heart murmur findings in real clinical notes, influencing the functional performance of the terminology model in this specific domain.

Table 3 The examples (AS Case vs. Non-AS Case) of annotations using four schemas

	AS Case	**Non-AS Case**
Textual Note	Heart: Loud 3 to 4/6 systolic ejection murmur heard best at the right upper sternal border. Absent of S2.	Heart: Regular rate and rhythmwith a 2/6 left upper sternal border systolic regurgitant murmur. P2 was slightly increased. There was an S4 but no S3. The apical impulse was not localizable.
SM Annotation	15157000:Cardiac murmur - intensity grade III (VI) procedure site: [117144008:upper parasternal region] laterality: [24028007:right] 25311008:Cardiac murmur - intensity grade IV (VI) procedure site: [117144008:upper parasternal region] laterality: [24028007:right] 77197001: Ejection murmur procedure site: [117144008:upper parasternal region] laterality: [24028007:right]	36680007:Cardiac murmur - intensity grade II (VI) procedure site: upper parasternal region laterality: [7771000:left] 31574009: Systolic murmur procedure site: [117144008:upper parasternal region] laterality: [7771000:left]
TM1 Annotation	Heart murmur: has cardiac cycle timing value: systolic timing has murmur intensity value: intensity grade III/VI has murmur intensity value: intensity grade IV/VI has point of maximum intensity: sternal border (laterality: right)	Heart murmur: has cardiac cycle timing value: systolic timing has murmur intensity value: intensity grade II/VI has point of maximum intensity: sternal border (laterality: left)
TM2 Annotation	Heart murmur: has cardiac cycle timing value: systolic timing has murmur intensity value: intensity grade III/VI has murmur intensity value: intensity grade IV/VI has point of maximum intensity: upper sternal border (laterality: right) has murmur quality value: loud	Heart murmur: has cardiac cycle timing value: systolic timing has murmur intensity value: intensity grade II/VI has point of maximum intensity: upper sternal border (laterality: left)
TM3 Annotation	Heart murmur: has cardiac cycle timing value: systolic timing has murmur intensity value: intensity grade III/VI has murmur intensity value: intensity grade IV/VI has point of maximum intensity: upper sternal border (laterality: right) has murmur quality value: loud has inferences to: ejection murmur	Heart murmur: has cardiac cycle timing value: systolic timing has murmur intensity value: intensity grade II/VI has point of maximum intensity: upper sternal border (laterality: left) has inferences to: regurgitant murmur

Table 4 The results of the evaluation metrics of the four datasets

	Precision (mean±sd)	Recall (mean±sd)	F-score (mean±sd)	Accuracy (mean±sd)
SM	74.2% ±13.7%	59.4% ±15.6%	64.5% ±12.7%	79.0% ±6.1%
TM1	67.5% ±14.9%	*44.6% ±13.8%	*52.1% ±11.5%	73.6% ±5.4%
TM2	71.0% ±14.0%	53.2% ±18.9%	59.0% ±15.3%	76.9% ±6.8%
TM3	80.0% ±12.2%	*69.8% ±14.6%	73.5% ±10.4%	*83.6% ±5.8%

*p< 0.05 (paired t-test)

Discussions

In this study, we developed an approach to compare and evaluate the domain coverage (indicated by the description completeness) of two semantic terminology models and their effects on the classification of real clinical notes. We found that the description completeness of the S-Model was better than the original T-Model with original value set, correspondingly the performance of the S-Model on classification was also better. The extensions of T-Model that improved the description completeness, did improve its performance on classification of clinical notes. We clearly demonstrated that the domain coverage of a terminology model was directly correlated with its performance on classification of clinical notes; this is not surprising.

We could see that the effect of a terminology model on its functional performance in a specific domain mainly depends on its ability to represent the contents of the domain. In other words, the key issue for a terminology model is how to achieve complete domain coverage. If two different terminology models could represent the contents of a domain to achieve the same coverage, their performances on classification of clinical notes should have no difference.

In original T-Model, the description of a hear murmur could be fully post-coordinated by a semantic structure of eight semantic attributes. With original value set, we found that its description completeness was sub-optimal. In the paper from which the model was derived [7], the authors stated that "to adequately capture the full spectrum of cardiac murmur descriptions, our model needed a complete list of murmur characteristics". So our first extension (TM2) completes the term values for all eight attributes of the original T-Model. The description of completeness was increased from 35.0% to 57.6%.

Thus, adding axes content to each attribute within the semantic structure did improve the domain coverage of the model; however, even with value completion, the original T-Model still could not achieve complete description for given corpus.

Therefore, we consider that the domain coverage of a terminology model depends not only on the full value set of its semantic structure, but also on the coverage of the semantic structure itself.

Our second extension (TM3) of the T-Model adds a semantic attribute together with its corresponding values. This did overcome the limitation of semantic structure of the original T-Model and achieves a complete description for given corpus. In other words, the extended structure allows a systematic examination of where content gaps exist (e.g. missing values of references to specific murmurs and etiological mentions) and also guides the "completion" of the terms or missing contents informed by the extended structure.

In S-Model, most of its contents are pre-coordinated, with the post-coordination only possible for two semantic attributes "procedure site" and "finding site". We did not extend the SNOMED CT model in a similar fashion since the model is an international standard although we believe that performance would be improved were it also extended. However, the extension of the model would be more complicated than that of template model because it involves both pre-coordination and post-coordination. We consider that the template model would be more applicable for achieving complete domain coverage. An important implication of these experiments is that a templatestyle terminology model more readily identifies gaps in coverage, and facilitates their completion for classification tasks.

Knowtator was used as our annotation tool and satisfied our purpose well, demonstrating the following merits. The first merit is that Knowtator uses the Protégé ontology editing environment to build the annotation schema. The frame-based knowledge representation system provides a flexible and expressive way to efficiently make schemas of the two model types in this study. The second merit is that Knowtator provides visualization of annotations, making the annotation task and confirmation process simple and efficient. The third merit is that the Java API of the system, which supports the annotation query that exports our coding of annotations to a classifier format automatically.

In order to improve the baseline performances on all standard evaluation measures, we performed control selection of clinical notes using strict criteria. This design did improve baseline performances (data not shown).

We regard the evaluation in this study in its comparative context across models; absolute measures of precision and recall are subject to factors beyond the scope of this study. A limitation of this study is that the annotations of clinical notes depends entirely on what clinicians decide to document for each patient, who they may or may not know has AS at the time. The local culture around documentation seems possible that these findings could be different on another corpus. Second, we only collected a relatively small size of clinical notes corpus given that the intensive annotation tasks were required. We consider that the annotation corpus is valid as both authors have clinical medicine background. Ten-fold cross validation used in this study may facilitate the efficient use of the data and get the best liability estimate. This kind of annotation corpus may be used to train a machine learning based annotation algorithm to build an automatic domain specific annotation tool. In addition, because it was not our intention to evaluate which classifier performed better, we only used a Weka implementation of the decision tree (J4.8) algorithm.

In conclusion, the domain coverage of the two models and their performance on classification clearly differ when applied to real clinical notes. Our approach provides an effective framework to evaluate the coverage and functional performance of the semantic terminology models in a specific domain for potential improvement. Future direction would focus on the scalability of the approach and the evaluation of interoperability among the different semantic terminology models.

Acknowledgements

This study is partly supported by NIH R01 LM07319. The authors would like to thank Philip V. Ogren, Serguei V.S. Pakhomov, Guergana K. Savova, Pauline Funk and James D. Buntrock for their support.

References

[1] Brown PJ, Sonksen P. Evaluation of the quality of information retrieval of clinical findings from a computerized patient database using a semantic terminological model. J Am Med Inform Assoc. 2000 Jul-Aug;7(4):392-403.

[2] Spackman KA, Campbell KE. Compositional concept representation using SNOMED: towards further convergence of clinical terminologies. Proc AMIA Symp. 1998;:740-4.

[3] Yu AC. Methods in biomedical ontology.J Biomed Inform. 2006 Jun;39(3):252-66.

[4] Zhou L, Tao Y, Cimino JJ, Chen ES, Liu H, Lussier YA, Hripcsak G, Friedman C. Terminology model discovery using natural language processing and visualization techniques. J Biomed Inform. 2006 Dec;39(6):626-36.

[5] URL: http://icn.ch/icnp.htm; last visited at December 29, 2006.

[6] Rosenbloom ST, Miller RA, Johnson KB, Elkin PL, Brown SH. Interface terminologies: facilitating direct entry of clinical data into electronic health record systems. J Am Med Inform Assoc. 2006 May-Jun;13(3):277-88.

[7] Green JM, Wilcke JR, Abbott J, Rees LP. Development and evaluation of methods for structured recording of heart murmur findings using SNOMED-CT post-coordination. J Am Med Inform Assoc. 2006 May-Jun;13(3):321-33. Epub 2006 Feb 24.

[8] URL: http://protege.stanford.edu/index.html; last visited at December 29, 2006.

[9] URL: http://bionlp.sourceforge.net/Knowtator/; last visited at December 29, 2006.

[10] URL: http://www.cs.waikato.ac.nz/ml/weka/; last visited at December 29, 2006.

Creation and Usage of a "Micro Theory" for Long Bone Fractures: An Experience Report

Howard S. Goldberg, MD[1], Vipul Kashyap, PhD[1], Kent A. Spackman, MD, PhD[2]
[1]Clinical Informatics R&D, Partners Healthcare System, Wellesley, MA, USA
[2]Oregon Health & Science University, Portland, OR, USA
{hgoldberg, vkashyap1}@partners.org, ksp@ihtsdo.org

We seek to leverage enhanced expressivity in OWL 1.1 via property chain axioms with right identities in order to organize and constrain anatomic concepts for use in clinical descriptions. Anatomic knowledge represented in SNOMED CT uses SEP triplets; we anticipate that property chains will allow a more parsimonious organization of anatomic concepts. However, these constructs may lead to unanticipated inference, especially when scaling to large numbers of concepts [1]. We used a bottom-up approach based on targeted use case questions to iteratively develop a "micro theory" that both identifies the sensible locations of fractures in long bones and also supports logic-based classification of fractures. Alternative representations of the statement "fractures occur in bone" were explored with the aim of creating rich clinical descriptors that support classification for inference and data mining. The process of creating this micro theory is discussed, where pragmatic decisions were made with an intention of both constraining data entry and enabling inferences within the scope of the use cases.

INTRODUCTION

OWL and other forms of description logics have been used extensively to model spatial relationships for anatomical knowledge [1-6]. The focus of these efforts has been either to investigate the computational properties of the description logic or to develop a generalized set of axioms or theories to support classification inferences for a wide variety of clinical decision support use cases. We seek to leverage the enhanced expressivity of OWL 1.1 [7] to organize anatomic concepts for use in creating clinical descriptions. In particular, we explore the use of property chain axioms with right identities to simplify a knowledge base of anatomy without limiting the inferences that can be computed. It has previously been demonstrated that for anatomical descriptions, inferences after addition of these axioms can remain computationally tractable [3]. In contrast with other approaches, such as SEP triplets [2], we anticipate that property chains will allow a more parsimonious organization of anatomic concepts. The downside to using property chains and transitivity, however, are that these constructs may lead to unanticipated inference, especially when scaling to large numbers of concepts [1].

For our initial investigation, we focused on a single use case limited to fractures of long bones. We adopted an iterative bottom-up process to developing a "micro-theory"—an axiomitization that yields sensible and logically correct inference in a limited domain. At each stage, we tested the incremental theory against the use case scenario. We re-used content from the Foundational Model of Anatomy (FMA) [8]. The various distinctions introduced in the FMA to model partonomy, i.e., systemic-part-of, regional-part-of and constitutional-part-of were explored. We attempted to design a theory that was compact and understandable and also gave us the correct intended behavior. The model accounts for both anatomic perspectives and functional clinical perspectives. We tested the model by computing the appropriate inferences based on the use cases.

Locative transfer over pathophysiologic processes is a fundamental property for ontologies that will be used for clinical decision support or data warehousing applications. Given the sheer number of anatomic concepts present in systems such as SNOMED CT, it is critical that modeling idioms yield predictable results in order to scale. An important goal of the current work is begin to understand the characteristics of these idioms in a limited domain.

Clinical Scenario and Use Case Questions

Typically, a physician creates a clinical descriptor that is of sufficient granularity to support a management plan—the clinical descriptor is an index for the general management plan for a given pathology. Within the contemporary electronic health record, the clinical descriptor may be reused as data to drive point-of-care decision support, or as warehouse data to support reporting. For instance, if we need to report the number of patients who had a fracture of the proximal femur, we should include the number of patients who had a fracture of the femoral neck. In both cases, the original descriptor should support detailed classification schemes.

With respect to bone fractures, it is desirable to describe fractures in detail with respect to the bone features involved—the clinical detail drives the management plan. The clinical detail may describe either a fracture involving an anatomic landmark or a functional region where all fractures act similarly. It is equally important that the clinical descriptor not admit any nonsensical description. While fractures may involve bony landmarks, we generally do not describe fractures of the periosteum—the bone lining—or the bone marrow. While these are parts of bones, they are not generally parts through which fractures are described to occur. The GALEN project used constraints called *sanctions* to specify the values that could sensibly be applied to relations such as *has-location* [6]. Similarly, we constructed our ontology fragment with the intent of logically defining the set of all and only locations for fractures.

Given a need to document, to classify, and possibly to obtain reference information, useful questions that might be posed include[†]

1. What bone regions and features are contained in the Distal Epiphysis of the Femur?
2. What parts of the Distal Epiphysis of the Humerus are covered by Articular Cartilage?
3. Is a fracture of the Femoral Neck also a fracture of the Proximal Femur (i.e., is a fracture through an anatomic feature a fracture of a functional region)?
4. Is a fracture of the Trochlea a fracture of the Distal Epiphysis of the Humerus?
5. Is a fracture of the Trochlea an intra-articular fracture?
6. Is a fracture of the Trochlea an intra-articular fracture of the Distal Epiphysis of the Humerus?

MATERIALS

We looked at the following two sources for creating the fracture ontology: (a) The SNOMED CT hierarchies spanning the femur and the humerus; and (b) The FMA hierarchies corresponding to the femur and the humerus. A brief description of the portion of these two knowledge sources is described below.

SNOMED CT

SEP-triplets are extensively employed in the anatomical part of SNOMED CT. For each SNOMED anatomical class representing one entire entity, called *entity (or entire)* class (E-class), there are two auxiliary classes, the *structure* class (S-class) and the *part* class (P-class). For example, in the femur hierarchy, we ideally would have the following classes defined:

$\mathsf{StructureOfFemur}$
$\mathsf{EntireFemur} \sqsubseteq \mathsf{StructureOfFemur}$
$\mathsf{FemurPart} \sqsubseteq \mathsf{StructureOfFemur} \sqcap \exists\mathsf{partOf.EntireFemur}$
$\mathsf{BoneStructureOfDistalFemur} \sqsubseteq \mathsf{FemurPart}$
$\mathsf{EntireDistalFemur} \sqsubseteq \mathsf{BoneStructureOfDistalFemur}$
$\mathsf{DistalFemurPart} \sqsubseteq \mathsf{BoneStructureOfDistalFemur}$
$\qquad \sqcap \exists\mathsf{partOf.EntireDistalFemur}$
$\mathsf{StructureOfDistalEpiphysisOfFemur} \sqsubseteq \mathsf{DistalFemurPart}$
$\mathsf{EntireDistalEpiphysisOfFemur}$
$\qquad \sqsubseteq \mathsf{StructureOfDistalEpiphysisOfFemur}$

The E-class is instantiated by entire anatomical objects (such as the entire femur), and the P-class by the proper parts of the referred objects (such as the distal femur). The S-class, finally, is instantiated by instances that are either entire objects or their parts. This definition explains the *is-a* links from the E-class and the P-class to the S-class, as well as the partOf link from the P-class to the E-class. The main idea underlying the SEP-triplet approach is to represent a part-whole relationship between two entity classes not by a part-of link between the E-classes, but rather by an is-a link between the S-class of the "part" and the P-class of the "whole". This is, however, sufficient to simulate transitivity of part-of through the inherently transitive relation is-a:

$\mathsf{EntireDistalEpiphysisOfFemur}$
$\sqsubseteq \mathsf{StructureOfDistalEpiphysisOfFemur}$
$\sqsubseteq \mathsf{DistalFemurPart}$
$\sqsubseteq \mathsf{BoneStructureOfDistalFemur}$
$\sqsubseteq \mathsf{FemurPart}$
$\sqsubseteq \exists\mathsf{partOf.EntireFemur}$

This allows us to conclude that every Distal Epiphysis of the Femur is part of some Femur. Since characteristics are inherited along the is-a hierarchy, the SEP-triplet encoding also allows us to simulate inheritance of characteristics along the part-of hierarchy. In our example, by connecting a fracture via the findingSite property to the S-class, we can ensure that a fracture located in the Distal Epiphysis of the Femur is classified as a fracture located in the Femur. Another advantage of the SEP encoding is that one can suppress such inheritance along the part-of hierarchy by connecting via findingSite to the E-class.

There are, however, several problems with the SEP-triplet encoding. First, from a formal ontological point of view, it partially conflates the is-a hierarchy with the part-of hierarchy, which may lead to unintended consequences since the two relationships are completely different by nature [9]. In SNOMED, it has indeed turned out that is-a links can be ambiguous, i.e., it is not always clear whether they are

[†] 1) The medial and lateral condyles of the Femur; 2) Trochlea and Capitellum; 3) Yes; 4) Yes; 5) Yes; 6) Yes.

introduced as part of the SEP-triplet approach, or are supposed to represent a genuine generalization relationship. Second, the SEP-triplet approach is error prone since it works correctly only if it is employed with a very strict modeling discipline. In SNOMED, triplets are often modeled in an incomplete way; in particular, the P-class and the part-of link to it from the E-class are missing in most cases. For example, the following axioms presented earlier were not actually asserted in SNOMED, but were included for pedagogical purposes (DistalFemurPart does not currently exist in SNOMED):

$$\begin{aligned}\mathrm{DistalFemurPart} \sqsubseteq{} & \mathrm{BoneStructureOfDistalFemur}\\ & \sqcap\ \exists \mathrm{partOf.EntireDistalFemur}\\ \mathrm{StructureOfDistalEpiphysisOfFemur} \sqsubseteq{} & \mathrm{DistalFemurPart}\end{aligned}$$

In addition, the auxiliary S-class is sometimes incorrectly used as if it were an *entire* entity class. Third, the approach introduces for every proper class in the ontology two auxiliary classes, which results in a significant increase in the ontology size. Finally, the SEP approach makes it much more difficult to define and maintain the set of sensible locations for fractures.

Foundational Model of Anatomy

The FMA ontology defines a set of partonomic relationships discussed in [10,11] for guiding the representation of anatomical parts. This is a smaller set than that used in GALEN [6], and thus one of the questions we seek to answer is whether it is sufficient for clinical modeling, Refinements of the generic part-whole relationships for anatomical structures are proposed, as anatomical structures have been decomposed based on several different contexts. A partition is defined as the decomposition of the entire body or any anatomical structure in a given context or viewpoint.

A *constitutional part* is defined as a primary partition of an anatomical structure into its compositionally distinct anatomical elements. In the context of the whole, an element is any relatively simple component of which a larger, more complex anatomical structure is compounded; i.e., the partition is compositional rather than spatial. For example, a stomach may be viewed as being partitioned into its wall and cavity. A *regional part* on the other hand is defined as a primary partition that spatially subdivides an anatomical structure into sets of diverse constitutional parts that share a given location within the whole; i.e., the partition is spatial rather than compositional. For example, a stomach may be viewed as being partitioned into its fundus, body and pyloric antrum to name a few of such parts. Constitutional parts are genetically determined, whereas regional parts are defined not only by genetically regulated developmental processes (e.g., lobe of lung, cortex of kidney, finger), but also by arbitrary landmarks or coordinates, such as used for demarcating the thoracic and abdominal parts of the aorta and the fundus of the stomach from adjacent parts of the corresponding wholes. A *systemic part* is defined as a secondary partition of an anatomic structure in accord with functional systems.

The distinction between regional parts determined by well defined genetically regulated processes and arbitrary landmarks and coordinates, is represented by associating the attributes *anatomical* or *arbitrary* with regional parts. Furthermore, these attributes provide the basis for the different views of regional partitions, as in the case of the liver, where its traditional partition into lobes based on *arbitrary* landmarks constitutes an arbitrary kind of regional view, while another partition based on the distribution of the tributaries of the hepatic veins or branches of the hepatic artery constitutes an *anatomical* regional view.

The FMA also supports topologic relationships supporting connectedness and containment. Connectedness describes whether structures are continuous with, attached to, or synapsed with other structures. Containment deals exclusively with the containment of a material anatomic entity within an anatomic space, e.g., Right lung -*contained in*- Right half of thoracic cavity. Connectedness and containment are orthogonal to regionality and constitutionality and do not confer parthood [12].

METHODS

We now present our approach to developing the long bone fracture ontology. We draw on the FMA as a primary source of anatomic content.

Regional vs. Constitutional Partitions

As previously discussed, the FMA ontology draws a distinction between a regional partition and a constitutional partition. We reviewed this content to determine whether it was suitable for reuse within our ontology fragment.

1. The ***regional partition*** of long bones is exemplified by the following regional parts of the Femur (regPartOf):

   $$\begin{aligned}&\mathrm{ProximalEpiphysisOfFemur} \sqsubseteq \exists\ \mathrm{regPartOf.Femur}\\ &\mathrm{DiaphysisOfFemur} \sqsubseteq \exists\ \mathrm{regPartOf.Femur}\\ &\mathrm{DistalEpiphysisOfFemur} \sqsubseteq \exists\ \mathrm{regPartOf.Femur}\\ &\mathrm{FemoralNeckOfFemur} \sqsubseteq \exists\\ &\mathrm{regPartOf.ProximalEndOfFemur}\end{aligned}$$

Regional parts of the femur include true anatomic parts (epiphyses, diaphysis) as well as functional parts defined by fiat boundaries (proximal end of femur), illustrating the FMA's anatomic and arbitrary types.

2. The ***constitutional partition*** of long bones is exemplified by the following constitutional parts of the Femur (constPartOf):

   BonyPartOfFemur ⊑ ∃constPartOf.Femur
   BoneOfFemur ⊑ ∃constPartOf.BonyPartOfFemur
   PeriosterumOfFemur ⊑ ∃constPartOf.BonyPartOfFemur
   MedullaryCavityOfFemur ⊑ ∃constPartOf.BonyPartOfFemur
   VasculatureOfBonyPartOfFemur ⊑ ∃constPartOf.BonyPartOfFemur
   ArticularCartilageOfDistalEpiphyisOfFemur ⊑ ∃constPartOf.Femur
   ArticularCartilageProximalEpiphysisOfFemur ⊑ ∃constPartOf.Femur
   VasculatureOfFemur ⊑ ∃constPartOf.Femur
   CavityOfFemur ⊑ ∃constPartOf.Femur

The constitutional parts of the femur include the multiple tissue types that combine to form a long bone—the bone proper, the articular cartilage, etc. Note the bone proper also decomposes to include the bone material itself, the periosteum, and the medullary cavity.

The regional partition includes the structures where clinicians locate fractures and the relationships between these structures. The constituents of long bone such as the periosteum, where fractures are not described to occur, are conveniently sequestered in the constitutional partition. We adopted the relevant portions of the regional partition for use in our model. However, we adopted a simpler representation for the incorporation of articular cartilage into the model for this initial iteration.

Modeling Design Choices
We now present some high-level classes and object properties that characterize the entities in which we are interested.

Bone
LongBone ⊑ Bone
Femur ⊑ LongBone
Humerus ⊑ LongBone
ObjectProperty(regionalPartOf)
reflexive(regionalPartOf)
transitive(regionalPartOf)
BoneRegion ⊑ ∃regionalPartOf.Bone
ObjectProperty(findingSite)
domain(findingSite) = Disorder
Disorder
Fracture ⊑ Disorder ⊓ ∃findingSite.BoneRegion

The class Bone is effectively the class BoneOrgan in the FMA. Within this initial iteration, we are neutral regarding the alignment of Bone with the Upper Ontology of FMA, i.e., aligning with the is-a hierarchy consisting of CavitatedOrgan, Organ, AnatomicalStructure, MaterialAnatomicalEntity, and AnatomicalEntity, as we did not see an impact of this in the context of the application at hand. The property findingSite aligns with the SNOMED CT relationship which assigns locations to clinical conditions

We declare the property regionalPartOf to be reflexive, thereby inducing Bone to be a BoneRegion. This has the important effect of unifying the treatment of entire long bones and bony landmarks with respect to findingSite—fractures may be declared to occur equally within the entire bone or at the landmark. We declare regionalPartOf to be transitive to support the interrelationships between discrete landmarks, larger regions of bone, and the entire bone.

Anatomical vs. Functional Partition
We add the following subclasses of BoneRegion into the model:

AnatomicBoneRegion ⊑ BoneRegion
FunctionalBoneRegion ⊑ BoneRegion

In clinical practice, pathology may be attributed to a true anatomic entity or a functional entity where unique pathologies behave similarly, are responsive to similar treatments, are aggregated for epidemiologic purposes, etc. In orthopedics, for example, several unique fractures all aggregate to fractures of the proximal femur. As previously noted, the FMA incorporates true anatomic regions and functional regions. We partition bone regions into either anatomic or functional components to support the independent enumeration of these features, as described in the use case.

Propagation of Locative Relationships
A key functionality that is required to support the use case questions discussed earlier is the ability to propagate the location of a fracture from a given region to all the regions to which it has regionalPartOf relationships. For instance, if a fracture is located in the femoral neck, it is also located in the proximal metaphysis of the femur as the femoral neck is a regional part of the proximal metaphysis of the femur. This is represented using the following axiom:

findingSite ∘ regionalPartOf ⊑ findingSite

It may be noted that the transfer of locative relationships is also propagated transitively due to the transitive nature of regionalPartOf.

$\text{regionalPartOf} \circ \text{regionalPartOf} \sqsubseteq \text{regionalPartOf}$

Articular Bone Regions

In order to explore articular fractures—the fracture of a bone region covered by articular cartilage, we incorporated the following concepts :

$\text{ArticularCartilage}$
$\text{ObjectProperty(coveredBy)}$
$\text{ArticularBoneRegion} \equiv \text{BoneRegion} \sqcap$
$\quad \exists\text{coveredBy.ArticularCartilage}$
$\text{ArticularFracture} \equiv \text{Fracture} \sqcap$
$\quad \exists\text{findingSite.ArticularBoneRegion}$

This representation provides a simple method to distinguish between articular and non-articular bone regions.

RESULTS

Using the initial ontology, we were able to create a series of detailed clinical descriptions which classified as expected. Some examples are discussed next.

Locative Transfer over Regional Parts

rA facture of the Femoral Neck is classified as a fracture of the Proximal Femur.

$\text{FemoralNeckFx} \equiv \text{Fracture} \sqcap \exists\text{findingSite.FemoralNeck}$
$\sqsubseteq \text{Fracture} \sqcap$

$\exists\text{findingSite.}(\exists\text{regionalPartOf.ProximalEndOfFemur})$
(Since $\text{FemoralNeck} \sqsubseteq$
$\quad \exists\text{regionalPartOf.ProximalEndOfFemur}$)
$\sqsubseteq \text{Fracture} \sqcap \exists\text{findingSite.ProximalEndOfFemur}$
(Since $\text{findingSite} \circ \text{regionalPartOf} \sqsubseteq \text{findingSite}$)
$\sqsubseteq \text{ProximalFemurFx}$

Transitive Locative Transfer

A fracture of the Femoral Neck is classified as a fracture of the Femur. Let's revisit the earlier example and begin with the following reformulation of FemoralNeck.

FemoralNeckFx
$\sqsubseteq \text{Fracture} \sqcap$
$\quad \exists\text{findingSite.}(\exists\text{regionalPartOf.ProximalEndOfFemur})$
$\sqsubseteq \text{Fracture} \sqcap \exists\text{findingSite.}$
$\quad (\exists\text{regionalPartOf.}(\exists\text{regionalPartOf.Femur}))$
(Since $\text{ProximalEndOfFemur} \sqsubseteq \exists\text{regionalPartOf.Femur}$)
$\sqsubseteq \text{Fracture} \sqcap \exists\text{findingSite.}(\exists\text{regionalPartOf.Femur})$
(Since $\text{regionalPartOf} \circ \text{regionalPartOf} \sqsubseteq$
regionalPartOf)
$\sqsubseteq \text{Fracture} \sqcap \exists\text{findingSite.Femur}$
(Since $\text{findingSite} \circ \text{regionalPartOf} \sqsubseteq \text{findingSite}$)
$\sqsubseteq \text{FemoralFx}$

The proof above indicates that we can represent direct relationships between bones and bone features and infer regional partonomy relationships between them.

Articular Fractures

Extending the model to describe and classify articular fractures is also accommodated by the model and creates no additional complications. The articular parts of the distal epiphysis of the humerus—trochlea and capitellum—are created as articular regions, while the non-articular parts—the medial and lateral epicondyle—are created as regular bone regions. The only caveat to this approach is that partially-covered regions are not considered articular regions; the distal epiphysis of the humerus is not considered an articular bone region by this criterion.

Fractures of the parts are created in the usual fashion by restricting the fracture finding site. Trochlear and capitellar fractures classify appropriately as articular fractures. General fractures of the distal humeral epiphysis and articular fractures are then created in the same way. Articular fractures of the distal humeral epiphysis are classified as subclasses of general fractures of the distal humeral epiphysis; trochlear and capitellar fractures are classified as further subclasses. Fractures of the epicondyles classify correctly as general fractures only. We did not specifically try to define non-articular fractures (fractures of parts not covered by articular cartilage).

Breach of the Model

We note one failure of the model in the subset of bones we examined. The FMA contains a regional part of the humerus, 'Nutrient Foramen of Humerus', literally, the hole where the nutrient artery enters the humerus. Because fractures are usually not described through this feature, this constitutes a failure of the model to constrain the set of sensible locations of fractures.

In the FMA, the nutrient foramen is a subclass of Immaterial Anatomic Entity. We can certainly remediate the model to additionally restrict bone regions to subclasses of Material Anatomic Entity. However, the appearance of the nutrient foramen as an arbitrary bone region bears further discussion.

DISCUSSION

We have begun to explore the creation of a 'micro-theory' for long bone fractures—an axiomitization that yields both sensible and logically correct inference. Using a set of framing axioms in combination with content from the FMA regional partition, we were able to describe and correctly classify a rich set of fractures, while maintaining a fairly parsimonious ontology. The resulting ontology is quite constrained as compared to an SEP triple approach.

Although this initial model successfully fulfills the use case, the breach raises significant questions. Because the FMA regional model admits arbitrary regions, there is no principled reason why an arbitrary region of a bone can sensibly be a fracture location. Our success seems to be an empirical finding—further analysis is necessary to see whether the model hold across all bones, or can extend to parenchymal organs such as the lung or the liver—organs made of the same 'stuff'.

Our model may succeed because bones exemplify a 'stuff/whole' partonomy, where arbitrary regions are compositionally homogenous. This does suggest that if regional parthood could be compositionally restricted, this would offer a more convincing model. One obvious possibility for implementing this restriction is to utilize the GALEN partitive attribute "hasSolidDivision", or perhaps a similar attribute with even more specialized meaning [13]. Currently, compositional properties are available through the constitutional partition of the FMA, i.e., if a part is composed of bone or a particular organ parenchyma, etc. Further work is necessary to reflect compositionality from constitutionality. We note that currently, wholes and parts have distinct roots in the FMA; in our model, it is important to treat parts and wholes similarly as bone regions.

This work provides initial insight into creating safe and effective inference over property chains for the purpose of creating and classifying clinical descriptions. Because of the tremendous change management implications of incorporating new idioms into terminologies such as SNOMED CT, it is important that we demonstrate that such idioms are safe, effective, and scale. Continuing work will investigate constraining the regional idiom with respect to homogenous compositionality, expanding our analysis to a larger set portion of the skeletal system, and examining the generalizability of the idiom to additional organ systems.

References

1. Seidenberg J and Rector A, Representing Transitive Propagation in OWL. Proc. 25th International Conference on Conceptual Modeling (ER 2006), November 2006.
2. Schulz S, Romacker M, and Hahn U, Part-whole reasoning in medical ontologies revisited: Introducing SEP triplets into classification-based description logics. Proc. 1998 AMIA Annual Fall Symposium.
3. Suntisrivaraporn B, Baader F, Schulz S and Spackmn K, Replacing SEP-Triplets in SNOMED-CT using Tractable Description Logic Operators. Proc. 11th International Conference on Artificial Intelligence in Medicine (AIME 07), July 2007.
4. Horrocks I and Sattler U, Decidability of SHIQ with complex role inclusion axioms. Artificial Intelligence 160(1-2), December 2004.
5. Schulz S, Hahn U and Romacker M. Modeling Anatomic Spatial Relations with Description Logics. Proc. 2000 AMIA Annual Fall Symposium.
6. Rector A, Bechhofer S, Goble C, Horrocks I, Nowlan W, Solomon W. The GRAIL concept modelling language for medical terminology. Artif Intell Med. 1997 Feb;9(2):139-71.
7. OWL Working Group, http://www.w3.org/2007/OWL/wiki/OWL_Working_Group
8. Foundational Model Explorer. http://fme.biostr.washington.edu:8089/FME/index.html
9. Patrick J. Aggregation and generalization in SNOMED CT. Proc. First Semantic Mining Conf. on SNOMED CT, 2006.
10. Mejino J L V, and Rosse C. Symbolic Modeling of structural relationships in the Foundational Model of Anatomy. Proc. First International Workshop on Formal Biomedical Knowledge Representation (KR-MED 2004), 2004.
11. Mejino J LV, Agoncillo A V, Rickard K L, and Rosse C, Representing complexity in part whole relationships within the Foundational Model of Anatomy. Proc. 2003 AMIA Annual Fall Symposium.
12. Smith B, Mejino JLV, Schulz S, et al. Anatomical Information Science. *Proceedings, Seventh International Conference on Spatial Information Theory COSIT 2005* Lecture Notes in Computer Science 3693, pages pp. 149-164.
13. Rector AL, Gangemi A, et al. The GALEN CORE model schemata for anatomy. Proceedings of Medical Informatics Europe, MIE 94, Lisbon, pp 186-189.

Representing clinical information using SNOMED Clinical Terms with different structural information models

David Markwell, MB, BS[a], Laura Sato MSc[b], Edward Cheetham MB ChB[c]
a) The Clinical Information Consultancy Ltd, b) NHS Connecting For Health, c) International Health Terminology Standards Development Organisation
david@clininfo.co.uk

Findings related to developing implementation specifications for the use of SNOMED Clinical Terms (SNOMED CT) in both HL7 and openEHR information models are summarized and compared. Common themes from this work, including overlaps between the expressivity of structure and terminology, are identified and discussed. Distinctions are made between aspects of meaning that are most readily represented by distinct structures, others where terminology offers greater flexibility and a 'gray-area' in which the relative merits are more balanced. Focusing on particular stages in the clinical information life cycle may suggest different points of balance and may lead to different approaches to integration. However, greater consistency is essential if clinical information is to be used effectively in electronic record systems. Consensus guidance documents of the type developed by the work described are only a first step. Mutually aware evolutionary refinement of structural and terminology standards is suggested as an enhancement to independent development.

INTRODUCTION

The last few years have seen the emergence of SNOMED Clinical Terms®[1] as the leading candidate for a controlled clinical terminology suitable for use in electronic health records[2]. In the same period, two structural information models have been advanced as standards for representing clinical information. The HL7 Reference Information Model has been used as the basis for a standard model of Clinical Statements[3] which is used in message specifications and in the HL7 Clinical Document Architecture, Release 2 (CDA)[4]. Meanwhile, the European standard for Electronic Health Records (EN13606) has been utilized by the *open*EHR Foundation[5] as a basis for developing a range of highly-constrained clinical statement and record composition models (called archetypes and templates).

These developments have been followed closely as part of the development of national specifications for capture and appropriate sharing of clinical information in the National Health Service (NHS) in England. NHS Connecting for Health (NHS CFH) chose SNOMED CT as the common clinical terminology to be used by all computers in the NHS in England. It also chose to utilize other relevant standards, including HL7 Version 3 for communications. More recently, *open*EHR-based archetypes and templates have been used to assist the specification of clinical data capture in regional NHS health record application projects. This required us to consider how SNOMED CT should be used with a combination of *open*EHR data models and existing HL7 Version 3 based models, as part of coherent 'end to end' system design specifications.

This paper draws together some of the findings of this work. It suggests general principles that may have wider applicability to when integrating terminologies with standard structural information models.

CONTEXT

SNOMED CT and HL7 Version 3

In 2004 it became apparent that there was widespread interest in the use of SNOMED CT in the HL7 community. The majority of the interest focused on how to integrate SNOMED CT with the emerging HL7 Version 3 standard. Following an initial meeting hosted by NASA, the HL7 Vocabulary Technical Committee launched the TermInfo Project to address this. The project was also supported by SNOMED International through an Associate Charter Agreement with the HL7 Board. The project has discussed a wide range of issues and prepared detailed guidance. After several ballot cycles, involving formal review and evaluation, the Guide to Use of SNOMED CT in HL7 Version 3[6] was accepted as a 'Draft Standard for Trial Use' in September 2007.

SNOMED CT, EN13606 and *open*EHR

During 2007, activities in the UK placed greater emphasis on the engagement of clinical experts in specifying content requirements for electronic health records. To facilitate this, the NHS in England has used *open*EHR archetype and template design tools. The underlying EN13606 architecture, like HL7 V3, is based on a fairly generic reference model. However, the *open*EHR tools for archetype and template design follow a paradigm that is similar to the design of a structured data collection form. This approach seems more familiar to clinical users than the design of HL7 message models. While this familiarity encourages greater clinical engagement, it does not guarantee consistency and reusability of the captured information. To address this, decisions need to be made about how the captured information is to be represented. This poses questions about the way in which SNOMED CT should be used in association with *open*EHR archetypes and templates. While the ways in which these questions are addressed may in some cases be specific to the archetype methodology, the underlying issues arising from combining structure and terminology are similar to those encountered by the TermInfo Project.

In addition to the theoretical similarities, there are practical reasons for considering the relationship between this work and the TermInfo Project. Health record content specified using this approach may subsequently be communicated using HL7 messages or documents. Consistent approaches to the integration of terminology with information models are likely to simplify any necessary transformations between these different structures.

METHODS

Identifying and managing overlaps

The TermInfo Project started by considering specific questions about how particular items of clinical information should be represented. In several cases, more than one option was found and discussion centered on which of these was the 'correct' or 'best' option. In each of these cases, the alternative approaches arose from the ability to express the same meaning, using either a structural element or a facet of the terminology. Therefore, the focus of the work shifted to identification of the areas of overlap between the semantics of HL7 Version 3 information models and SNOMED CT. The HL7 Clinical Statement Pattern[3], a common model for clinical information representation within HL7, was used as the practical point of reference for examples.

This allowed systematic analysis of alternative solutions to particular types of issue, leading to more consistent resolution. Where overlaps were identified, the options shown in Table 1 were considered.

	HL7 Representation	SNOMED Representation
1	Required	Required
2	Optional	Required
3	Required	Optional
4	Required	Prohibited
5	Prohibited	Required
6	Optional (either or both)	
7	Optional (either one but not both)	

Table 1 – Options for overlaps.

Depending on which of these options is chosen different rules are required to derive one form from the other or to validate the consistency of dual representation. If both representations mean precisely the same, then either option is equally acceptable. However, in many cases there are differences in the precise nature of the information or level of detail. Ambiguity may also arise when both representations are permitted because the second representation could be interpreted as a restatement or a combinatorial factor (e.g. "a request to request …", "finding … not absent", "family member has family history of ...").

The TermInfo recommendations address the most common overlaps with specific guidance on preferred representations that resolve these ambiguities.

Identifying and managing gaps

In some cases, neither the information model nor the terminology may offer a way to meet a particular requirement. In theory, a gap is easier to address than an overlap because it simply requires a decision on which component should be extended to meet the requirement. However, requirements for resolution to meet an immediate business need may force the use of an interim measure or work-round. More detailed analysis, by those responsible for the relevant standard component, may lead to a different recommended approach. The end result may be to turn a gap into a future overlap as the work-round is replaced by a more appropriate solution.

To minimize the risk of short-term decisions turning into new legacy issues, gaps were documented and passed to the relevant organization or expert for rapid evaluation. Even where the release cycle for contributing standards makes a short term fix essential, this type of approach reduces the likely impact of future substantive correction.

Binding terminology to specific structures

The NHS work took detailed *open*EHR templates specified by clinical groups as a starting point. The objective was to identify appropriate ways to bind elements of SNOMED CT to information model nodes in order to represent the intended meaning.

There was an urgent business requirement to apply codes to a set of pre-existing templates. However, the need for a more consistent approach was also recognized. To facilitate this, the short-term exercise of coding specific templates was augmented with a more systematic review to identify the types of issues encountered and to propose a more systematic and scalable approach for future NHS development.

RESULTS

General comment

This section summarizes some common themes arising from the activities described. Our intention is to highlight some key findings rather than to provide an exhaustive list of all the issues encountered.

Managing semantic granularity

A general challenge for using a terminology with an information model is aligning classes and attributes in the model with the expressivity supported by the terminology. There is a requirement to match the semantic granularity of coded expression from the terminology with the slots in the structural model. If the information model provides a single coded attribute to represent a particular concept, this assumes that the terminology contains a code to represent that precise concept.

SNOMED CT allows codes to be post-coordinated to create expressions representing more specific concepts. The model for these post-coordinated expressions is described in 'SNOMED CT Abstract Models and Representational Forms'[7] and approved domain and range constraints are published in the 'SNOMED CT Technical Reference Guide'[8]. SNOMED documents also specify transformation rules that can be applied to normalize expressions to enable computation of equivalence and subsumption[9]. Post-coordination can only be used if the information model provides a structure that can accommodate this type of representation. Similarly, the rules for normalization have a dependency on any semantics embedded in the surrounding structures.

In most cases, each class in the HL7 Clinical Statement pattern represents a unit of information that can be readily coded using a single SNOMED CT expression. Furthermore, the HL7 coded data types support post-coordination. Thus the level of coding granularity was relatively easy to align with the classes in the model. Some HL7 classes also contain additional coded attributes which, while necessary when using other code systems, duplicate information present in a single SNOMED CT expression. Most of these attributes are optional and can be refined out of specific models to minimize potential confusion.

In contrast the *open*EHR related work involved review of specific archetypes and templates. The intention of this work was to assign appropriate terminology bindings to each coded node in the template. Initial review of these identified a wide range of different structural granularities. As a result, the appropriate SNOMED CT expression may depend on the values entered in three or more separate but related nodes in a branch of the template. This presents a significant problem for terminology binding, since, if the individual slots in the template are coded independently, similar types of information may be coded quite differently. More importantly, these different representations would not be amenable to normalization without a clear understanding of the semantic relationships between the separate coded slots. It may be possible to apply more rigorous semantics to the design process to preemptively reduce these variations. However, for the purposes of the current work, the chosen approach was to retrospectively identify the units of clinical meaning that could be appropriately captured by SNOMED CT expressions. The co-dependencies between different nodes in the archetypes and templates were captured and linked to the appropriate SNOMED CT constructs using XPATH.

Context, situations and sections

Alternative representations of contextual information were another common finding from both activities. The SNOMED CT concept model includes attributes that allow representation of various clinical situations such as family history, past history and current findings. The objective of this part of the model is to clearly distinguish between the same finding in difference contexts. For example, to ensure that 'family history or asthma' is subsumed by 'family history of respiratory disorder' but not by 'past medical history of asthma'.

Both HL7 and *open*EHR provide structural conventions for representing these types of contextual information. Structural options include the use of a document section, a specific entry in a template and references to the subject to whom the information applies.

Each of these approaches has distinct merits. A section-based approach matches the way many clinicians work when capturing and reviewing data. Structures that allow references to specific family members are more flexible for representing genetic information. The SNOMED CT approach allows a single coded expression to unequivocally represent family history.

The key to managing these differences seems to be to allow them to be safely combined by ensuring that the way terminology is bound to the structures facilitate transformation to a common normal form. If a family history section is used, this must be bound to the SNOMED CT representation of family history so that the disorder concepts listed within the section can be reliably transformed into appropriate SNOMED CT expressions for analysis. Similarly, if a structural model is used to represent relationships to specific people, the types of relationship (e.g. parent, brother, sister, etc.) should still be represented using SNOMED CT.

Detailed entries, summaries and check-lists

Structural models for representing clinical information may include assumptions about the level of detail captured. For example, some models assume different structures for the detailed story of presenting complaint, a summary of past history and a general review of symptoms affecting body systems. This approach aligns with the way that paper records are written and with the design of data collection forms for specific types of condition or consultation.

In contrast, SNOMED CT provides concepts at different levels of detail that can be used in a range of situations. The structure of the terminology allows more detailed refinement to be added where this is appropriate. This approach assists with retrieval for decision support or analysis, as the way in which the data is recorded is not specific to the way in which it is captured.

It is possible to combine these approaches by binding lists of summary values in a template to relevant concepts in the terminology. However, in both the HL7 and *open*EHR related work, this raised important questions about the intention behind a chosen data collection paradigm and information model structures that mimic it. These issues, which apply to many types of structured data collection, are seen most clearly in relation to check-lists.

There is a clear consensus that check-lists are a useful or even essential tool for effective data collection. However, in both pieces of work it was evident that there are different views about the representation of the information captured using check-list. These views can be characterized as:

a) Representing the information as captured.
b) Representing the information independently of the way in which it was captured.

View (a) represents each entry in the list as the name of the check-list item (i.e. either text or a linked code) and a value (e.g. 'true', 'false', 'not known') based on the response given. This approach is concerned with capturing information about the completion of the check-list and also ensuring the reviewer knows how the data was acquired.

View (b) represents the meaning implied by each entry in that same way, as if that information was captured in another way (e.g. by selecting a code from a terminology search). This approach seeks to ensure the information can be used to return reliable answers to questions irrespective of the nature of the user-interface. View (b) can be seen as a representation of what Rector[10] describes as the 'model of meaning' while view (a) is a specific 'model of use'.

Strong arguments can be advanced for meeting both sets of requirements. However, the balance between them depends on the rationale for using a check-list, and the value of reusing the captured data.

Further investigation of the use of check-lists identified a range of reasons for specifying requirements using check-lists:

- To remind the clinician to ask or consider a question.
- To record whether a question was asked or considered.
- To allow rapid entry of common significant information without recourse to searches.
- To provide an example of the type of information that should be recorded – presuming that other entries can be added as needed.
- As a single place to look for and maintain key information – assumes that the check-list may be populated from previously collected data.

Even within the same NHS *open*EHR template, the reasons for using check-lists varied. These differences may influence decisions on terminology binding. Depending on the reason for using a check-list approach, there may also be a requirement to represent view (a) to audit the process of care and/or data collection. Irrespective of the process, if the information is to be reusable for clinical purposes, the consistency offered by view (b) also needs to be supported.

Interdependencies between multiple data nodes

As noted earlier in this paper, there may be differences in semantic granularity between structural and terminological components. These differences mean that in some cases multiple nodes in the structure need to be considered to generate a single SNOMED CT expression. However, this is only one of the types of interdependency noted during this work.

The value applied to one node may constrain the potential range of coded expressions that can be applied to another node.

An example of this is the case where the structural model provides separate attributes for 'disease', 'site', and 'laterality'. Depending on the specified disease the site may either be superfluous (e.g. appendicitis) or essential (e.g. 'fracture') and the relevant of 'laterality' may depend on the selected site. Even in the case of disorders without a fixed site, a post-coordinated expression might contain the site and/or laterality.

Many interdependent constraints may be expressed by reference to the SNOMED CT concept model. However, this depends on the assumption that the specific nodes in the structural model are aligned with the relevant attributes in the concept model.

In other cases, the existence of a preferred or mandatory form or representation in the structural model may indirectly constrain the use of terminology.

The HL7 Pharmacy models represent the action of administering a substance in one class (Act) and the substance administered in another associated class (Entity). Both the nature of the action and the entity can be represented using SNOMED CT concepts. However, SNOMED CT also supports expressions that include the substance administered as a refinement. To avoid conflicts with the models, coded expressions that incorporate the substance may need to be prohibited.

> For example, the concepts 'subcutaneous injection' and 'insulin' might be used in the two associated classes but the concept 'subcutaneous injection of insulin' might not be permitted.

DISCUSSION

Terminology, structure and meaningful records

Electronic health records offer a range of potential benefits. Many of these depend on being able to consistently process meaningful clinical information within those records. Two distinct threads have developed to address this requirement – a structural thread and a terminology thread.

The structural thread places emphasis on the set of specific items of data that express a particular class of clinical information. In contrast, the terminology thread seeks to provide reusable codes or labels for events or ideas. These two threads have developed and work together in almost all areas in which information is processed. This symbiotic co-existence is apparent at all stages in the life cycle of an item of clinical information - data entry, display, storage, communication and retrieval. Different approaches to the use of structure and terminology have developed in proprietary clinical systems and efforts to develop standards have tended to separate terminological and structural aspects.

Previous work on binding between information models and SNOMED CT reported by Sundvall[11] noted the value and limitations of simple equivalence binding between a node and a terminology concept. It emphasized the need for a powerful constraint binding formalism to address these limitations.

Clinical information life-cycle perspectives

Different approaches to representing clinical information often arise as a result of perspectives that are influenced by particular stages in the life-cycle of that information (see figure 1). All three components considered by the work described in this paper have the broad ambition of representing meaningful clinical information. However, each of them has a significantly different perspective. The focus of HL7 Version 3 is on interoperable communication and thus it specifies static and dynamic models related to interaction between discrete applications. SNOMED CT takes a retrieval perspective; by representing subsumption and interrelationships between the different concepts, it enables effective subsequent retrieval for multiple purposes. EN13606 archetypes have a similar role to the classes of the HL7 RIM. However, *open*EHR archetype and template design, are more directly influenced by the data capture perspective. Each template reviewed walks through the typical process of collecting data during a particular type of clinical encounter. As shown in Figure 1, these perspectives are interdependent.

The primary rationale for binding SNOMED CT to structured clinical information is to enable selective retrieval and reuse of information.

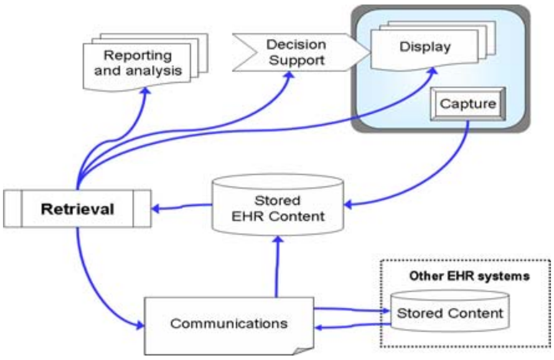


Figure 1 – Clinical information life-cycle (summary)

The process of integrating terminology with structure may also involve some normalization of the structure to address the anticipated retrieval requirements. Structural differences may obscure semantic similarities and binding a code to a field will not necessarily deliver the full potential of the terminological component. For example, a template may structure some items of current and past clinical history in the form of a checklist, some as codes chosen from a picking list, and others as more detailed collections of coded and textual data items. A degree of normalization may be essential to ensure that the record can be used to answer questions such as, 'does the patient have a past clinical history of respiratory problems?'

Balancing the use of structure and terminology

Both structural and terminological approaches have specific strengths and weaknesses. Those wishing to exploit the strengths of a particular structural or terminological approach may differ in their perception of the appropriate balance between these components. However, there is general agreement that some facets of clinical information are best represented using structure, while others are more effectively expressed using terminology.

Figure 2 summarizes a consensus position agreed in the *open*EHR related terminology binding project. Different facets of clinical information were identified and assigned to one of five categories indicating whether a terminological or structural approach was recommended and the strength of that recommendation. The two outer categories encompass facets that can only be effectively represented using one of the approaches. Two further categories include facets for which one approach has a clear advantage but the other approach is also possible. Between these is a 'gray area' in which the relative merits of the two approaches are more finely balanced or may depend on a specific use case.

Practical principles for terminology binding

The following principles are suggested as a basis for detailed recommendations on integration between any combination of a terminology and a structural model. These principles are based on those agreed by the HL7 TermInfo Project. They have been revised and extended to take account of more recent practical experience summarized in this paper.

1. Understandability

The recommendations must be understandable by implementers who are familiar with the use of the terminology and structural models being integrated.

The integration recommendations need not repeat general advice on the underlying components but should not require other pre-existing knowledge.

2. Reproducibility

The recommendations should be tested on members of the intended target audience of implementers to ensure they are interpreted and applied consistently.

3. Usefulness

The recommendations need not cover all possible use cases but should cover all the most common scenarios encountered in the intended scope of use.

4. Reusability and common patterns

Representations that can be reused consistently in many contexts should be recommended in preference to those that are specific to a particular context.

> For example, the representation of a finding should follow a similar pattern whether recorded as a problem, a new diagnosis, an item of past medical history, detailed documentation of presenting complaint or a discharge diagnosis.

5. Transformability and normal forms

If alternative representations are permitted, rules should be specified to unambiguously transform these into a common representation.

Terminology model only

Specific concepts:

> For example, diseases, symptoms, signs, procedures, drugs, etc.

Semantic relationships between concepts

> For example, relationship between 'viral pneumonia', 'lung', 'virus', 'infectious disease'.

Representation of constraints on use of terminology

> For example, concept model and value-set definition formalism.

Terminology model preferred (structural model deprecated)

Constraints on combination of concepts in instances including abstract model of post-coordination and permissible attributes and ranges for refinement of concepts in specified domains:

> For example, restrictions on 'finding site' refinement of 'appendicitis', conventions on representing laparoscopic variants of a procedure.

Gray area (preference unclear or use case dependent)

Representation of contextual information related to instances of clinical situations

> For example, family history, presence/absence, certainty, goals, past/current, procedure done/not-done.

Representation of additional constraints on post-coordination of concepts for specific use cases

> For example, constraints on terminology use specific to immunization and related adverse reaction reporting.

Structural model preferred (terminology model deprecated)

Representation of relationships between distinct instances of record entries and other classes

> For example, assertions of causal relationships between entries, grouping of entries related by timing, problem or other organizing principles.

Structural model only

Attributes with specific data types

> For example, dates, times, durations, quantities, text markup.

Identifiable instances of real-world entities

> For example, people, organizations, places.

Overall record and/or communication architecture

> For example, EHR extract, EHR composition, openEHR reference model, CDA documents, HL7 messages.

Representation of constraints on use of particular classes or attributes in given use cases

> For example, formalism for templates applied to constrain openEHR archetypes or HL7 CDA documents.

Figure 2 – Strengths of structure and terminology

6. Tractability

Requirements for tooling to transform or validate instances that conform to the recommendations should be computational tractable.

7. Practicality

Existing tools and applications, either in their current form or with reasonable enhancements, should be able produce the recommended instances.

8. Scalability

Recommendation should not require a combinatorial explosion of pre-coordinated concepts.

> For example, the model should not require the creation of the cross product of "Allergic to" and all drugs and substances.

9. Limiting arbitrary variation

Optionality should be restricted where possible to limit arbitrary variations. Where more than one approach appears to be equally valid based on other criteria, a single approach should be recommended to avoid unnecessary variation.

> If one approach has already been successfully implemented and the other has not, the approach that has been implemented should be selected.
>
> If two or more approaches have already been implemented, one should be recognized as the preferred form. Other approaches that are already in use may be permitted but should not be recommended for new implementations.

10. Responsive participating standards

The participating structural and terminology standards should provide prompt mechanisms to enable notification and correction of gaps and inconsistencies. These mechanisms should be used rather than local work rounds, to avoid increasing the number alternative representations. Implemented systems and participating standards should be sufficiently agile to allow rapid and reasoned development of effective compositional solutions.

Requirements for specific guidelines

The principles outlined in this paper are only a foundation. Practical implementation requires detailed specific guidelines for integration between SNOMED CT and an information model. The first detailed guide on use of SNOMED CT with HL7 Version 3 is now available as a Draft Standard for Trial Use[6]. Detailed guidance related to a trial set of *open*EHR archetypes and templates is under review but has yet to be finalized and more widely published.

Dependency-aware evolution

An original design goal of SNOMED CT was usability in applications with different information models. Likewise, the standard information models of HL7 Version 3 and EN13606 were designed to enable use of different terminologies. Thus HL7 specifications include coded attributes that need to be bound to specific value sets before implementation. Similarly, *open*EHR (a development based on EN13606) states[5] that its fundamental building blocks (archetypes) are 'terminology neutral' and that a single archetype can be 'bound to more than one terminology'.

This mutual openness between alternative code systems and information models seems an attractive proposition. However, we contend that the extensive overlaps and interdependencies demonstrated by the work described in this paper point to a requirement for closer mutually aware development of information models and terminologies. While tools and guidelines for binding are necessary to address the interface between current information models and terminologies, they are unlikely to be sufficient unless future development of information models and terminologies take due account of the need to work together rather than as independent variables.

Acknowledgments

The HL7 TermInfo Project was facilitated by Health Level 7 of Ann Arbor, USA. The work described in on binding terminology to *open*EHR archetypes and templates is related to models under development at NHS Connecting for Health.

References

1. IHTSDO, SNOMED CT User Guide, IHTSDO Copenhagen, Denmark 2007, http://www.ihtsdo.org/ fileadmin/user_upload/Docs_01/Technical_Docs/snomed_ct_user_guide.pdf [cited 2008 January]
2. Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. Methods Inf Med. 1998;37(4–5):394–403.
3. HL7 PC TC. Clinical Statements, Health Level 7, Ann Arbor, USA 2007. http://www.hl7.org/v3ballot/html/domains/uvcs/uvcs.htm [cited 2008 January].
4. HL7 SDTC. Clinical Document Architecture Release 2, Health Level 7, Ann Arbor, USA 2005, http://www.hl7.org/v3ballot/html/infrastructure/cda/cda.htm [cited 2008 January].
5. Beale T, Heard S, Kalra D, Lloyd D, Introducing openEHR. Great Britain, The openEHR foundation 2006, http://www.openehr.org/releases/1.0.1/ openEHR/introducing_openEHR.pdf [cited 2008 January].
6. HL7 Vocabulary TC, Using SNOMED CT in HL7 Version 3; Implementation Guide release 1.4), Health Level 7, Ann Arbor, USA 2007. http://www.hl7.org/v3ballot/html/infrastructure/terminfo/terminfo.htm [cited 2008 January]
7. IHTSDO, SNOMED CT Abstract Models and Representational Forms, IHTSDO, Denmark 2007. http://www.ihtsdo.org/fileadmin/user_upload/Docs_01/Technical_Docs/abstract_models_and_representational_forms.pdf [cited 2008 January].
8. IHTSDO, SNOMED CT Technical Reference Guide IHTSDO, Denmark 2007.

9. IHTSDO, SNOMED CT Transformations to Normal Forms, IHTSDO Copenhagen, Denmark 2007. http://www.ihtsdo.org/fileadmin/user_upload/Docs_01/Technical_Docs/transformations_to_normal_forms.pdf [cited 2008 January].

10. Rector A, Qamar R, Marley T. Binding Ontologies & Coding systems to Electronic Health Records and Messages, KR-MED 2006, http://www.cs.man.ac.uk/~qamarr/papers/Terminology-binding-KRMED-rector-final.pdf [cited 2008 January].

11. Sundvall E, Qamar R, Nyström M, Forss M, Petersson H, Åhlfeldt H, Rector A, Integration of Tools for Binding Archetypes to SNOMED CT, Semantic Mining Conference on SNOMED CT, Copenhagen, Denmark, 2006, http://www.hiww.org/smcs2006/proceedings/12Sundvall SMCS2006final.pdf [cited 2008 January]

Strategies for Updating Terminology Mappings and Subsets using SNOMED CT®

John Mapoles, Ph.D., Corey Smith, Jane Cook, Brian Levy, M.D.
Health Language, Inc., Aurora, Colorado 80011

Abstract:
SNOMED CT® (SCT) is a large, comprehensive medical terminology with many applications in the health care IT sector. SCT is often mapped to existing billing classifications as well as to proprietary terminologies in order to support access to and from SCT from existing applications. Subsets of SCT are used to reduce complexity and size. These subsets can vary from small sets that will be used to populate drop down lists in electronic medical record applications and larger lists that are used for reference, e.g. the SCT Non-human subset. There are costs and time limitations in maintaining mappings and subsets, particularly after each SCT release when concepts are retired and new concepts are added. There is a need for a careful strategy to identify changes, determine which changes need to be reviewed, and to rank changes so they can be reviewed systematically in order of importance. Here we outline our updating strategies for the Health Language Medical Specialty Subsets, a list of 10,000 SCT terms grouped into 45 subsets. These strategies can be used for any subset of SCT as well as for mappings created to and from SCT.

Introduction:

Electronic health records (EHR) and other healthcare IT applications rely on controlled medical terminologies to provide well defined concepts for accurate and consistent encoding of records and data mining. SNOMED CT® (SCT) provides broad coverage of all medical domains with approximately 280,000 active concepts. A great deal of attention has been focused on the models and strategies to implement SCT[1,2,3]. Most applications will use defined subsets of SCT for specific use cases rather than exposing all of SCT to all users. Subsets are collections, lists, of SCT concepts or terms. Applications using SCT will need to be able to store these lists for purposes of maintenance and delivery to EHR interfaces.

Mappings are often created between SNOMED CT and other terminologies such as billing classifications, ICD-9-CM and ICD-10, as well as local and proprietary terminologies. Mappings can be represented as a pair of codes, the source and target of the map or relationship. These mappings can be used for translation of information from one set of codes to another.

These mappings and subsets represent work that a local site or user is performing to the standard – in this case SCT. Thus, SCT is distributed from the standards body, and local users such as EHR vendors or hospitals, need to add value to their SCT version with mappings and subsets. It is critical that the local user adopts the next version of SCT in order to prevent semantic drift – multiple versions of a terminology being used that drift in meaning enough to be incompatible. In the paper, Oliver[4], et. al. discuss some of the principles of localization of terminologies and also explores the impact of migration of localized terminologies to the next version of the standard. Updating subsets and mappings as described in this paper actually only involves a small subset of the many kinds of changes that occur to the terminology. Cimino[5], et. al. discuss the types of changes that need to be considered such as refinement, name changes, code-reuse and more that impact the updating of terminologies and content based on them.

The Semantic Web work is now introducing new issues with regards to ontology versioning. Liang[6], et. al. discuss the impact of changes in ontologies to existing applications that depend on them. In this semantic web paper, a middle layer to monitor and detect changes is proposed to be used between the underlying ontologies and the dependant applications. The work we present here with subsets uses a terminology service and specialized scripts to serve as this middle layer between the standard ontology, SCT in this case, and the resulting applications – EMRs for example that depend on the subsets.

Maintenance of the mappings and subsets are costly and time dependent because the content is often in production. Maintenance can involve changes that are dictated by the applications that use the content but also because the underlying SCT data model has changed. Semiannual updates to SCT can change the SCT model to varying degrees, sometimes substantially. An SCT update involves new concepts and terms, retirement of concepts and terms, and addition and deletion of relationships. Each mapping and subset needs to be evaluated so that only the specific groups of changes that immediately affect the content are considered. This analysis is essential for performing the update quickly and efficiently.

Health Language (HLI) provides a terminology server, or Language Engine (LE) that serves up terminologies and related content to other applications. HLI also provides modeling and mapping tools to allow users to update and localize the content in their terminology server. Health Language (HLI) has developed a set of 45 subsets, Medical Specialty Subsets (MSS), that represent "clinically friendly" terms used in medical practice. These subsets represent 10,000 entries that must be rapidly updated for production release. Changes to SCT must be filtered so that changes that impact the MSS are reviewed, without the distraction of changes that are not important to the MSS. This paper will discuss the principles used to update the subsets. These same principles are also used to update our various mapping projects which include maps between SCT, billing classifications, and proprietary terminologies.

Methods

Medical Specialty Subsets: The MSS consist of 45 subsets containing 10,000 SCT terms. The subsets contain “clinically friendly” SCT terms most often encountered in clinical practice. The subsets were constructed so that SCT concepts were specific to the level of granularity commonly needed by clinicians in that specialty practice. The top approximately 150 diagnoses and procedures applicable to each specialty were incorporated into the subsets. Claims data as well as medical domain expert knowledge was used to determine these concepts that are applicable for the subsets. Concepts with low incidence based on claims data, e.g. rabies, were not included.

Only concepts mapped to billable ICD-9-CM and CPT codes were used to construct the subset, using the College of American Pathologists SCT to ICD-9-CM cross maps and HLI SCT to CPT cross maps. This increases the possible utility of the subsets for billing purposes-

Some concepts may be too specific for one subset, but applicable in another. For example, *Acute anterior myocardial infarction* may be too specific for the Family Practice subset, but applicable in the Cardiology one.

The MSS are stored in the HLI LE® database. Members of the subsets can be viewed and managed using the HLI browser and editing tool LExScape® or the HLI Java application programming interface (API). HLI APIs are a full set of Java classes, interfaces, and methods that allow access and management of data in an HLI LE database. The APIs can be bundled into complete applications such as LExScape and other HLI management tools or they can be used to terminology enable applications that require both terminology support and other functionality not related to terminology. The APIs are collected into standard jar files and can be accessed and using standard Java programming methods.

Terms included in the MSS must meet the following requirements:

a. Subsets related to disease take terms from concepts in the SCT Clinical Findings taxonomy.
b. Subsets related to procedures take terms from concepts in the SCT Procedure taxonomy.
c. Terms must have an SCT description status of 0 on a concept with a status of 0. These are referred to as active terms. A status of 1 or greater is retired or limited in use.
d. A subset can contain only one term from each concept.
e. Concepts that contain terms in the disease subsets must have a valid relationship to ICD-9-CM as defined by the College of American Pathologists produced SCT to ICD-9-CM cross mapping. Only those concepts that have a cross map to a billable ICD-9-CM code are considered.
f. Concepts that contain terms in the procedure subsets must have a valid relationship to CPT as defined by the HLI SNOMED - CPT relationships. Only those concepts that have a cross map to a billable CPT code are considered.

SNOMED CT: The SCT release of January 31, 2007 was downloaded from the College of Amercian Pathologists and was transformed into the HLI's data structure and stored in the LE database.

Changes to SCT: Changes to SCT between releases are calculated by comparing consecutive versions of SCT using the HLI Java APIs. The SCT core vocabulary consists of three broad object types: concepts, terms (descriptions), and relationships. Although concepts and terms are never deleted from SCT they can change status from active to limited or retired. Concepts can also move between SCT hierarchies. The number of changes that occur each time SCT is updated varies widely and is distributed throughout all SCT taxonomies.

Within the scope of this project we are concerned with term changes, because the subsets are lists of SCT terms. Concept changes must also be tracked because if a term's concept moves out of a subset's target hierarchy the term can no longer be used in the subset. Changes in the defining relationships of the concepts generally do not impact the placement of the concept in the MSS because if the concept changes in meaning completely, then it will be retired in SCT. The types of changes that are considered were:

a. Concepts or terms that became limited or retired.
b. New terms in the target hierarchy.
c. A term's concept no longer has a relationship to a reference terminology, i.e. ICD-9-CM or CPT.

These change types were then analyzed by creating custom Java scripts written to the HLI APIs. These scripts can then be re-run for each SCT update. The output of these scripts are then fed into the HLI modeling and mapping tools. Each change type is presented to the modeler in the tool as a separate project of affected concepts; thus a collection of small update projects is generated for each SCT release.

Once the update is completed, the MSS are then versioned as a set and released. Versioning collects all subset changes, including new terms, and stamps them with a version number applicable for the HLI product. Each version number is then tied to a release of SCT.

Results

The goal of any update is to isolate changes that bear directly on the data sets of interest because the SCT update is generally so large that all changes can not be reviewed in a timely fashion. Once the changes that impact the data sets are isolated they must be categorized and arranged so that the most important changes are reviewed first.

For the MSS the following change types are considered important:

1. Invalid concepts or terms: these are concepts or terms that have been retired, that have changed to status limited SCT status value 6 - representing a classification or administrative concept, that are no longer mapped to a reference hierarchy, or that are still active but have been moved out of the target taxonomy.
2. New terms on concepts in the subset: Because terms from these concepts are already in the subset a new term on these concepts are especially interesting as a possible replacement.
3. New terms on descendents that have a relationship to ICD-9-CM or CPT: Descendents of concepts that are in the subset are of special interest because they are likely to contain terms that are more specific than the term in the subset.
4. All other new terms: These are all new terms in the target taxonomy that are not part of the type-2 and type-3 group. These are due to the addition of new concepts and the addition of new terms on existing concepts.

Changes should be reviewed in the order of importance. Limited time and resources dictate that essential changes must be repaired first.

Changes in relationships and qualifiers in SNOMED CT are not considered during the udpates process. Minor changes in relationships are not considered to change the meaning of the concept. Major changes usually cause the concept to be retired and replaced. Changes to IsA relations, and indirectly to defining non-IsA relations, are considered in change types 3 and 4. Changes to qualifying relations are not considered because these relations are not definitional

Management of the process

Changes of type-1 must be reviewed first because they break the requirements of the MSS. These terms must be removed and possibly replaced. Whenever possible the SCT historical relationships (e.g. SAME AS, MAYBE A, REPLACED BY, etc.) are used to identify an active term replacement. For example, in an earlier SCT update, the concept of *Coronary artery thrombosis* was retired and placed into the *Ambiguous concept* hierarchy. This retired concept now has a *MAYBE A* relationship to the active *Myocardial infarction* concept. The Java scripts written to identify the changes include an algorithm to locate these historical relationships and potential replacement SCT concepts. Modelers then can view these potential replacement concepts in the tool.

Once the requirements of the subset are satisfied, possible new or replacement terms can be considered. Changes of type-2 are new terms on concepts in the subsets. These concepts are the most likely to contain replacement terms for existing terms.

SCT is arranged into hierarchies. Concepts become more specific when moving from the top to the bottom of a hierarchy. Changes of type-3 leverage the SCT hierarchy. Terms that are descendents of terms in the subset are likely to be of interest. They are likely to be more specific than existing terms providing replacements or valuable additions. Only concepts that have mapping relations to ICD-9-CM or CPT are considered in this review.

Changes of type-4 include all other new terms in the target taxonomies. These are the terms that are least likely to be of interest but all new terms must be reviewed to ensure each subset contains the most recent SCT content. Terms of type-4 may also contain new ideas for inclusion in subsets. An important part of the management of the process is that all terms of type-4 are reviewed together once and considered as a group for inclusion in all subsets. This is a very large group and reviewing only once is a valuable time saver.

The data for all of the subsets is not presented here. Of the 45 subsets, 24 were not changed at all. Table-1 presents a sample of change data for four subsets.

Subset Name	# of Members	Type-1 Changes	Type-2 Changes	Type-3 Change
Critical Care - Disease Subset	558	3	3	137
Neurology - Disease Subset	367	4	2	185
Gastroenterology - Procedure Subset	185	6	4	3
Neurosurgery - Procedure Subset	327	1	1	7

Table 1: Changes in selected subsets

Subset Name	Changes Made, Type-1 Changes	Changes Made, Type-2 Changes	Changes Made, Type-3 Changes	Changes Made, Type-4 Changes
Critical Care - Disease Subset	3	3	2	1
Neurology - Disease Subset	4	1	0	1
Gastroenterology - Procedure Subset	6	2	2	1
Neurosurgery - Procedure Subset	1	1	0	2

Table 2: Changes made to selected subsets based on type

Data from Table-2 demonstrates how this approach focused the update tasks. The review of each subset is limited to a controlled number of review tasks that are specific to the subset.

Changes of type-1 always result in a corresponding action (Tables 1 and 2). Changes of type-1 break the rules of the subsets so they must result in a change.

Changes of type-2, type-3, and type-4 are new terms. Changes of type-2 are new terms on the concepts that have a term already in a subset. Of the 14 candidates 7 were added to the subsets. The percentage of type-3 changes added to the subsets is much smaller. These are terms on descendents of concepts already in the subsets. They are added at a much lower rate because modelers consider them too specialized for the subsets.

Table-2 also demonstrates that Type-4 changes are important. They add valuable new content to the subsets. The percentage of added terms here is low as would be expected since these terms are on concepts that are not in the subsets or their descendents. There is pressure to keep the subsets tightly defined to a core group of diagnoses and procedures.

The update task for the HLI MSS has been broken into a series of specific tasks and one large task (type-4). In this release, January 2007, of SCT there were over 3000 new concepts and 3400 new terms on existing concepts as well as 6500 concepts and terms retired. Thus despite the large number of changes that occurred in SNOMED CT in this past release, only a small number of changes actually affected the subsets, limiting the amount of review required.

Discussion

Development of any data model requires time and resources as part of a pre-production effort.

Updates are post-production and under the additional constraint that updates must be finished in a timely fashion so that new production data is available as soon as possible.

We present a strategy here for the maintenance of SCT subsets based on the experience of maintaining the Medical Specialty subsets. These strategies can be adapted to any subset or mapping created with SCT or other terminologies. The basic premise is that it is not feasible to review all entries in all subsets after each new SCT release. Thus, we identified the types of changes that occur in SCT that would be more relevant to the subsets. These 'units' of change are then ranked in order of importance. Automated processes are employed whenever possible followed by manual review of the changes when necessary. The SCT hierarchies are also leveraged to allow for identification of possible new, more specific terms to be included in the subsets.

Similar strategies can be used to update mappings to and from SCT as well. As in the case for subsets, a set of requirements defines the rules used to create the mappings. Many of the same change types described above that affect subsets also may affect mappings as well. But, in addition to changes in SCT, consideration needs to be given toward similar changes in the other terminology being mapped to or from SCT. For example, a mapping between ICD-9-CM and SCT needs to be updated for changes to ICD-9-CM as well as SCT. Therefore, evaluation for other change types particular to the non-SCT terminology need to be considered. Modelers then need only review a portion of the maps that are affected by the changes. In the case of various mapping projects that HLI has performed, these reviews usually only include hundreds of concepts instead of tens of thousands.

Changes to medical domains in the real world define the requirements for medical terminologies such as SNOMED CT. New terms that are added to SNOMED CT to meet these requirements. This paper discusses how HLI manages these changes in the MSS There will always be cases where the medical world moves faster than the terminology. SNOMED CT provides a mechanism for post-coordination and extension to fill these gaps. HLI however has decided, as a rule, that the Medical Specialty Subsets only use pre-coordinated concepts. HLI does make submissions to the International Healthcare Terminology Standards Development Organization (IHTSDO) whenever a new concept is required to fill a gap between the medical domain and the terminology. Using the HLI framework MSS users can make additions and deletions to their local copy of the MSS to account for local conditions.

As SCT becomes more widely used, managing its changes and effects on content based on SCT will be critical. Using efficient processes such as those identified here will help manage SCT changes.

References

1. Richesson R., Young K., Guillette H., Tuttle M., Abbondondolo M, and Krischer J. Standard Terminology on Demand: Facilitating Distributed and Real-time Use of SNOMED CT During the Clinical Research Process. AMIA Annu Symp Proc. 2006; 2006: 1076.
2. Vikström A., Skånér Y., Strender L-E., and Nilsson G. Mapping the categories of the Swedish primary health care version of ICD-10 to SNOMED CT concepts: Rule development and intercoder reliability in a mapping trial. BMC Med Inform Decis Mak. 2007; 7: 9. Published online 2007 May 2. doi: 10.1186/1472-6947-7-9.
3. Jacobs A., Quinn T., and Nelson S. Mapping SNOMED-CT Concepts to MeSH Concepts. AMIA Annu Symp Proc. 2006; 2006: 965.
4. Oliver DE, Shahar Y. Change Management of Shared and Local Versions of Health-Care Terminologies. Meth Inform Med 2000;39:278-290.
5. Cimino JJ. Formal Descriptions and Adaptive Mechanisms for Changes in Controlled Medical Vocabularies. Meth Inform Med 1996;35:202-210.
6. Klein M., Fensel D*., Ontology Versioning on the Semantic Web,* Proceedings of the International Semantic Web Working Symposium (SWWS), Stanford University, California, USA, July 30 -- Aug. 1, 2001.

Using SNOMED CT as a Mediation Terminology: Mapping Issues, Lessons Learned, and Next Steps Toward Achieving Semantic Interoperability

Sarah Maulden, MD, MS[1], Patty Greim, RN, MS[1], Omar Bouhaddou, PhD[1,2], Pradnya Warnekar, RPh, MS[1,3], Laura Megas[4], Fola Parrish, PharmD[5], Michael J. Lincoln, MD[1,6]

1 Department of Veterans Affairs, Veterans Health Administration, Chief Health Informatics Office, Salt Lake City, UT
2 Electronic Data Systems, Plano, TX
3 dNovus RDI, San Antonio, TX
4 Northrop Grumman Corporation, Chantilly, VA
5 Department of Defense, Tricare Management Activity, Falls Church, VA
6 Department of Biomedical Informatics, University of Utah, Salt Lake City, UT
sarah.maulden@va.gov, patricia.greim@va.gov

ABSTRACT

The Clinical Data Repository / Health Data Repository (CHDR) project is a combined effort of the Department of Veterans Affairs (VA) and the Department of Defense (DoD) to exchange clinical information between our Electronic Health Records (EHR). CHDR exchanges standardized, computable data, as opposed to textual data that is only human readable. CHDR utilizes mediation terminologies for health data exchange. For allergy reactions data, CHDR uses SNOMED CT in conformance with Health Information Technology Standardization Panel (HITSP) recommendations. This paper reports how we implemented this solution.

Business rules for mapping allergy reactions were established jointly. Each agency independently mapped its legacy data to the same version of SNOMED CT. CHDR has since been implemented in seven locations where VA and DoD have joint patient care environments. Statistics on actual patient data from February-June 2007 showed a 74-99% mediation success rate for allergy reactions data.

Examination of mediation failures exposed issues related to mapping and SNOMED CT concept modeling. In addition, we emphasize the significance of adherence to a detailed terminology mediation strategy, desirability of a standard SNOMED CT-based subset for allergy reactions, and the creation of this subset for publication and distribution.

INTRODUCTION

The President has ordered Federal agencies to promote improved healthcare quality and efficiency through secure and standard-based data exchange[1]. When clinicians exchange data, interoperable meaning is possible because clinicians share structures of clinical practice and familiar clinical language[2]. Similarly, meaningful electronic data exchange requires a shared structure for transmission and a common electronic vocabulary[3], which yields Computable Semantic Interoperability (CSI)[4]. CSI makes order checks and electronic alerts possible across institutions, and is an essential component of a longitudinal EHR that protects patient safety.

The CHDR project is a Congressionally-mandated, combined effort which aims to exchange standardized, computable data, as opposed to textual data that is only human readable. Computable data exchange enables "semantic interoperability" and permits utilization of electronic decision support tools on the sum of local and remote data at either agency[6]. CHDR currently exchanges pharmacy and allergy data elements and the agencies are working to share laboratory data elements by the end of fiscal year 2008.

CHDR has informed the Health IT Standards Panel (HITSP) that designates interoperability standards for EHRs. VA and DoD use different internal data standards for allergies, and under CHDR utilize a common, HITSP-specified mediation terminology. CHDR exchanges pharmacy, drug allergens, and allergy reactions, and will soon exchange laboratory (chemistry/hematology) data. CHDR exchange of comprehensive pharmacy information[7] and drug allergy reactant information[8] have been well described.

The CHDR strategy for exchange of allergy reactions (signs and symptoms) data uses SNOMED CT, in conformance with Consolidated Health Informatics (CHI) and now HITSP recommendations. We now report how VA and DoD have used SNOMED CT successfully as a mediation terminology, and describe the results.

METHODS

Initial work for allergy reactions under the mediation approach included the commitment at each agency to normalize legacy terms, using a list of centrally maintained concept terms[9]. Allergy reactions were comprised primarily of signs and symptoms, but could also include disorders or clinical conditions attributable to exposure to a drug reactant. Each agency mapped its legacy allergy reactions data to SNOMED CT[10]. The four-part terminology mediation strategy was outlined as follows[11]:

1. Select a mediation terminology compliant with CHI/HITSP standards (if possible).
2. Map each agency's terms to concepts within the mediation standard.
3. Exchange the mediation codes.
4. Coordinate content maintenance plans.

Table 1 shows the CHI standard terminologies and releases designated for the four domains at the start of the CHDR project.

Business rules for mapping allergy reaction legacy terms to SNOMED CT concepts were developed jointly[12]. For example, SNOMED CT hierarchies were prioritized in order of preference for mapping as follows: 1) Findings, 2) Disorders, 3) Morphologic abnormality, 4) Observable entity, 5) Context Dependent Category. Mappings from specific to more general terms (and vice versa) were avoided, because of the bidirectional nature of the data exchange. For instance, mapping "nasal burning" to "burning sensation of mucous membrane (finding)" creates either a loss of the clinical detail "nasal" when translated (for an outbound message), or forces the translation of a general term "mucous membrane" to a specific one--"nasal"--(for an inbound message). Local terms not found in SNOMED CT were collected for potential submission to the SNOMED development organization. Other mapping rules governed misspellings, qualifiers, synonyms, ambiguous terms, and outdated terms.

Table 2 shows a sample of VA allergy reaction terms with their VA unique identifiers (VUIDs) and SNOMED CT mappings.

Once mapping rules were established, terminologists at each agency manually mapped allergy reaction terms to SNOMED CT. VA used Apelon's TermWorks tool and SNOMED's CliniClue® browser, and DoD used the Terminology Service Bureau (TSB) and the CliniClue® browser.

Table 1. CHDR Domains and Designated Standards.

Domain	Mediation Terminology (CHI Standard)
Pharmacy	RxNorm Jun 2005
Drug Allergens	UMLS Jan 2005AA
Allergy Reactions	SNOMED CT Jan 2005
Lab (Chemistry & Hematology)	LOINC 2.14 Jan 2005

Table 2. VA Unique Identifiers, Allergy Reaction Text, and Corresponding SNOMED CT Mappings.

VUID	VUIDText	SNOMED CT ID	SNOMED CT Text
4637123	BLISTER	339008	Blister of skin AND/OR mucosa (finding)
4543527	ORTHOSTASIS	271648003	Postural drop in blood pressure (finding)
4696326	ASEPTIC NECROSIS OF BONE	398199007	Aseptic necrosis of bone (disorder)
4538635	RASH	271807003	Eruption of skin (disorder)
4538640	SEIZURES	91175000	Seizure (finding)
4539274	NOSEBLEED	249366005	Bleeding from nose (finding)

For mapping validation of allergies terms (both reactions and reactants), two reviewers conducted three separate reviews (10 hours each for a total of 60 experts' hours) and identified various discrepancies in about 5% of the total number of terms. All discrepancies were corrected[13]. An independent review of concepts common to both agencies was performed to ensure accurate translation and calculate expected mediation success rates[14]. See Table 3.

Terminology "translation" and "mediation" are described as follows by Bouhaddou et al.:

"The mediation success rate defines the percentage of data in one system that is understood and computable by the other system. For each direction of the data exchange, inbound or outbound, there is a different mediation success rate. For mediation to succeed, two translations have to be successful. First, the source agency has to translate from its vocabulary to the mediation terminology. Then, the target agency has to translate from the mediation terminology to its native vocabulary without loss of meaning[15]."

Mediation success rates are calculated by multiplying the translation success rates of each agency. When coded mediation fails, the CHDR project exchanges allergy reaction data as text without a mediation code.

RESULTS

Terminology translation and mediation statistics were compiled for allergy reactions data during a 5-month period in 2007. The numbers of translation and mediation attempts fluctuated from month to month, but generally showed an increasing trend as the project was implemented at additional sites over the 5-month timeframe. Table 4 shows translation and mediation success rates for allergy reactions sent from VA to DoD. Table 5 shows statistics for allergy reactions sent from DoD to VA. Overall, mediation success rates varied from 74% to 99%.

Table 3. Common and Unique Allergy Reaction Concepts Determined by Each Agency Mapping to SNOMED CT.

Agency	Total	Common Terms	Mapped Terms Unique to Each Agency	Unmapped Terms
VA	346	299	25 (7%)	22 (6%)
DoD	456	299	47 (13%)	110 (24%)

Table 4. VA-to-DoD Mediation Statistics for Allergy Reactions, Feb-June 2007.*

VA-to-DoD	February	March	April	May	June
Total VA-to-SNOMED CT translation attempts	168	193	338	959	502
Translation failures (VA-to-SNOMED CT)	4	0	1	13	1
Total VA allergy reactions sent to DoD	164	193	337	946	501
Translation Success Rate: VA-to-SNOMED CT	**98%**	**100%**	**100%**	**99%**	**100%**
Total allergy reactions received by DoD	164	193	337	946	501
Translation failures (SNOMED CT-to-DoD)	17	17	34	121	5
Total VA allergy reactions sent to DoD CDR†	147	176	303	825	496
Translation Success Rate: SNOMED CT-to-DoD	**90%**	**91%**	**90%**	**87%**	**99%**
MEDIATION SUCCESS RATE: VA-to-DoD	***88%***	***91%***	***90%***	***86%***	***99%***

*Yellow areas designate translation services performed by VA. White areas designate translation services performed by DoD. †CDR=Clinical Data Repository.

Table 5. DoD-to-VA Mediation Statistics for Allergy Reactions, Feb-June 2007.*

DoD-to-VA	**February**	**March**	**April**	**May**	**June**
Total DoD-to-SNOMED CT translation attempts	1,509	1,788	2,025	3,521	4,030
Translation failures (DoD-to-SNOMED CT)	306	467	432	432	107
Total allergy reactions sent to VA	1,203	1,321	1,593	3,089	3,923
Translation Success Rate: DoD-to-SNOMED CT	**80%**	**74%**	**79%**	**88%**	**97%**
Total allergy reactions received by VA	1,203	1,321	1,593	3,089	3,923
Translation failures (SNOMED CT-to-VA)	1	0	8	11	69
Total DoD allergy reactions sent to VA HDR†	1,202	1,321	1,585	3,078	3,854
Translation success rate: SNOMED CT-to-VA	**100%**	**100%**	**99%**	**100%**	**98%**
MEDIATION SUCCESS RATE: DoD-to-VA	***80%***	***74%***	***78%***	***87%***	***96%***

*Yellow areas designate translation services performed by VA. White areas designate translation services performed by DoD. †HDR=Health Data Repository.

Analysis of the causes of the mediation failures revealed the following issues, listed in order of frequency of occurrence:

1. SNOMED CT concept modeling issues were exposed. For example, a search for “nosebleed” in SNOMED CT’s CliniClue® browser returns more than one option within the “finding” hierarchy: “bleeding from nose” vs. “nosebleed/epistaxis symptom.” Another example of a modeling issue: the “Situation with Explicit Context” hierarchy was not addressed in the original VA/DoD mapping rules, as this hierarchy evolved within SNOMED CT after the initiation of the mapping.
2. New legitimate allergy reaction terms were added independently within each agency, which led to mediation failures in the time interval between synchronization and updating of each agency’s files.
3. Maintenance and versioning issues emerged when SNOMED CT released new versions with new concept statuses (e.g., “erroneous”, “limited”, “duplicate”, “ambiguous”) during the project. If agency updates were not synchronized, mediation failures would result.
4. Allergy reaction concepts and terms were sometimes deemed appropriate by one agency but not the other. For example, the concept “systemic disease” was used at one agency, but the other agency felt this term added no valuable information about an allergic reaction and did not include it in its list of selectable reactions for use by providers.
5. Divergent approaches to SNOMED mapping existed between VA and DoD, despite shared business rules. For instance, “hypertension” was mapped to “finding of increased blood pressure (finding)” at one agency, and to “Hypertensive disorder, systemic arterial (disorder)” at the other.

DISCUSSION

We begin with a list of lessons learned.

1. Mapping rules must always be tailored to the specific purpose of the mapping. These rules may be influenced by non-terminological issues, such as the potential for the entire message to fail if one component fails. We must recognize that mappings are often purpose- or use case-driven, as well as built by semantic nuances of context.

2. Even with established rules in place, there is a clear need for continued communication between agencies. We were unable to discern any major consistent reason for the mapping rule violations. One possibility is that VA and DoD initially used

different mapping tools. Another is that the process of finding the correct map for a term is variable and influenced by syntax and linguistic features of the search engine. In several cases, the Clue browser yields an apparently correct result (for example, a search for "orthostatic hypotension" returns "orthostatic hypotension (disorder)") but the term is located in the disorder hierarchy, rather than the findings hierarchy (to be used in preference if possible). It may not be immediately apparent that an alternative mapping exists ("postural drop in blood pressure (finding)") in another hierarchy. The clinical knowledge, background, and familiarity with SNOMED hierarchies and features of CliniClue® also are likely to influence search results. Ideally, a common team, process, and toolset would be used to produce the mapping. Perhaps the mapping could become a service of the Standards Development Organization, as is the case with RxNorm.

3. SNOMED CT modeling issues were probably the most difficult to address, as these require a sophisticated knowledge of concept modeling and of the evolution of SNOMED hierarchies over time.

4. Maintenance plans for using mediation terminologies need to include specific plans for synchronizing updates to the standard reference terminology, in this case SNOMED CT, and also for synchronizing updates to each agency's mapping file.

A significant outcome of this project is the generation of a new, unique SNOMED CT subset specific for Allergy Reactions (signs and symptoms) which could potentially be submitted for inclusion in SNOMED CT as an official subset. It could also be published and shared among federal agencies and non-federal partners.

In December 2007, HITSP designated the VA/Kaiser Permanente (KP) Problem List subset (16,430 entries) as the recommended standard for allergy reactions, a departure from previous CHI recommendations to use the VA/DoD Allergy Reactions subset (864 entries)[16]. While many of the VA/DoD Allergy Reactions terms are contained within the Problem List subset, use of the Problem List subset to record allergy reactions (signs and symptoms) may prove problematic, as is the case whenever data is used for a purpose other than that originally intended. Consider the terms "circumoral paresthesia (finding)" and "edema of pharynx (disorder)." These terms are appropriately found within the VA/DoD Allergy Reactions subset, but not within the VA/KP Problem List subset. The sheer size and complexity of the Problem List subset, compared to that of the Allergy Reactions subset, may unnecessarily complicate data entry for providers and result in unwanted entry of inappropriate terms as Allergy Reactions. The smaller subset could enable more precise data constraints and greater computing speed, without sacrificing data integrity. Communication with HITSP is ongoing regarding this issue. We propose that a new study be undertaken to evaluate the VA/KP Problem List and compare it to the VA/DoD Allergy Reaction subset, documenting content gaps, areas of overlap, and suitability for use as a mediation terminology.

In conclusion, we point out that the expense of mapping VA's and DoD's legacy terms (and maintenance of same) was relatively substantial—even for the limited list of Allergy Reactions. As CHDR expands to include more VA and DoD sites, the terminology maintenance requirements will continue.

Adopting the HITSP standard internally as a representation for allergies and reactions would be a more efficient method of working toward true semantic interoperability. Using a phased approach, legacy terms can be mapped to the standard, presented for adoption by the Standards Development Organization (SDO), and eventually migrated to the standard representation itself with deprecation of invalid legacy terms.

The use of mediation terminologies for computable data exchange is a dynamic and evolving process. It is prone to pitfalls, but is an effective, practical method for advancing the goal of semantic interoperability.

REFERENCES

1. Bush GW. Executive order: Promoting quality and efficient health care in federal government administered or sponsored health care programs [Internet]. Washington, D.C.: Office of the Press Secretary; 2006 Aug 22 [cited 2008 Jan 28]. Available from http://www.whitehouse.gov/news/releases/2006/08/20060822-2.html.

2. Mead CN. Data interchange standards in healthcare IT-Computable semantic interoperability now possible but still difficult, do we really need a better mousetrap? Journal of Healthcare Information Management 2006 Winter; 20(1): 71-78.

3. Ibid.

4. Ibid.

5. CHDR Training Home Page [Internet]. Brecksville, OH: VA Learning University Office of Information National Training and Education Office; last updated 2007 Jan 4 [cited 2008 Jan 28]. Available from: http://vaww.vistau.med.va.gov/VistaU/chdr/default.htm.

6. Insley M. Summary of VA/DoD sharing/interoperability initiatives. Dept of Veterans Affairs internal communication. 2007 May 17.

7. Parrish F., Do N, Bouhaddou O, Warnekar P. Implementation of RxNorm as a terminology mediation standard for exchanging pharmacy medication between federal agencies. AMIA Annu Symp Proc. 2006;2006:1057.

8. Warnekar P, Bouhaddou O, Parrish F, et.al.Use of RxNorm to exchange codified drug allergy information between the Department of Veterans Affairs and the Department of Defense (DoD). AMIA Annu Symp Proc. 2007;2007:781-785.

9. Mandel J. CHDR Fact Sheet: Terminology Mediation vs. Common Terminology. Dept of Veterans Affairs internal communication. 2007 Jan 12.

10. Ibid.

11. Bouhaddou, Omar. CHDR and Terminology Mediation Services. Dept of Veterans Affairs internal communication. 2006 Sep 7.

12. Bouhaddou O, Warnekar P, Parrish F, et al. Exchange of Computable Patient Data Between the Department of Veterans Affairs (VA) and the Department of Defense (DoD): Terminology Mediation Strategy. Journal of the American Medical Informatics Association (JAMIA) Vol. 15 No. 2, Mar/Apr 2008.

13. Ibid.

14. Ibid.

15. Ibid.

16. ANSI Public Document Library [Internet]. Washington, D.C.: American National Standards Institute. HITSP Summary Documents Using HL7 Continuity of Care Document (CCD) Component, v2.1; 2007 Dec 13 [cited 10 Mar 2008]; p. 49. Available from: http://publicaa.ansi.org/sites/apdl/Documents/Standards%20Activities/Healthcare%20Informatics%20Technology%20Standards%20Panel/Interoperability%20Specification/IS-Released%20for%20Imp.%20and%20Recognized/IS03%20-%20Consumer%20Empowerment/HITSP_V2.1_2007_C32%20-%20Summary%20Documents%20Using%20CCD.pdf.

Using SNOMED-CT For Translational Genomics Data Integration

Joel Dudley[1-3], David P. Chen[1-3], Atul J. Butte[1-3], M.D., Ph.D.,
[1]Stanford Center for Biomedical Informatics Research, Department of Medicine,
[2]Department of Pediatrics, Stanford University School of Medicine, Stanford, CA/USA
[3]Lucile Packard Children's Hospital, Palo Alto, CA/USA
{jdudley,dpchen,abutte}@stanford.edu

As industrial, governmental, and academic agencies place increasing emphasis on translational research, biomedical researchers are now faced with entirely new challenges in regards to both biomedical data integration and knowledge discovery. There is now both a strong need and a tremendous opportunity to apply translational bioinformatics to address the fundamental challenges in integrating the vast bodies of -omics and clinical data. Here we report on our preliminary work in utilizing SNOMED-CT as both a tool for translational data discovery, and a major component in a framework for the large-scale integration of gene expression microarray data and clinical laboratory data. Annotations from microarray experiments in NCBI GEO were mapped to SNOMED-CT terms using UMLS, and these mappings were joined to clinical laboratory data using ICD9CM to SNOMED-CT mappings within UMLS. We find that microarray experiments characterizing 211 distinct diseases can be mapped to clinical laboratory data measurements for 13,452 distinct patients. We maintain that this work represents critical first steps in providing a foundation for large-scale translational data integration, and underlines the important role that controlled clinical terminologies, such as SNOMED-CT, can play in addressing such problems.

INTRODUCTION

Our ability to generate high-quality biomolecular data has advanced at considerably faster rate than our ability to investigate the data generated. This imbalance, driven primarily by rapid advances in high-throughput biological data acquisition technologies and plummeting per-experiment costs, has created an entire spectrum of informatics challenges that are, in many instances, as intangible and complex as the fundamental biological questions that these technologies were designed to address. As a consequence, our ability to formulate and investigate important biological and medical questions is currently limited by our ability to manage and integrate the profusion of biomedical data.

Problems in data integration are moving towards the forefront of biomedical research, driven foremost by the sheer diversity of measurement technologies now available, and the tremendous volumes of such measurements finding their way into the public domain. The situation is further complicated by the fact that the majority of the public biomolecular data is annotated using unstructured free-text, making it difficult to discern the various biological and medical contexts of the data in an automated fashion. In previous work we demonstrated the feasibility of using controlled terminologies and straightforward text-mining techniques to elucidate clinical, environmental, and phenotypic contexts from free-text annotations associated with public microarray data[1, 2]. The establishment of experimental context is critical to linking genes to environment, phenotype, and ultimately medicine.

While most major types of biomolecular data can be found in the public domain, it is traditionally difficult for researchers to gain access to clinical data. This is unfortunate as the data generated on a daily basis by hospitals and clinicians is perhaps the richest source of phenotypic biomarker data currently available. Fortunately modern Electronic Health Record (EHR) systems such as the Stanford Translational Research Integrated Database Environment (STRIDE)[3] and the University of Virginia Health System Clinical Data Repository (CDR)[4] grant institutional researchers access to large volumes of de-identified, quantitative clinical data in digital form. In recent work, we demonstrated the utility in applying bioinformatics methods to quantitative clinical data to draw new inferences about disease severity[5], and elucidate novel biomarkers[6].

Genome Wide Association studies have revealed that for many complex diseases, the pathogenesis of the disease may be facilitated by relatively minor changes across a large number of genes interacting through as of yet poorly understood mechanisms[7]. These findings have therefore highlighted the importance of linking biomolecular data with phenotypic quantifications in order to uncover the full complexity of disease etiology. Recent work in integrating these two data types has offered new insights into disease etiology and pathology with direct clinical implications. Segal and colleagues correlated imaging traits from computed tomography (CT) images of liver cancers with gene expression data to reconstruct global expression signatures in cancer tumors that are linked to diagnosis, prognosis and treatment[8]. A number of studies have demonstrated the utility of patient microarrays in identifying gene expression patterns linked to disease diagnosis[9], subtypes[10, 11], outcome[12], and treatment[13, 14]. As significant as the aforementioned findings are, their underlying methods are limited by the fact that, in all instances, they require that the biomolecular and clinical data be derived from the same patient. Given the current high costs and logistical complexities involved in acquiring patient data in a clinical setting, it would be prohibitively expensive to scale the same approaches to address the broad spectrum of human disease. Furthermore, such an approach implicitly eschews the great wealth of public biomolecular data readily available.

A major problem in integrating clinical and biomolecular data derived from disparate sources is to identify attributes by which they can be appropriately joined. This task is complicated by the fact that the majority of biomolecular data is annotated around the concepts of genes and gene products, whereas clinical data is centered on the concept of a patient. We find one concept shared among both clinical data and vast amounts of biomolecular data, and that is the concept of a *disease*. Therefore it is possible to integrate anonymous biomolecular data characterizing an aspect of a particular disease state with quantitative clinical data derived from patients being treated for the same disease.

Central to this approach is the need for a comprehensive controlled disease terminology through which the biomedical and clinical data is joined in a systematic fashion. In general, we would want this disease terminology to maximize three primary criteria: coverage, defined by the number of unique disease terms defined; expressiveness, which is the richness of relationships between disease terms; and resolution, which is the level of detail offered by the terminology structure. A deficiency in any of these could negatively impact the amount and diversity of data that could be integrated, and potentially limit the types of analyses that can be performed on the data downstream. There are a number of well-established disease terminologies in active use that satisfy the above criteria to varying degrees. Chief among these are the International Classification of Diseases (ICD), Medical Subject Headings (MeSH), and the Systemized Nomenclature of Medicine-Clinical Term (SNOMED-CT). Each of these is suited for data integration, yet each of them present particular pros and cons.

The ICD terminology, evolved from a lineage that spans more than 100 years, is the most widely utilized disease terminology, with widespread adoption among a large number of major healthcare providers, the U.S. Federal Government, as well as the World Health Organization. Consequently, the majority of clinical data is codified using ICD codes. Unfortunately the ICD is poorly suited for data integration as the approximately 14,000 unique terms codified by ICD is quite small compared to other terminologies. Furthermore, the ICD is more a compendium of diagnosis and procedure codes, as it lacks any significant hierarchical or relational structure.

MeSH, which is used primarily for the purpose of indexing publications, is only slightly larger than ICD in terms of size with more than 22,000 unique terms. However, the design of MeSH is much more structured and diverse compared to ICD. MeSH terms are arranged into a hierarchy of 14 distinct top-level categories that organize terms by Anatomy, Disease, Chemicals and Drugs, and Geography among other things. MeSH also contains a set of qualifier terms that can be used to narrow the specificity of a descriptor term (e.g. "Measles/epidemiology"). While MeSH possesses many of the attributes desirable for translational data integration, its attributes modest in comparison to those of SNOMED-CT.

SNOMED-CT was born from a medical terminology lineage that traces back more than 75 years, and is currently in use by pathologists worldwide to perform precise classifications of human disease[15, 16]. With more than 340,000 unique biomedical concepts organized into 19 relational hierarchies linked by more than 1.3 million relationships, it is by far the most expansive and expressive disease terminology in existence. The sheer number of concepts coupled with the rich relational architecture in SNOMED-CT offers attributes superior to other disease terminologies. For example, SNOMED-CT establishes that a *clear cell carcinoma of the kidney* is both a *malignant tumor of the kidney* and *a malignant tumor of the retroperitoneum*. The ICD version 9 (ICD-9) simply asserts that a *malignant neoplasm of the kidney* is a *malignant neoplasm of the genitourinary organs*, which is a much coarser designation. Therefore assert that SNOMED-CT is currently the best-suited terminology for integrating biomolecular and clinical data by disease.

In this study we investigate the feasibility of using SNOMED-CT to integrate gene expression data from a public microarray repository with de-identified clinical laboratory data obtained from a hospital EHR system by disease. We propose that SNOMED-CT is well suited for this approach as it is the largest disease vocabulary currently available. We evaluate the effectiveness of this approach based on the extent of data successfully joined.

METHODS

A high level representation of the data integration approach is detailed in figure 1. The microarray experiment data was obtained from the NCBI GEO FTP site (downloaded 11/27/2007), which was parsed into a relational structure and stored in a MySQL database. The de-identified clinical laboratory data was obtained from the Lucile Packard Children's hospital via STRIDE as delimited text files. UMLS release 2007 AA was used as the vocabulary source. The integration steps were performed as follows.

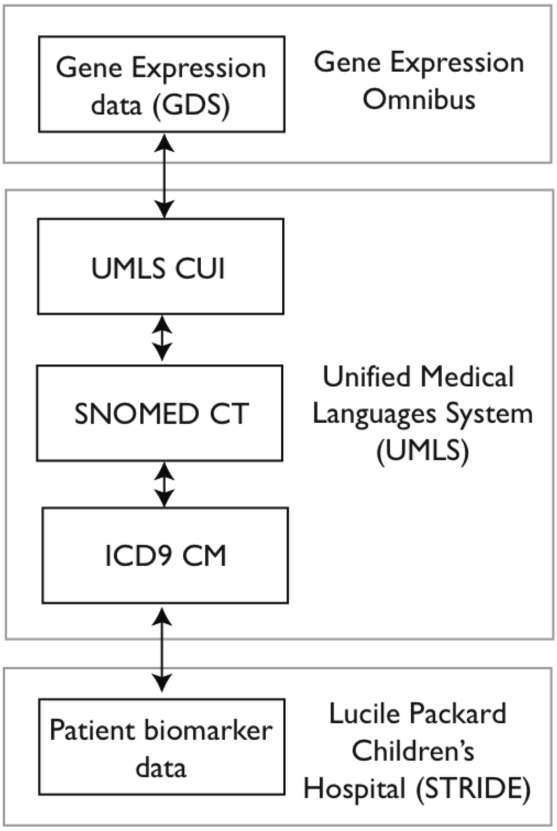
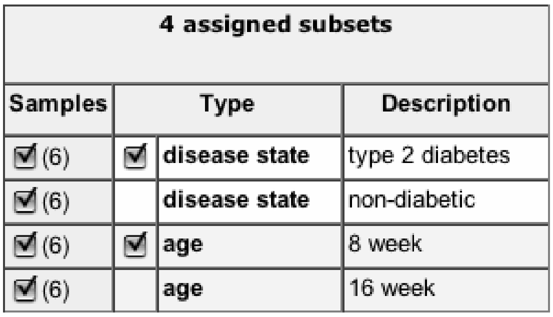

Gene Expression data (GDS)
Gene Expression Omnibus
UMLS CUI
SNOMED CT
Unified Medical Languages System (UMLS)
ICD9 CM
Patient biomarker data
Lucile Packard Children's Hospital (STRIDE)

Figure 1 – Schematic representation of the approached used to join gene expression data with clinical laboratory data. Annotations from GDS are first mapped to UMLS CUIs that map to at least one SNOMED CT term, and the ICD9 CM codes from the patient records are mapped to SNOMED CT terms using the relational architecture of UMLS.

Mapping microarray experiments to diseases

Clinically relevant microarray data was identified using a previously described method[17]. In brief, we queried the NCBI Gene Expression Omnibus (GEO)[18] to obtain all GEO DataSet experiments with associated PubMed identifiers. For each PubMed identifier we obtained the associated MeSH headings using NCBI eUtils. Each of the MeSH headings was mapped to a UMLS CUI using the MRCONSO table. Using the MRSTY table, we obtained the semantic type identifier (TUI) for the mapped CUIs, and if any MeSH term is found to have a semantic type among Injury or Poisoning (T037), Pathologic Function (T046), Disease or Syndrome (T047), Mental or Behavioral Dysfunction (T048), Experimental Model of Disease (T050), or Neoplastic Process (T191) then the associated experiment is determined to be disease-associated and therefore clinically relevant. This resulted in the positive identification of 737 disease-associated experiments.

The disease-associated experiments are investigated by a second previously described text-mining technique that examines GEO DataSet (GDS) subset annotations to identify when a disease state is being compared to a normal control state[2]. GDS are higher-level representations of microarray experiment in which samples are organized into biologically informative collections known as subsets. The subsets are representative of the experimental axis under examination (figure 2). An attempt is made to map the free-text annotations associated with the GDS subsets to SNOMED-CT disease terms using UMLS concepts. These mappings are subsequently manually reviewed for accuracy, where erroneous codifications are corrected if found.



4 assigned subsets

Samples	Type		Description
☑ (6)	☑	disease state	type 2 diabetes
☑ (6)		disease state	non-diabetic
☑ (6)	☑	age	8 week
☑ (6)		age	16 week

Figure 2 – Example of microarray data subsets defined by GEO GDS experiments.

Mapping patient laboratory data to diseases

Clinical laboratory data for pediatric patients from the Lucile Packard Children's Hospital was obtained digitally from the STRIDE system. All of the laboratory measurements were received pre-encoded with ICD-9 codes. These ICD-9 codes were mapped to SNOMED-CT codes by first querying UMLS to find the CUI identifier associated with the ICD-9 code. We then took advantage of the inter-terminology mappings provided by the UMLS (MRMAP) table to translate the ICD-9 codes into SNOMED-CT concepts using associated CUIs.

Joining the microarray and patient lab data by disease

The GDS subsets with mappings to SNOMED-CT disease CUIs were joined with the clinical laboratory data using the UMLS CUIs derived from mapping the ICD-9 codes to SNOMED-CT terms using the UMLS MRMAP table. Of the 238 unique disease concepts mapped to the microarray data, 90% were mapped to quantitative clinical laboratory data for at least one patient.

RESULTS

Using automated methods, were able to identify 737 GDS microarray experiments in NCBI GEO related to human disease. The GDS subsets were investigated for terms related to UMLS concepts that were linked to a SNOMED-CT disease term, resulting in the identification of 238 unique human disease concepts. In total, 29,451 microarray samples were codified with SNOMED-CT disease identifiers. Note however that method was restricted to include only those GDS for which a disease and normal control subset could be identified. This restriction ensures that a disease vs. normal vector of change can be extracted from the data to establish a baseline disease expression signature for downstream analysis.

Disease	SNOMED Terms	ICD9CM Terms	Ind
Allergic asthma	1	1	2240
Asthma	1	1	2240
Allergic asthma NEC	1	1	2240
Esophageal Reflux	1	1	1895
H. pylori infection	1	2	1322
Colitis	1	1	1299
Primary Hypertension	1	1	1017
Hypertension	1	1	1017
Obesity	2	1	1010
Type 1 diabetes	1	1	843

Table 1 – Top ten data mappings ordered by the number of patient lab records matched.

We retrieved quantitative clinical laboratory data representing diagnostic biomarkers for 49,414 patients across 9,997 distinct diagnosis codes. These codes mapped to 20,049 distinct UMLS CUIs. It is interesting to note that in mapping ICD to UMLS we find that twice as many UMLS concepts as ICD-9 terms are found. This likely resulted from the fact that ICD-9 is generally a more high-level terminology, and therefore terms related to rare genetic disorders, for example, may only be represented by one ICD-9 code, whereas UMLS may allow for more fine-grained attribution of specific rare genetic disorders.

In joining the ICD-9 disease codes from the clinical laboratory data to the microarray data using SNOMED-CT disease codes, we find that 211 of the unique disease concepts annotating the microarray data can be mapped to clinical laboratory data. In total, clinical laboratory data for 13, 452 patients was mapped to SNOMED-CT disease codes that were used to annotate the microarray GDS experiments. Table 1 shows the top diseases by the number of patients mapped.

Disease	SNOMED Terms	ICD9CM Terms	Ind
Follicular lymphoma	4	3	136
Hamman-Rich syndrome	4	2	18
Mycobacterial infection	3	2	26
Mixed hyperlipidemia	3	2	90
Hepatoma	3	2	67
Fetal alcohol syndrome	3	1	10
Diabetic nephropathy	3	2	30
Megakaryocytic leukemia	2	2	125
Acute monocytic leukemia	2	1	7
Status epilepticus	2	1	84

Table 2 – Top ten data mappings sorted by the number of SNOMED-CT terms matched.

As evident from the data listed in table 1, there are cases in which distinct SNOMED-CT terms will map to the same ICD-9 term. To explore the ambiguities of mapping terms between the SNOMED-CT and ICD-9 using CUIs, we investigated the overall pattern of the mapping cardinalities. Table 2 shows cases in which a single UMLS CUI maps to multiple

SNOMED-CT terms. This could indicate that there is some degree of ambiguity in the SNOMED-CT to ICD-9 UMLS mappings, and perhaps a dampening of SNOMED-CT term resolution when using UMLS concepts.

To better understand the influence of UMLS CUI definitions with regards to source identifier consolidation, we calculated summary statistics for several terminologies with UMLS and restricted the results to CUIs representing a disease. The summary statistics are listed in table 3.

Source	**Total disease concepts**	**Identifiers per concept**
SNOMED-CT	74,611	1.4
ICD-9-CM	12,631	1.1
NCI	12,257	1.0
MeSH	6,613	1.0

Table 3 – Summary statistics for select disease terminologies sorted by total number of disease concepts (CUI).

DISCUSSION

The profusion of large public data repositories of genome-scale measures, coupled with the pressing imperative to translate such data into medicine, has precipitated the need to develop informatics tools and techniques for integrating disparate forms of biomolecular and clinical data. The purpose of this investigation was to explore the feasibility of using SNOMED-CT for such integrative efforts. We assessed the feasibility of SNOMED-CT as a translational joining factor by using it to integrate anonymous gene expression data from a public microarray repository with de-identified clinical laboratory data by disease.

We find that SNOMED-CT is effective as a disease terminology for integrating these two types of biomolecular and clinical data. The cases in which microarray data could not be mapped to clinical laboratory data largely reflect the fact that only pediatric data was used. The unmapped terms contain diseases such as *Parkinson's disease*, *macular degeneration*, *Alzheimer's disease* and other diseases not generally found in children. Other failed mappings represent relatively rare disorders, such as *Yersiniosis* and *Luteoma*. Better mappings might be obtained by leveraging the relational structure of UMLS to map terms that are parent or child relationships to the disease terms.

The many-to-many and many-to-one SNOMED-CT to ICD-9 mappings using UMLS CUIs do present an interesting problem. These could lead to ambiguities in the mappings such that a highly specific disease variant is mapped to a more generalized disease category. This could have a negative impact on the downstream utilization of the integrated data. The data in table 3 suggests that large source vocabularies like SNOMED-CT have been constrained and compressed by the smaller vocabularies within UMLS to the degree that original source vocabulary resolution is lost. This may suggest and alternative strategy in which the biomolecular samples are labeled only with SNOMED-CT identifiers and the translation between SNOMED-CT and ICD-9 is performed outside of UMLS CUI constraints.

There are several caveats in the interpretation of the results. First off, the data sets were not generalized in that the clinical laboratory data only represented pediatric patients and the microarray experiments were limited to those in which a disease and a normal control distinction was evident. Furthermore, this study offered only a focus on SNOMED-CT and did not apply the same techniques to the alternative disease terminologies mentioned to offer any quantitative comparison. Although the investigation revealed that SNOMED-CT was capable of joining the two data types, it offers no statistical characterization of the joining to assess its overall quality and reliability. Of course we also acknowledge that the text mining aspects of this approach are prone to errors, such as miscodings of the data.

The results demonstrate that current and future translational data integration endeavors can leverage existing clinical terminologies, such as SNOMED-CT, to integrate clinical and biomolecular data types and shift valuable efforts to downstream discovery. Furthermore, this study provides support for the continued development and use of SNOMED-CT for translational data integration, and brings to light the importance inter-terminology mappings resources such as UMLS. As demonstrated by our own work, and the work of others, the straightforward act of integrating data from the molecular and clinical worlds can have profound and direct impact on human health.

Although our initial work focused on the integration of microarray data and patient lab data specifically, we are now working to expand the application of the underlying system to integrate additional data types. In order to integrate new forms of biomolecular data into our current framework we must develop improved text-mining methods to map the underlying experimental data to SNOMED-CT identifiers. From the clinical perspective we will continue to integrate new data obtained from the STRIDE system and look to incorporate additional clinical data types as well.

We must also develop methods to test and improve the reliability of the clinical data, as hospital workers will inevitably miscode a small percentage of the data. We must also account for the fact that the application of clinical codes is subject to a number of non-scientific influences, such as hospital billing policies, insurance companies, and pharmaceutical regulations. Any future work in this area should also entail the development of statistical metrics to evaluate the joining terminology, such that a principled decision can be made to identify the most appropriate terminology for a particular integration scenario.

ACKNOWLEDGEMENTS

This work was supported in part by the Lucile Packard Foundation for Children's Health, National Library of Medicine (K22 LM008261), National Institute of General Medical Sciences (R01 GM079719), National Human Genome Research Institute (P50 HG003389), Howard Hughes Medical Institute, and the Pharmaceutical Research and Manufacturers of America Foundation. The authors would also like to thank Alex Skrenchuck for High Performance Computing support.

REFERENCES

1. Butte AJ, Kohane IS. Creation and implications of a phenome-genome network. Nature biotechnology. 2006 Jan;24(1):55-62.
2. Dudley J, Butte AJ. Enabling Integrative Genomic Analysis of High-Impact Human Diseases Through Text Mining. Pacific Symposium on Biocomputing. 2008.
3. STRIDE. [http://stride.stanford.edu/STRIDE/]
4. CDR. [https://cdr.virginia.edu/]
5. Chen DP, Weber SC, Constantinou PS, Ferris TA, Lowe HJ, Butte AJ. Clinical Arrays of Laboratory Measures, or "Clinarrays", Built from an Electronic Health Record Enable Disease Subtyping by Severity. AMIA Annual Symposium Proceedings. 2007.
6. Chen DP, Weber SC, Constantinou PS, Ferris TA, Lowe HJ, Butte AJ. Novel Integration of Hospital Electronic Medical Records and Gene Expression Measurements to Identify Genetic Markers of Maturation. Pacific Symposium on Biocomputing. 2008.
7. Pickrell J, Clerget-Darpoux F, Bourgain C. Power of genome-wide association studies in the presence of interacting loci. Genetic epidemiology. 2007 Nov;31(7):748-62.
8. Segal E, Sirlin CB, Ooi C, Adler AS, Gollub J, Chen X, et al. Decoding global gene expression programs in liver cancer by noninvasive imaging. Nature biotechnology. 2007 Jun;25(6):675-80.
9. Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, Angelo M, et al. Multiclass cancer diagnosis using tumor gene expression signatures. Proceedings of the National Academy of Sciences of the United States of America. 2001 Dec 18;98(26):15149-54.
10. Pandita A, Zielenska M, Thorner P, Bayani J, Godbout R, Greenberg M, et al. Application of comparative genomic hybridization, spectral karyotyping, and microarray analysis in the identification of subtype-specific patterns of genomic changes in rhabdomyosarcoma. Neoplasia (New York, NY. 1999 Aug;1(3):262-75.
11. Lapointe J, Li C, Higgins JP, van de Rijn M, Bair E, Montgomery K, et al. Gene expression profiling identifies clinically relevant subtypes of prostate cancer. Proceedings of the National Academy of Sciences of the United States of America. 2004 Jan 20;101(3):811-6.
12. Chen HY, Yu SL, Chen CH, Chang GC, Chen CY, Yuan A, et al. A five-gene signature and clinical outcome in non-small-cell lung cancer. The New England journal of medicine. 2007 Jan 4;356(1):11-20.
13. Potti A, Dressman HK, Bild A, Riedel RF, Chan G, Sayer R, et al. Genomic signatures to guide the use of chemotherapeutics. Nature medicine. 2006 Nov;12(11):1294-300.
14. Komatsu M, Hiyama K, Tanimoto K, Yunokawa M, Otani K, Ohtaki M, et al. Prediction of individual response to platinum/paclitaxel combination using novel marker genes in ovarian cancers. Molecular cancer therapeutics. 2006 Mar;5(3):767-75.
15. SNOMED Intl. [http://www.snomed.org]
16. Chute CG. Clinical classification and terminology: some history and current observations. J Am Med Inform Assoc. 2000 May-Jun;7(3):298-303.
17. Butte AJ, Chen R. Finding disease-related genomic experiments within an international repository: first steps in translational bioinformatics. AMIA Annual Symposium proceedings / AMIA Symposium. 2006:106-10.
18. Barrett T, Suzek TO, Troup DB, Wilhite SE, Ngau WC, Ledoux P, et al. NCBI GEO: mining millions of expression profiles--database and tools. Nucleic acids research. 2005 Jan 1;33(Database issue):D562-6.

A Methodology for Encoding Problem Lists with SNOMED CT in General Practice

Francis Lau, Ph.D., Ray Simkus, M.D., Dennis Lee, M.Sc.
School of Health Information Science, University of Victoria, Victoria, B.C., Canada
fylau@uvic.ca, ray@wmt.ca, dlhk@uvic.ca

ABSTRACT

This paper describes a methodology for encoding problem lists used in general practice with SNOMED CT. Our intent is to help general practitioners to incorporate SNOMED CT into their existing Electronic Medical Record (EMR) systems with minimal disruption as a first step, thus allowing them to assess its impact prior to full-scale conversion. We started with 1,713 original unique terms that made up the problem lists from the general practice EMR used in the study. We ended with 1,468 unique concepts after two cycles of matching and revisions that led to 1,347 or ~92% successful matches. The remaining terms were revised to tease out modifiers or secondary concepts that could be used to provide equivalency through post-coordination. While skeptics of reference terminology systems often balk at their unwieldy size and complexity for local adoption, this study has demonstrated that, using our methodology, it is possible to create a manageable subset of SNOMED concepts for problem lists used in general practice with immediate tangible value.

INTRODUCTION

The problem list is the keystone of the medical record. In general practice settings, the type of problems presented by patients can be quite diverse. Examples range from non-specific symptoms such as headaches with unknown cause, to a diagnosis of coronary disease that can be expressed in different ways such as heart attack and myocardial infarction. The choice of terms used in problem lists becomes an important design issue for the electronic medical record (EMR), since the level of granularity selected for defining the problems and the actual terms entered into the system can affect one's ability to retrieve the information afterwards, thus impacting the overall quality of the EMR system.

There have been many studies on the design and use of controlled terminology to encode the problem lists in EMR systems and their impact on practice [1-8]. Most of these studies are focused on large institutions involving a substantive number of clinical terms in order to accommodate the needs of a wide range of clinicians in the institution. For example in their study of diagnosis and problem lists in a computerized physician order entry system, Wasserman [9] reported that 88.4% of their 8,378 clinical terms were found in SNOMED CT. With the addition of 145 site-specific terms they were able to achieve 98.5% overall content coverage. With the formation of the International Health Terminology Standards Organization (IHTSDO), the historical barriers to SNOMED CT related to cost and the proprietary nature of the product have now been removed, and national initiatives related to EMR's are emerging to use SNOMED CT as a clinical terminology in several countries around the world.

Despite such impressive development, the effort to adopt SNOMED CT in Canada has been minimal to date. There continues to be a concern especially in the primary care setting where most general practices are made up of small groups of practitioners, of whom few are equipped with an EMR. Critics often balk at the enormous size and complexity of SNOMED CT, considering it as too unwieldy and costly for local adoption and use. But a review of data collected from several sites by one author showed the number of codes needed to cover disorders of at least 1:100,000 occurrence would be under 5,000 [10]. Work is underway with IHTSDO and the WICC group of WONCA to finalize this list as a potential primary care SNOMED subset [11].

In this paper, we describe a methodology that we have developed based on an ongoing study to encode problem lists using SNOMED CT (July 2007 release) for a local general practice in Canada. The intent of this methodology is to enable general practitioners to incorporate SNOMED CT into their existing EMR systems within minimal disruption as a first step, thus allowing them to assess its potential impact prior to full-scale conversion.

METHODS

Design and Setting

For this study, we included all the problem list (PL) terms from the commercial EMR system used by a local general practice in British Columbia, Canada. This setting is typical of many general practices across the country, which are made up of small groups of general practitioners working in a private medical office, mostly on a fee-for-service basis. The medical office in this study has four general practitioners who have worked as a group for 30 years in a township with a population of 100,000 located east of Vancouver, British Columbia. The practice has had 8 years of experience using an EMR. At least two of the practitioners record all of the information on their patients on a daily basis at the time of encounter or shortly thereafter. Laboratory and imaging results and consult reports from external sources – both electronic and on paper – are entered into the EMR either by the practitioners themselves or the medical office assistant.

Matching Algorithms

We applied four matching algorithms used in an earlier SNOMED CT to ICD-10 mapping project to find matching SNOMED concepts for each of the PL terms [12]. Three are lexical techniques for exact-match, match-all and partial-match. The fourth is semantic matching that involves retrieving the current concepts based on historical relationships if the initial SNOMED concepts found were inactive. These algorithms are summarized in Table 1.

Algorithm	Explanation
1. Exact match	Exact string match where all words are same and in same sequence, including punctuation
2. Match all	String match where all words are same but not necessary in same order; additional words allowed
3. Partial match	String match where one or more words is found
4. Semantic match	For inactive concepts use historical relationships Was-A, Same-As, May-Be-A, Replaced-By to find current concepts
5. Unmatched	Assigned when no match is found

Table 1. Matching algorithms used in this study

Normalization Steps

In addition to applying the matching algorithms to the original PL terms, we reran the algorithms after we normalized the PL and SNOMED terms to remove "noise" using the Unified Medical Language System (UMLS 2007 version) normalization steps, shown in Table 2a [13,14]. To improve matching, we expanded step-2 to remove both "stop words" and "exclude words" and SNOMED prefixes, shown in Table 2b. For step-5 we included the lookup and stemming methods to uninflect the phrase. The lookup method uses the UMLS SPECIALIST Lexicon's inflection table with ~1 million entries, whereas the stemming method is a computational technique that reduces word variants to a single canonical form [15,16].

No	Step	Example
1	Remove genitive	Hodgkin***'s*** disease, NOS → Hodgkin diseases, NOS
2	Remove stop words	Hodgkin diseases, ***NOS*** → Hodgkin diseases,
3	Convert to lowercase	***H***odgkin diseases, → hodgkin diseases,
4	Strip punctuation	hodgkin diseases**,** → hodgkin diseases
5	Uninflect phrase	hodgkin disease***s*** → hodgkin disease
6	Sort words	**hodgkin** **disease** → disease hodgkin

Table 2a. UMLS normalization steps [8, slide20]

Matching PL Terms

The process of matching the PL terms involved cycling through the matching algorithms one at a time to find the best candidate SNOMED CT concepts. For each algorithm we always began with the original terms, then the UMLS normalized terms, followed by the stemmed terms. During each cycle, we would review the candidate concepts found to determine if it was a match, and if so, what type of match it was based on the algorithm applied. When no matching concepts were found, we would label the term as unmatched. Our experience with the matching algorithms had been that, the sooner we could find a match in the cycle, the greater confidence we would have that the candidate concept is appropriate. The preferred order of matching selected is always exact first, then all, followed by partial. For exact-match and match-all if only inactive concepts are found then a semantic-match is done to find their corresponding current concepts through the historical relationships.

Step-5	Explanation
Stop words	Frequent short words that do not affect the phrase: and, by, for, in, of, on, the, to, with, no, (nos)
Exclude words	Words that may change meaning of the word but if ignored help to find a term otherwise missed: about, alongside, an, anything, around, as, at, because, before, being, both, cannot, chronically, consists, covered, does, during, every, find, from, instead, into, more, must, no, not, only, or, properly, side, sided, some, something, specific, than, that, things, this, throughout, up, using, usually, when, while
SNOMED Prefixes	[X] – concepts with ICD-10 codes not in ICD-9 [D] - concepts in ICD-9 XVI and ICD-10 SVII [M] – morphology of neoplasm concepts in ICD-O [SO] – concepts in OPCS-4 chapter Z in CTV3 [Q] – temporary qualifying terms from CTV3 [V] – concepts in ICD-9 and ICD-10 on factors influencing health status and contact with health services (V-codes and Z-codes)

Table 2b. Expanded UMLS normalization step-2

Encoding the Problem Lists

The process of encoding the problem lists extracted from the EMR followed these steps: (a) tabulating the frequency of occurrences for all of the original PL terms; (b) cataloguing all of the unique words across the PL terms present; (c) examining all unique words and PL terms to identify and revise for acronyms, abbreviations, spelling variants and errors; (d) matching the PL terms to SNOMED CT concepts using matching algorithms described earlier; (e) producing detailed and summary outputs to show the type of matches found; (f) reviewing/verifying the matched concepts one term at a time for accuracy; (g) repeating steps (c-f) until no further matches could be found; (h) examine remaining partial-matches for post-coordination; (i) create an index table of all PL and matched SNOMED terms. As part of this study, we also explored navigating within the SNOMED hierarchy to examine how the super-types and relations could be used to improve the quality of recall using the matched SNOMED concepts.

RESULTS

Summary of PL Terms and Matches

A total of 7,833 PL entries were extracted from the EMR for this study. The majority of these entries were recorded by one practitioner over a 7-year period. Of these entries, there were 1,713 unique PL terms present. Based on the frequency distribution of the entries, the top 10 PL terms were hypertension, hypercholesterolemia, diabetes mellitus, hypothyroid, asthma, atrial fibrillation, gastroesophageal reflux, depression, congestive heart failure and chronic kidney disease. After the second cycle we had 1,296 (88.23%) exact-matches where the PL terms are exactly the same as the SNOMED terms found. There were 51 (3.47%) match-all where all the words in the PL terms are present in the SNOMED terms but not necessarily in the same sequence. There were 120 (8.17%) partial-matches where one or more words matched the SNOMED terms. Another 20 (1.42%) SNOMED terms were found with semantic matches. Between the two cycles partially-matched terms were revised to tease out qualifiers and secondary concepts if present in order to explore post-coordination. A summary of the PL terms and the SNOMED matches found is shown in Table 3.

Characteristics of Encoded PL Terms

In Table 4 we have examples of the frequently used PL terms with their SNOMED terms found by exact, all and semantic matches. Also shown are the matches after revision and post-coordination of the original and partially-matched PL terms. For most exact-matches we selected the preferred terms from SNOMED CT as they are identical or closest to the original PL terms, such as Atrial fibrillation. In some cases we chose the synonym terms, such as Hypertension instead of the preferred term which is Hypertensive disorder. For match-all and some partial-matches we selected the SNOMED terms that were closest to the PL concept involved, such as GERD gastro-esophageal reflux disease. For semantic matches we looked up the current concepts of the matched but inactive SNOMED terms through their historical relationships, such as Cirrhosis. For post-coordination we added qualifier and refinement terms to SNOMED concepts or combined those that are lexically closest to the original PL terms, such as Atrial fibrillation+Chronic, Kidney disease+Chronic, and Headache+Migraine. After the second cycle any remaining partial-matches were treated as unmatched. Initially there were eight PL terms not found in SNOMED CT. Five were spelling errors and were revised for the second cycle (e.g. hepatomegal*l*y → hepatomegaly); three were legitimate missing terms – vasculopath, pyocystitis and hypotestosteronemia, where we had to modify the PL term or tag as local extensions. Using these outputs we created an index table to link the PL terms to their matched SNOMED terms, shown in Table 5. Each row contains the PL-termId, conceptId, descriptionId, relationship-typeId match-type, and post-coordination-sequenceId.

Description	**Frequency**	
No. of patients	2,894	
Total PL entries	7,833	
Total words in PL terms	16,455	
Unique words	1,764	
Longest word	Hypercholesterolemia, 20 characters	
Median length	8 characters	
Most common word	Hypertension, 585 times	
Matching Algorithm	**Initial Cycle Frequency (%)**	**2nd Cycle Frequency (%)**
Exact-Match	905 (52.83%)	1,296 (88.23%)
Match-All	167 (9.75%)	52 (3.47%)
Partial-Match	633 (36.95%)	120 (8.17%)
Semantic-Match	49 (2.86%)	20 (1.42%)
Unmatched	8 (0.47%)	2 (0.14%)
Post-coordination	Not done	In-progress
Total unique PL terms	**1,713**	**1,468**

Table 3. Summary of PL terms and matches. For frequency %, once a match has been found it is not included as part of the next matching algorithm

Revision of PL Terms

Manual revisions were done on the 1,713 unique PL terms after the initial cycle. By selecting the PL terms that were not matched in SNOMED CT, we were able to identify entries that were misspelled, idiosyncratic local terms or ambiguous concepts. A number of spelling mistakes were corrected. The CliniClue Browser [17] was used to find matches for each term. A few terms were found in our problem lists but not in SNOMED CT. Some were local terms that needed to be reconsidered but there were also terms that would be submitted for inclusion in SNOMED CT. One example is "chronic kidney disease" which seems to be the preferred term in common usage. Yet the closest SNOMED term is "chronic renal failure." In this revision we also noted parts of some PL terms could be removed as qualifiers or modifiers, thus increasing the number of exact matches found. Examples include left, right, lower, midline, chronic, recurring, active, query and multiple. These modifiers seemed to be clustered around the concepts of time course, number, location and severity. We found 313 such instances in our PL terms. In another 89 instances we found post-coordination of two SNOMED concepts produced a good match.

Navigating the SNOMED Hierarchy

As part of this study, we explored ways to navigate the SNOMED hierarchy to determine if it could improve one's ability to retrieve related concepts. Of the 1,296 exact matches found for the 1,468 unique PL terms present, we selected a subset of 32 PL terms related to cardiovascular disorders for this analysis. First, we did frequency counts of these PL terms to show how often they were present in the EMR system. For each PL term present, we navigated up the hierarchy until we reached the super-type "49601007|Disorder of cardiovascular system." We then pruned the tree to include only those concepts with a positive frequency count, but left their immediate super-types intact. This partly-instantiated cardiovascular disorder hierarchy is shown in Figure 1. The value of this tree is that it shows the SNOMED concepts that are actually present in the EMR and how often they occur via the frequency counts based on the PL terms recorded. This tree can aide in the retrieval of relevant concepts recorded using different PL terms. For instance, by specifying the concept "56265001|Heart disease" in the query, one should expect to retrieve all sub-types under "5754005|Acute myocardial infarction" and "12026006|Paroxysmal tachycardia." On the other hand, by specifying the concept "57054005|Acute myocardial infarction" in the query, the sibling concept "12026006|Paroxysmal tachycardia" should automatically be excluded.

DISCUSSION

A proposed Methodology

Drawing on the lessons learned from this study, we propose the following steps for general practitioners to encode problem lists from their EMR in SNOMED as a first step for review before full-scale conversion:

1. Extract all PL entries from the EMR and tabulate the frequency of the PL terms present;
2. Catalogue all unique words across the PL terms;
3. Examine all unique words and PL terms to identify and revise for acronyms, abbreviations, spelling variants and errors;
4. Match the PL terms to SNOMED concepts using the matching algorithms outlined in this paper (contact authors for copies of the algorithms);
5. Create detailed and summary outputs to show the exact, all, partial and semantic matches found;
6. Review matched SNOMED terms for accuracy; remove successful exact-match and match-all terms from further matching cycles;
7. Repeat steps 3 through 6 for remaining partial matches until no further matches found;
8. Post-coordinate remaining PL terms with qualifier, refinement and combined concepts;
9. Create a pruned PL hierarchy tree showing all concepts with positive frequency counts and immediate super-type concepts;
10. Create index table containing unique identifiers for the PL and matched SNOMED terms.

Implications

Post-coordination is thought to be a feature that is difficult to implement. Yet based on the small number of SNOMED concepts used in this study to post-coordinate our PL terms, it seems feasible to achieve. We did note the use of pre-coordination in SNOMED CT is unpredictable, and it seems common to include acronyms within SNOMED descriptions. Careful use of modifiers such as laterality, chronicity and severity should be considered. Further studies are needed.

Critics often balk at the unwieldy size and complexity of SNOMED CT as too impractical for local use. In Canada the vendor and general practice communities, which are often small in size, are reluctant to adopt SNOMED CT, questioning their return on value for the effort required. From this study, we have shown it is feasible to incorporate SNOMED CT into EMR in the general practice setting. The methodology we have outlined is practical even for small medical offices with an EMR in place. We have also shown the potential use of SNOMED CT to improve the quality of recall from its hierarchy. The ability to demonstrate return on value, as in our encoding of problem lists with SNOMED CT to improve recall, is an important first step for practitioners to consider before full-scale conversion of their EMR.

Limitations

There are several limitations to this study. First, the PL terms used have been established over the years mainly by one practitioner from a single setting, which are likely to vary between practices. Second, our current matching algorithms do not take into account subtype hierarchy to limit searches, which could otherwise restrict unlikely choices such as Physical Object and Substance. Third, the evaluation of this methodology is incomplete to date; the full extent of the post-coordination effort required to encode the entire set of PL terms in this EMR should be further examined and reported. Fourth, the use of our partly instantiated hierarchy tree to improve recall quality, while promising, requires more thorough investigation into its utility with more complex real-life cases. Its design should also be aligned with the existing SNOMED navigation hierarchy feature that is already in place as part of the new RefSet release.

Next Steps

We are developing a Web-based mapping tool made up of the matching algorithms described earlier to allow the matching of clinical terms to SNOMED CT in an interactive or batch mode. With our focus continued to be on general practice EMR systems, there are several steps ahead to be considered. For instance, we need to expand the use of SNOMED terms to other parts of the EMR such as procedures, medications and billing. We also need to refine our encoding methodology to take into account specific contexts such as past/family history and health risks, and to use subtype hierarchy to improve search precision. The inclusion of frequency statistics on the distribution of matched SNOMED CT terms across the hierarchies would be useful to validate the results. These efforts should aid in the eventual creation of a primary care SNOMED subset, and eventually a concept model in the primary care domain. But most important, we should continue to exploit ways by which the use of SNOMED CT in the EMR can actually enhance patient care.

ACKNOWLEDGMENTS

Funding support for this project has been provided by the Canadian Institutes for Health Research Strategic Training Initiative.

Original PL Term	Type of Match	Identifier	Id Type	SNOMED Term	Descn Type	Descn Status
Atrial Fibrillation	Exact	49436004	C	Atrial fibrillation (disorder)	F	0
		82343012	D	**Atrial fibrillation**	**P**	**0**
Hypertension	Exact	38341003	C	Hypertensive disorder, systemic arterial (disorder)	F	0
		1215744012	D	Hypertensive disorder	P	0
		64176011	D	**Hypertension**	**S**	**0**
Gastroesophageal Reflux - GERD	All	235595009	C	Gastroesophageal reflux disease (disorder)	F	0
		2535970019	D	**GERD – Gastro-esophageal reflux disease**	**S**	**0**
Cirrhosis	Semantic	155809006	C	Cirrhosis	U	4
		19943007	C	Cirrhosis of liver (disorder)	F	0
		33568015	D	**Cirrhosis of liver**	**P**	**0**
Atrial Fibrillation - Chronic	Post, Exact	**82343012**	D	**Atrial fibrillation**	**P**	**0**
		288524001	C	Courses (qualifier value)	F	0
		428182017	D	**Courses**	**P**	**0**
		90734009	C	Chronic (qualifier value)	F	0
		150360019	D	**Chronic**	**P**	**0**
Chronic Kidney Disease - CKD	Post	90708001	C	Kidney disease (disorder)	F	0
		150315015	D	**Kidney disease**	**P**	**0**
		263502005	C	Clinical course (attribute)	F	0
		391753013	D	**Clinical course**	**P**	**0**
		90734009	C	Chronic (qualifier value)	F	0
		150360019	D	**Chronic**	**P**	**0**
Headache Migraine	Post, Exact	37796009	C	Migraine (disorder)	F	0
		63055014	D	**Migraine**	**P**	**0**
		246090004	C	Associated finding (attribute)	F	0
		367802015	D	**Associated finding**	**P**	**0**
		25064002	C	Headache (finding)	F	0
		41990019	D	**Headache**	**P**	**0**

Table 4. Examples of matched PL and SNOMED terms by exact, all, semantic and post-coordinated matches. Legend: Identifier (contains ConceptId or DescriptionID depending on Id-Type); Id Type (C- Concept, D-Description); Descn-Type (P-preferred, S-synonym, F-fully specified name, U-undefined); Descn-Status (0-current, 4-ambiguous); note that all selected SNOMED terms are shaded and in bold

Rec	PL-Id	PL-Term	ConceptId	DescriptionId	Match	AttributeId	SequenceId
1	160	Atrial Fibrillation	49436004	83243013	Exact	0	0
2	789	Hypertension	38341003	64176011	Exact	0	0
3	685	Gastroesophageal Reflux - GERD	235595009	2535970019	All	0	0
4	32666	Chronic Kidney Disease CKD	90708001	150315015	Post	0	0
5	32666	Chronic Kidney Disease CKD	90734009	150360019	Post	263502005	1
6	431	Cirrhosis	19943007	33568015	Semantic	0	0
7	1044	Headache Migraine	37796009	63055014	Post, Exact	0	0
8	1044	Headache Migraine	25064002	41990019	Post, Exact	246090004	1

Table 5. Examples of the index table linking the original PL terms to matched SNOMED terms. Legend: SequenceId indicates the relative ordering of the post-coordinated records

Two sets of post-coordinated terms shown above

- ***49601007 Disorder of cardiovascular system (disorder) - 1***
  - 128487001 Acute disease of cardiovascular system (disorder)
    - 127337006 Acute heart disease (disorder)
      - 57054005 Acute myocardial infarction (disorder)
        - ***70211005 Acute myocardial infarction of anterolateral wall (disorder) - 1***
        - ***73795002 Acute myocardial infarction of inferior wall (disorder) - 5***
        - ***307140009 Acute non-Q wave infarction (disorder) - 5***
      - ***12026006 Paroxysmal tachycardia (disorder) - 1***
  - 9904008 Congenital anomaly of cardiovascular system (disorder)
    - 363028003 Congenital anomaly of cardiovascular structure of trunk (disorder)
      - ***13213009 Congenital heart disease (disorder) - 1***
        - 10818008 Congenital malposition of heart (disorder)
          - ***27637000 Dextrocardia (disorder) - 1***
  - 27550009 Disorder of blood vessel (disorder)
    - 359557001 Disorder of artery (disorder)
      - 72092001 Arteriosclerotic vascular disease (disorder)
        - ***53741008 Coronary arteriosclerosis (disorder) - 9***
      - 414024009 Disorder of coronary artery (disorder)
        - ***53741008 Coronary arteriosclerosis (disorder) - 9***
  - 55855009 Disorder of pericardium (disorder)
    - ***3238004 Pericarditis (disorder) - 2***
      - ***15555002 Acute pericarditis (disorder) - 1***
  - ***56265001 Heart disease (disorder) - 1***
    - 127337006 Acute heart disease (disorder)
      - 57054005 Acute myocardial infarction (disorder)
        - ***70211005 Acute myocardial infarction of anterolateral wall (disorder) - 1***
        - ***73795002 Acute myocardial infarction of inferior wall (disorder) - 5***
        - ***307140009 Acute non-Q wave infarction (disorder) - 5***
      - ***12026006 Paroxysmal tachycardia (disorder) - 1***

PL-Id	Original PL Term	Concept Id	Fully Specified Name
4435	Dextrocardia	27637000	Dextrocardia (disorder)
10086	Heart Disease	56265001	Heart disease (disorder)
10087	Heart Disease Congenital	13213009	Congenital heart disease (disorder)
1035	MI Inferior Myocardial Infarction	73795002	Acute myocardial infarction of inferior wall (disorder)
12653	Myocardial Infarction Anterolateral	70211005	Acute anterolateral myocardial infarction (disorder)
1591	Myocardial Infarction Subendocardial (Non Q wave)	307140009	Acute non-Q wave infarction (disorder)
1202	Pericarditis	3238004	Pericarditis (disorder)
13641	Pericarditis Acute	15555002	Acute pericarditis (disorder)
15976	Tachycardia Paroxysmal	12026006	Tachycardia paroxysmal (disorder)

Figure 1. A partial SNOMED hierarchy for cardiovascular disorders derived from a set of original PL terms. The upper figure portion shows the partial SNOMED hierarchy for cardiovascular disorders; the lower figure portion shows the original PL terms with the matched SNOMED concepts and their fully specified names. In the hierarchy, concepts that are bold and italicized are exact matches for the PL terms, followed by the frequency of how often they appeared in the EMR.

REFERENCES

1. Chute CG, Elkin PL, Fenton SH, Atkin GE. A clinical terminology in the post modern era: pragmatic problem list development. *Proceedings AMIA Ann Symposium* 1998; 795-9.
2. Warren JJ, Collins J, Sorrentino C, Campbell JR. Just-in-time coding of the problem list in a clinical environment. *Proceedings AMIA Annual Symposium* 1998; 280-4.
3. Petersson H, Gunnar N, Strender LE, Ahlfeldt H. The connection between terms used in medical records and coding system: a study on Swedish primary health care data. *Medical Informatics* 2001; 26(2):87-99.
4. Wang SJ, Bates DW, Chueh HC, Karson AS, Maviglia SM, Greim JA, Frost JP, Kuperman GJ. Automated coded ambulatory problem lists: evaluation of a vocabulary and a data entry tool. *International Journal of Medical Informatics* 2003; 72, 17-28.
5. Fabry P, Baud R, Ruch P, Despont-Gros C, Lovis C. Methodology to ease the construction of a terminology of problems. *International Journal of Medical Informatics* 2006;75:624-32.
6. Meystre S, Haug PJ. Automation of a problem list using natural language processing. *BMC Medical Informatics and Decision Making* 2005;5(30):1472-6947/5/30.
7. Elkin PL, Brown SH, Husser CS, et al. Evaluation of the content coverage of SNOMED CT: ability of SNOMED clinical terms to represent clinical problem lists. *Mayo Clinic Proceedings* 2006;81(6):741-8.
8. O'Halloran J, Miller GC, Britt H. Defining chronic conditions for primary care with ICPC-2. *Family Practice* 2004;21(4):381-6.
9. Wasserman H, Wang J. An applied evaluation of SNOMED CT as a clinical vocabulary for the computerized diagnosis and problem list. *Proceedings AMIA Symposium* 2003, 699-703.
10. Comment: one author reviewed datasets provided by colleagues from hospitals in Buenos Aires in Argentina, Kaiser Permanente in United States, and Sherbrooke in Canada during Fall 2007.
11. Comment: WONCA is the World Organization of Family Doctors, and WICC is the WONCA International Classification Committee. URL http://www.globalfamilydoctor.com/; Jan20/08.
12. Lee DHK. *Reverse Mapping ICD-10-CA to SNOMED CT*. UVic Master of Science research project report, Oct 2007. Unpublished.
13. Wang Y, Patrick J, Miller G, O'Halleran. Linguistic mapping of terminologies to SNOMED CT. *Semantic Mining Conference on SNOMED CT* Oct 2006, Copenhagen, Denmark.
14. Kleinsorge R, Willis J, et al. UMLS Overview – Tutorial T12. *AMIA Annual Symposium* 2006. http://165.112.6.70/research/umls/pdf/AMIA_T12_2006_UMLS.pdf. Jan15/08.
15. National Library of Medicine. *The SPECIALIST Lexicon*. http://lexsr3.nlm.nih.gov/LexSysGroup/Projects/Summary/lexicon.html. Jan15/2008.
16. Goldsmith JA, Higgins D, Soglasnova S. *Automatic Language-specific Stemming in Information Retrieval.* Springer-Verlag Berlin Heidelberg 2001.
17. CliniClue. The Clinical Information Consultancy, Ltd., UK. http://www.cliniclue.com/software. available for download. Jan22/08.

Post-Coordination in the Mapping of Interface Terms of a Clinical Wound Documentation System to SNOMED CT

Martin Boeker[a], Stefan Schulz[a], Thilo Schuler[a,b]
[a] Dept. of Medical Biometry and Medical Informatics,
University Medical Center Freiburg,
[b] Department of Dermatology,
University Medical Center Freiburg, Germany
martin.boeker@uniklinik-freiburg.de

The objective of this work is to provide a formalization of the semantics of SNOMED CT's refinement rules in Description Logics and to exemplify their usage on a real world wound documentation system.

The goal of unambiguous documentation and communication of medical information with explicit semantics can be reached by combining standards and terminologies. Information Models (e.g. the HL7 Clinical Document Architecture on Level 3) together with terminology systems (e.g. LOINC and SNOMED CT) are promising candidates for building a semantically interoperable framework for electronic health records.

We investigated how LOINC and SNOMED CT concepts can unambiguously and completely cover user interface terms of an existing electronic, form-based documentation system used in clinical dermatology. Especially, the feasibility of post-coordinating complex expressions according to the SNOMED CT terminology model is target of our investigations. Besides analyzing completeness and uniqueness of the mappings and the user-friendliness of the mapping process, we discuss the different ways of post-coordination (refinement types) presented in SNOMED CT's technical documentation. Where post-coordination was required, we adhered to the SNOMED CT terminology model refinement types and the "SNOMED Compositional Grammar" syntax.

The manual mapping process proved to be time consuming and prone to ambiguous solutions where post-coordination of SNOMED CT expressions was necessary. However, for most user interface terms a complete semantic representation could be generated. A coverage of nearly 100% of clinical user interface terms shows the appropriateness of SNOMED CT as a reference terminology for the domain under scrutiny. The natural language descriptions of refinement types in the SNOMED CT documentation were formalized in Description Logics and reduced to four basic patterns.

Problems with coding and post-coordination can be explained by weak documentation and poor tool support. The structure of the documentation forces users to collect necessary information from several SNOMED CT reference documents. Although mechanisms for post-coordination allowed to express a substantial amount of terms we suggest that tool support and formalized documentation for post-coordination (refinement) is enhanced. Tool support should reduce browsing complexity, support post-coordination and give clear advice how to use SNOMED CT according to the SNOMED CT compositional grammar and refinement rules. Furthermore, we recommend a thorough redesign of the post-coordination guidelines which entails the clarification of SNOMED CT's logical and ontological foundations.

Using SNOMED CT Concepts for PAIRS

A.M. Mohan Rao, MB BS., MS,
Logic Medical Systems, Hyderabad, AP, India
ammohanrao@logicmedicalsystems.com

SNOMED CT medical vocabulary can be used to identify complementary features in a database. This functionality is used to develop a natural language processor (NLP) for PAIRS (Physician Assistant Artificial Intelligence Reference System). Although about 99% of concepts in PAIRS are present in SNOMED CT some features missing in it makes it unacceptable for any diagnostic decision support system (DDSS). Here we show that implementation of another NLP along with SNOMED CT makes it practically useful.

INTRODUCTION

Ideally medical databases must have clinical entities whose features comprise of medical domain , apart from microbiological, pathological, radiological and surgical domains. Computerization of such a data enables many interesting functionalities. Apart from being a learning tool it can also be a new source of knowledge which is of use in diagnosis. For example, feature-disease or feature-feature links can be deduced from disease-feature links. One can aim to achieve a uniform usage of clinical terms from such databases.

Utility of a database is limited by its completeness. Use of such a database helps one to design a Natural Language Processor which helps in data extraction from different data sources. Here, we show how SNOMED CT is a source of uniform clinical terms that not only can be used to simplify PAIRS database but also helps code a NLP that can be used for further evaluations. In this paper we use the terms PAIRS and SNOMED as representing their respective databases.

SNOMED vocabulary is used for several applications including Electronic Patient Records (EPR). However, as an evolving database many clinical terms are still missing in SNOMED which may be an impediment in developing a fully functional NLP. Here we show that this problem can be overcome by using a substitute algorithm (AINLP) (in addition to SNOMED algorithm) that works on PAIRS database.

PAIRS is an internet based DDSS that gives diagnosis for a given patient data. It is available free for evaluation from www.lmspairs.com upon request. Its artificial intelligence (AI) is based on a variational probabilistic belief networks as developed by Jaakkola & Jordan [1]. PAIRS database has 547 internal medicine diseases, 3700 unique features and 40 000 disease-feature links. Feature-disease links of a patient data are extracted from PAIRS database by AINLP. This process is limited by ability of AINLP algorithm to find a complementary feature in the database [2]. This limitation is rectified by SNOMED CT algorithm that works on SNOMED CT database.

There are several technological difficulties involved in NLP which may be caused by the algorithm, its implementation or its run times. These aspects are discussed at the end. Here we show how SNOMED CT can be used as a NLP for a DDSS.

MATERIALS

The computations are performed using HP Pavilion Entertainment PC with (1.6 GHz Intel processor, 1014 MB RAM) MS Vista operating system. We used a Perl program eutils written by Oleg Khovayko of National Library of Medicine to download about 2.5 million abstracts from PUBMED. NCBI Clinical Queries Research Methodology Filters developed by Haynes RB. et al is used. Routine search and extraction processes are done by MS Visual C++ programming language. A customized database PAIRS-DB is used to store and extract data programmatically. We obtained an International affiliate license for SNOMED CT from National Library of Medicine, USA. Sun Micro-systems Net Beans 5.5.1. IDE along with Visual Web Pack is used for internet enabling PAIRS. PAIRS database is owned and developed by Logic Medical Systems (www.logicmedicalsystems.com).

PAIRS

PAIRS comprises about 75000 disease-feature links for over 1700 diseases. This database is developed over a decade from various text sources and journals. We used a perl program eutils (written by Oleg Khovayko of National Library of Medicine) to download up to 2500 abstracts for each of the 1700 diseases from PUBMED. Queries based on Research Methodology Filters developed by Haynes RB. et al., are used for searching PUBMED [3]. We programmatically searched for each of the features in these abstracts. Initially MS Access is used to store the data as disease-feature links. It has over 7000 unique features. About 540 diseases having substantial number of features (about 40 000) are identified for application of artificial intelligence (AI) for diagnosis.

METHODS

SNOMED CT indices

We used SNOMED algorithm for assigning indices for PAIRS features. Over 3100 features are used as input data for this analysis. The search is facilitated by using a customized database known as PAIRS-DB. PAIRS-DB is programmatically loaded with SNOMED CT indices of duel_keywords, keyword and index-relationships. Each of the feature's duel_keyword and keyword indices are found using SNOMED algorithm. The indices (base-indices) are looked for inter-relationships in SNOMED CT relationships table. Those indices (relative-indices) having maximum frequency are selected as representative of the clinical feature. If a feature does not have any relative-index its base-indices are used for searching a concept. We used 250 clinico-pathological cases (CPC) published in New England Journal of Medicine (NEJM) between 1996-2003 for studying SNOMED CT based NLP functionality.

AINLP

AINLP is a substitute NLP that works independently of SNOMED algorithm. Its component database tables include: (a). general words: valid medical word and word pieces, (b).abbreviations: meanings of abbreviations. (c). synonyms, (d). antonyms and (e). feature-count: number of feature-disease links for given feature in PAIRS database. Input features from a patient data file are separated into their constituent words and their derivative word pieces. Word pieces are derived by deleting single letters from the end of each word in a cyclical fashion. These are checked against "general word list" for selecting only valid medical words. Words are searched for abbreviations and if found their meanings are selected. Further, synonyms and their antonyms are searched for the given words. Finally, the input data is searched for their complementary features in PAIRS database.
Computational times for AINLP are far shorter than SNOMED CT based NLP by several degrees. For example, for a list of 10-15 input features AINLP can find complementary features in 1-3 seconds where as same for SNOMED CT involves much long (sometimes as much as 1-3 seconds for each feature). Hence, AINLP is always run and if no complementary feature is found then only SNOMED CT based NLP is used.

SNOMED CT vs AINLP

For a given input feature duel keywords, keywords, their indices and relationships between indices are generated using SNOMED CT algorithm as described in SNOMED CT Clinical Terms Technical Implementation Guide [4]. Tables used for index search are: sct_concepts_duelkeyindex_20070731, sct_concepts_wordkeyindex_20070731. PAIRS-DB has 27 folders alphabatically labelled (plus a base folder that takes numerical and non-alphabetic data). Each of the folder is again assigned 27 folders. To reduce the computational time both the tables are programmatically loaded into PAIRS-DB. Finally, sct_relationships table has only those indices that are represented by PAIRS features. SNOMED CT based NLP runs in the following way: (a). find duelkeyindices, (b). find keyword indices that share duelkey indices . Finally, find maximally represented indices in sct_relationships table which is a PAIRS complementary feature for the input feature.
For a given input feature AINLP converts it into its words, and word pieces (by deleting single letters in a loop from the end). General words or abbreviations in this pool are identified and a search of PAIRS list of features is made. Complementary features in PAIRS database are identified by finding those that represent maximally in a search. Since volume of information processed in AINLP is much smaller than SNOMED CT based NLP AINLP computational times are much shorter.

Evaluation of NLP

PAIRS NLP has two components: AINLP and SNOMED CT. We tested the functionality of each in a two stage process. Firstly, we tested each component's ability to identify complementary features in PAIRS database. Secondly, we verified its function by testing PAIRS diagnostic output. We used 250 CPC cases of NEJM for this study. Each of the case shares some features from 3100 unique features of PAIRS database. Since AINLP computational times are much shorter than those of SNOMED CT based NLP AINLP is used in PAIRS always. SNOMED CT based NLP is used only if no complementary feature is generated by AINLP.

RESULTS

Functionality of SNOMED CT indices

Initially we included 3100 features of PAIRS in our analysis and identified 31 concepts (1 out of 100) that are not represented in SNOMED CT. Table 3 gives the features in PAIRS database that are identified as missing concepts in SNOMED CT. Here, we show results of search of SNOMED CT concepts in PAIRS.
Multiple indices are assigned to a given concept in SNOMED CT (see table 1). This can be problematic

in assigning an appropriate index number to a complementary feature in PAIRS. This problem can be resolved by use of relationships table.Our analyses of finding SNOMED CT concepts in PAIRS, we look for common relationships between indices/concepts rather than indices/concepts themselves. A number that is generated at a higher frequency in our search is preferred in assigning it to a feature in PAIRS.

For example "acute abdomen" may imply "acute abdominal pain". Same index number (9991008) is shared between three different concepts. Index for "Acute abdomen" (920005) is also assigned to "Acute abdominal pain syndrome". If we check the relationships table, only common relationship between two different indices of "Acute abdominal pain" (9991008 & 116290004) is "Abdominal pain" (21522001). Given two different indices for "Abdominal pain" (9991008) and "Abdominal pain through to back" (74704000) the only common relationship could be one index number identified by concepts: "Abdomen" or "Abdominal structure" (113345001).

SNOMED CT can be used as a form of NLP by implementing its algorithm and assigning a unique identifier index for a complementary feature in a given database. Any abbreviation or concept in a database can be accessed by using this functionality.

For example, index numbers for the feature ESR is given in table 2. By assigning 416838001 index for the feature "Erythrocyte sedimentation rate raise" in PAIRS, a user can access correctly by either entering "ESR" or "Erythrocyte sedimentation rate".

SNOMED CT missing concepts

Several radiological terms used in routine clinical practice are missing in SNOMED CT database (Table 3). Some of these features occur in significant number of diseases making them indispensable for any DDSS. For example, on chest x-ray "tree in bud pattern", miliary infiltrates, (contrast) enhancing lesion, pulmonary nodules represent 18,19,12 and 45 diseases respectively in PAIRS database. "Focal neurological lesion" missing concept represents about 27 PAIRS diseases. "Hypopigmented macules" and "heliotrope rash eyelids" are examples of other important missing concepts.

Evaluation of SNOMED CT indices

We assigned a SNOMED CT index to each of 3100 complementary features in PAIRS. We selected features from each patient data of 250 CPCs (NEJM) and looked for them in PAIRS using SNOMED CT algorithm. We are able to find the feature if input feature is represented in SNOMED CT. However, if a feature is not in SNOMED concepts, we are unable to find it in our results. We overcome this issue by using AINLP along with SNOMED CT. By testing features from each of patient data in 250 CPCs (NEJM) we are able to generate representative features in PAIRS. These representative features constitute not only the exact input feature but also features related to it.

Functionality of NLP

PAIRS NLP has two components: AINLP and SNOMED CT whose implementation of the algorithms is described in methods. AINLP is best suited for 3100 features listed in PAIRS. Advantages of it include the computational times which are rapid in finding complementary features. However, for those features which are not part of 3100 features, AINLP is not tested. In case where AINLP fails to find a complementary feature SNOMED CT algorith is used, thus the limitations of AINLP are supplemented by SNOMED CT.

We test the functionality of NLP by 250 CPC cases of NEJM. Each of the case has some of the features of 3100 features listed in PAIRS. Each case has about 10-30 features. The computational times are related

to the number of features in a case. If the features are more the time NLP takes more time to complete.

Index No.	SNOMED CT concept
9209005	Acute abdomen
9209005	Acute abdomen, NOS
158499006	[D]Acute abdomen
163250006	O/E - acute abdomen
207221008	[D]Acute abdomen
207255006	[D]Acute abdomen
268942007	O/E - acute abdomen
9991008	Spasmodic abdominal pain
9991008	Acute abdominal pain
9991008	Colicky abdominal pain
9209005	Acute abdominal pain syndrome, NOS
9209005	Acute abdominal pain syndrome
83132003	Upper abdominal pain
74704000	Abdominal pain through to back
71850005	Abdominal pain worse on motion
54586004	Lower abdominal pain
21522001	Abdominal pain
21522001	AP - Abdominal pain
116290004	Acute abdominal pain
111985007	Chronic abdominal pain
102614006	General abdominal pain-symptom
102614006	Generalised abdominal pain
102613000	Localised abdominal pain

Table 1 Multiple index numbers representing same SNOMED CT concept. This complexity can be resolved by finding relationships between them in relationships table.

Index No.	SNOMED CT concept
103208001	Erythrocyte sedimentation rate
103208001	ESR - Erythrocyte sedimentation rate
104154005	Erythrocyte sedimentation rate, non-automated
104155006	Erythrocyte sedimentation rate, automated
142875000	Erythrocyte sedimentation rate
165464006	Erythrocyte sedimentation rate
165466008	Erythrocyte sedimentation rate
165467004	Erythrocyte sedimentation rate
165468009	Erythrocyte sedimentation rate
165468009	Erythrocyte sedimentation rate
365649001	Finding of erythrocyte sedimentation rate
365649001	Erythrocyte sedimentation rate
365649001	Erythrocyte sedimentation rate - finding
416103000	Elevated erythrocyte sedimentation rate
416560009	Erythrocyte sedimentation
416838001	Erythrocyte sedimentation rate measurement
416838001	ESR - Erythrocyte sedimentation rate
416838001	Erythrocyte sedimentation rate measurement

Table 2 Natural language processor functionality by assigning unique identifier of SNOMED CT. See text for explanation.

During the test, the maximal computational times are in range of 15-20 seconds and each input feature finds its complementary feature correctly. This satisfactory results do not rule out possibility of features that may not give complementary features both by AINLP and SNOMED CT. It needs further testing to identify such features and take necessary steps.

DISCUSSION

SNOMED CT as NLP

SNOMED CT concepts are useful in many applications including EPR. Its application in Diagnostic Decision Support Systems (DDSS) is limited by presence (or absence of) a concept. It can simplify querying in a clinical database [5]. Its clinical vocabulary can be used for computerized diagnosis and problem list [6]. Almost complete coverage (98.5%) of concepts in SNOMED CT is reported. Out of 5000 features, about 92.5% concepts are covered in SNOMED CT [7]. Here, we report about 99% concept coverage in SNOMED CT in present study.

Missing concept in SNOMED CT	PMID
Air under diaphragm	7509402 (4)
Ataxia, sensory	8036880 (2)
Ataxia, stance	14561428 (1)
Spontaneous bleeding	14979383 (4)
Bowel wall thickness	16632735 (3)
Cervical sounds	
Edema of face	16340761 (6)
Exercise intolerance/ Exertional intolerance/ Effort intolerance	16689370 (5)
Nephromegaly	12621244 (1)
Focal neurological signs	16499723 (27)
Pulmonary oligemia	15658055 (4)
Heliotrope rash eyelids	10770031 (4)
Hypopigmented macules	15884465 (9)
Immobile diaphragm	*(2)*
Infundibular pinching	*(4)*
Leucoerythroblastic changes	*(4)*
Miliary infiltrates	11555380 (19)
Pre syncope	16195623 (9)
Pseudo fractures	6147751 (4)
Pulmonary nodules	15875070 (45)
Enhancing lesion	15891158 (12)
Secondary achalasia	11176337 (3)
Toxic granulocytosis	
Transient erythema	
Unilateral tongue weakness	12490688 (1)
Vertebral tenderness	7895748 (5)
Chest x-ray Hyperinflated_lungs	16338298 (5)
Chest x-ray honey comb appearance	*(5)*
Chest x-ray tree in bud pattern	*(18)*
Chest x-ray space occupying lesion of lung	*(2)*
X-ray skull space occupying lesion of brain	*(9)*

Table 3 PAIRS concepts missing in SNOMED CT. PMID shows related abstract number in PUBMED. Number in parenthesis shows number of diseases the feature is present in PAIRS. Feature-disease links for each are: (1). acute appendicitis (2). sub acute axonal polyneuropathy. (3). Wernicke-Korsakoff syndrome.(4). acute leukemia. (5). Crohn disease. (6). aortic incompetence. (7). aortic arch syndrome. (8). cardiac failure and dilated cardiomyopathy. (9). polycystic kidney disease. (10). AIDS related lymphoma. (11). CREST syndrome.(12). mixed connective tissue diseases. (13). ataxia telangiectasia. (14). liver abscess. (15). Hodgkin lymphoma. (16).megaloblastic anemia due to folic acid deficienc. (17). breast cancer. (18). arrhythmia. (19). cholestasis jaundice. (20). actinomycosis. (21). anterior

> *cerebral artery syndrome. (22). gastric carcinoma. (23). gram -ve septicemia.* (24). *alcoholism (25). cerebral ischemia. (26). ankylosing spondylitis. (27*). *amyloidosis. (28). bronchiectasis. (29). allergic angiitis. (30). lung cancer. (31). epilepsy.*

PAIRS is a DDSS which has a database on which an artificial intelligence (AI) system works [2]. It has over 43 000 disease-feature links which are quantified and work in AI to give a diagnosis. Its AI system is based on variational probabilistic belief networks. For this, we require a robust NLP that can filter a clinical feature given any input. SNOMED CT is an ideal data source for such an application. However, its application is limited by a number of features missing. Many of these features may belong to radiology domain, which is crucial for diagnosis. For example, pulmonary nodules as a feature may be present in as many as 45 diseases in PAIRS but is missing in SNOMED CT. This suggests that use of SNOMED CT alone is insufficient NLP for PAIRS.

AINLP as NLP

For features which are exact complements of PAIRS or abbreviations of them, AINLP gives good results. However, results not shown here suggest that on its own AINLP is insufficient as NLP for PAIRS. This prompted us to use both AINLP and SNOMED CT for PAIRS. It is expected that those deficiencies in SNOMED CT are supplemented by AINLP. SNOMED CT has vast database and hence takes considerable amounts of time (up to about 3 seconds for each feature alone), we run AINLP first and if necessary (i.e., if a feature is not found by AINLP) then SNOMED CT is run. Thus, we check their computational run times.

SNOMED CT missing concepts

Typically diagnosis of a case involves not only history and physical examination but also interpretation of radiological data apart from pathological and microbiological data. Terms such "honey comb appearance" to describe a set of disease patterns is common in clinical setting. It is preferable to have these included in SNOMED CT concepts. As reported in results (see Table 2) many such concepts missing in SNOMED CT makes it impossible to use for a DDSS. Hence, AINLP is used along with SNOMED CT which substitutes the missing functionality. It is sometimes possible that a feature may not have its complementary for AINLP. In such a case SNOMED CT is allowed to run. We are not yet come across a feature that has no complentary for both AINLP and SNOMED CT. Presumably such a thing can happen either as a bug in the program or PAIRS-DB.

PAIRS as a DDSS

PAIRS is an internet enabled and can be used for diagnosing difficult cases. Features entered in a text area are processed by AINLP and SNOMED CT to select complementary features in PAIRS. These features are further processed by an AI system to give probabilities of diagnoses. These diagnoses are based on Bayesian probabilities and depend on age, gender and geographic parameters. Effectiveness of PAIRS functionality is limited by several technical and user difficulties. Generally, users like to enter patient data in a free form rather than choose from a table and they expect NLP to recognize the complementary features in the database for any given feature. For example, "myalgia" may suggest "bodypain", "bodypains", "body pain" or "body pains". However, "body" and "pain" are common for many (upto 30 000) concepts in SNOMED CT and hence its runtime process may become unacceptable (over 300 seconds). This problem is solved in PAIRS by using AINLP. PAIRS NLP consisting of both AINLP and SNOMED CT algorithms are tested using over 3100 unique features. Both the algorithms are complex and hence may yeild unpredictable outcomes in rare cases. The AI of PAIRS involves convex analysis and gives its diagnoses using a complex process. Therefore, activity of NLP may or may not affect diagnostic ability of PAIRS. It may not affect adversely if for example, "abdominal pain" finds a feature "abdominal pain, upper". But it may be otherwise if a complementary feature is not found at all. Many of these difficulties are minimized by use of AINLP followed if a complementary feature is not found by SNOMED CT algorithm.

Key advantages of PAIRS as a DDSS include not only its ability to generate a differential but also suggest procedures and features to look for in the patient for a given diagnosis. Its diagnostic process includes age, gender and geographic criteria based on epidemiological data from NHS, NCHS and WHO. PAIRS judgment on a diagnosis is graded into 7 heirarchial levels (certainly, as far as evidence goes, probably, necessarily, presumably, possibly and impossibly) on basis of variational probability, age, gender, geographic data, precipitating cause, duration, pathogenesis and system/ nonsystemic involvement. Highest grade prediction for a selected disease is attained only if all criteria match to the real data. For example, tuberculosis as selected disease does not match a geography "United States of America" because this presumption does not support epidemiological data of NCHS. Hence, PAIRS judment never be "certainly" for such a diagnosis. PAIRS ability to suggest features to look for in a patient for given diagnosis is also very useful in arriving at a possible diagnosis.

PAIRS has a free component PAIRS-LM (which covers 980 diseases of which 580 are common internal medicine diseases) that gives links to about 2.5 million abstracts of PUBMED in National Library of Medicine (NLM). Each disease has about 2500 abstracts categorized into diagnosis, features, genetics, treatment, complications, prevention, incidence, nationality etc. This information is useful in the process of arriving at a diagnosis.

Two of the gold standards suggested for any DDSS include procedures to be performed for a given case and ability to extract data from EPR and give an output [8]. PAIRS suggests procedures for a given case. It also has a facility to select a case from multiple cases and give a diagnostic/ procedural output. However, several of technical difficulties discussed by Berner [8] such as correctness of diagnosis, quality of differential, user acceptness and amount of use are critical issues that still remain. Several additional advantages of PAIRS make it a possibly useful tool for arriving at a diagnosis atleast in difficult cases.

Technical problems

Main difficulty while using SNOMED CT arises because it has multiple indices assigned to a given concept. However, from its relationships table one can derive parent and child relationship between various indices (see results). Sometimes the relationships table may not yeild a clear parent/child ontologies for a given concept [9]. In such a case one may have problem in assigning an index to the complementary feature in user database. Results shown here are similar to those reported by others[9-10].

Computational times involving AINLP and SNOMED CT vary depending on the input feature. Typically AINLP run-times are much shorter than those for SNOMED CT and is run only if AINLP cannot generate any complementary feature. These difficulties prompt suggestions for users of PAIRS to get familiarize with PAIRS list of features before evaluating its diagnostic functionality and preferably to limit to those in 3100 list of PAIRS features as input data.

Acknowledgments

I thank National Library of Medicine, USA for giving an International Affiliate license for SNOMED CT. I thank Oleg Khovayko of National Library of Medicine, USA for Perl program e-utils.

References

1. Jaakkola, TS and Jordan, MI. Variational methods and QMR-DT database. J. Artificial Intelligence. 1999:10,291-322.
2. Mohan Rao.AM. PAIRS:a diagnostic decision-support engine. BJHC & IM. 2004:30-2.
3. Wilczynski, NL. and Haynes, R.B. Developing Optimal Search Strategies for Detecting Clinically Sound Causation Studies in MEDLINE. AMIA Annu Symp Proc 2003: 719-723.
4. SNOMED Clinical Terms Technical Implementation Guide, The International Health Terminology Standards Development Organization. July 2007. International Release.
5. Lieberman, MI. The Use of SNOMED CT Simplifies Querying of a Clinical Data Warehouse.AMIA Annu Symp Proc. 2003; 2003: 910.
6. Wasserman, H., Wang, J. An Applied Evaluation of SNOMED CT as a Clinical Vocabulary for the Computerized Diagnosis and Problem List. AMIA Annu Symp Proc. 2003; 2003: 699–703.
7. Elkin, PL, Brown, SH et al. Evaluation of the Content Coverage of SNOMED CT: Ability of SNOMED Clinical Terms to Represent Clinical Problem Lists. Mayo Clin Proc.2006; 741-748.
8. Berner, ES. J Am Med Inform Assoc.2003, 10, 608-610.
9. Bodenreider, O. Smith, B.Kumar, A and Burgun, A., KR-MED 2004,12-20.
10. Ceusters, W., Smith, B., Kumar A and Dhaen, C. Mistakes in Medical Ontologies: Where Do They Come From And How Can They Be Detected? Ontologies in Medicine:Proc Workshop on Medical Ontologies. 2003.

A Mapping Between SNOMED-CT and the OXMIS Coding Scheme

Jeffery L. Painter, B.S.
GlaxoSmithKline, Research Triangle Park, NC

Abstract

The General Practice Research Database[1] (GPRD) contains a mixture of legacy data coded in both the Oxford Medical Information Systems (OXMIS) coding scheme and various versions of the Read codes. The state of GPRD today offers a unique opportunity to make use of the SNOMED CT coding scheme in order to integrate this mixed coding into a single view for all of the data contained therein. One mechanism we offer is to make use of the UMLS Metathesaurus to facilitate a mapping between OXMIS and the SNOMED CT nomenclature. Providing a map between Read is more easily handled since those codes have been "absorbed" into the SNOMED CT coding scheme itself.

Objective

Our mapping is a multi-stage process focused on providing a high degree of accuracy while minimizing the number of unmapped codes between these two schemas. This is facilitated through the use of a semantic similarity metric and association maps utilizing concepts found within the UMLS Metathesaurus[1].

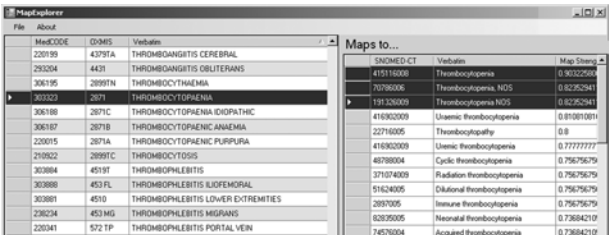
Our alignment is comprised of the following steps:

1. Direct Lexical
2. Plural Variants
3. Normalized Form
4. Stop-words removed
5. bi-gram comparison for probabilistic match

This procedure is focused on providing a direct map between a "foreign" scheme and SNOMED CT[2] directly for use in a relational database or similar application.

We use the ideas of semantic similarity and "concept" mapping to align two distinct coding schemes, namely OXMIS[3] and SNOMED CT via the Metathesaurus. This method allows one to progress in mapping two schemas by using only the verbatim representations given by the code/term pairs found within each.

While we strive to provide a fully automated map, this is not entirely possible. Consequently, a browser has been constructed to enable the clinician to explore our maps and determine with their medical expertise the "correct" association. When a direct map cannot be found, the probability (0-1 *with 1 being an exact match*) values are displayed and the results are ranked from most to least likely match.



	MedCODE	OXMIS	Verbatim
	220199	4379TA	THROMBOANGIITIS CEREBRAL
	293204	4431	THROMBOANGIITIS OBLITERANS
	306195	2899TN	THROMBOCYTHAEMIA
▶	303323	2871	THROMBOCYTOPAENIA
	306188	2871C	THROMBOCYTOPAENIA IDIOPATHIC
	306187	2871B	THROMBOCYTOPAENIC ANAEMIA
	220015	2871A	THROMBOCYTOPAENIC PURPURA
	210922	2899TC	THROMBOCYTOSIS
	303884	4519T	THROMBOPHLEBITIS
	303888	453 FL	THROMBOPHLEBITIS ILIOFEMORAL
	303881	4510	THROMBOPHLEBITIS LOWER EXTREMITIES
	238234	453 MG	THROMBOPHLEBITIS MIGRANS
	220341	572 TP	THROMBOPHLEBITIS PORTAL VEIN

	SNOMED-CT	Verbatim	Map Streng
	415116008	Thrombocytopenia	0.90322580
	70786006	Thrombocytopenia, NOS	0.82352941
▶	191326009	Thrombocytopenia NOS	0.82352941
	416902009	Uraemic thrombocytopenia	0.81081081
	22716005	Thrombocytopathy	0.8
	416902009	Uremic thrombocytopenia	0.77777777
	48788004	Cyclic thrombocytopenia	0.75675675
	371074009	Radiation thrombocytopenia	0.75675675
	51624005	Dilutional thrombocytopenia	0.75675675
	2897005	Immune thrombocytopenia	0.75675675
	82835005	Neonatal thrombocytopenia	0.73684210
	74576004	Acquired thrombocytopenia	0.73684210

Figure 1: Mapping of Thrombocytopaenia.

The example shown in Figure 1 demonstrates the relationship between the OXMIS code "2871" for Thrombocytopaenia and the resulting list of codes it is associated with in the SNOMED CT coding scheme. The first code "415116008" has an alternate spelling for the term, but the probability of match is quite high, 90.3%. Additional codes, such as "Thrombocytopenia, NOS" are also provided as possible candidates for this particular code.

Results

There are 16,920 unique OXMIS codes that actually appear in the coded data. These codes represent over 65.7 million data entries. The mapping procedure attained significant data coverage mapping approximately 96% of the records to SNOMED CT.

References

[1] U.S. National Library of Medicine, Unified Medical Language System, `http://www.nlm.nih.gov/research/umls/`.

[2] SNOMED CT is copyrighted by the International Health Terminology Standards Organization (IHTSDO)

[3] Perry J, ed: OXMIS Problem Codes for Primary Medical Care. Oxford Headington. 1978.

[1] General Practice Research Database is maintained by the (UK) National Health Service Information Authority

Achieving Consensus on a Common Vocabulary for Patient Health History and Exam Questions and Responses

A.W Forrey[1](forraw@u.washinton.edu), Peter Elkin[2], Gretchen Murphy[1]
[1]University of Washington, [2]Mayo Clinic

Keywords: Health History, Exam, Vocabulary

Abstract

This paper describes an approach to a vocabularies for the questions/responses for patient health histories and examinations that are applicable not only to present paper-based records but also to the Electronic Health Record (EHR) and across health specialty disciplines. It is also relevant to the interoperable use of the EHR across settings of care and is an initial step in the use of these vocabularies in implemented systems. The project uses the biomedical terminologic expertise present in the healthcare professional schools of the Washington, Wyoming, Alaska, Montana, Idaho (WWAMI) region served by the University of Washington Health Sciences Center. The intention is to develop cross-health specialty professional consensus on these two small specific vocabularies as an example of not only the process for arriving at common vocabularies for use in the evolution of the EHR and its support of the Basic Patient Care Scenario but also to deal with the practical problems of the implementation of the uses of such vocabularies in real healthcare enterprises and the transition to new information storage media. These vocabularies have been submitted for ASTM E-31 ballot for ASTM E-1633 Standard Specification for Coded Values to be Used in the Electronic Health Record and to be associated with specific data elements already identified in the E-1384 Standard Practice for the Structure and Content of the EHR. Other national/international informatics standards issues are recognized, including the potential applicability of SNOMED-CT.

Formal and continuing professional education aspects are also recognized and steps are taken to help address these problems. This effort contributes to the national common conventions for the EHR by providing an initial value set for the key data elements for patient assessment on the Patient Care Scenario described in E-1384. The Patient Care Scenario for use of the Electronic Health Record (EHR), or its current paper analogs, in medicine and dentistry identifies the capture/updating of the patient's health history as a key initial step of a healthcare encounter followed by capturing examination observations for use in patient assessment. The idea for a common specific vocabulary for both the questions asked in a patient health history and the responses to be given, as well as the examination observations made, has been openly discussed for some time. Both vocabularies are derived from widely used paper forms and organized by the categories for Body Systems already stated in E-1633 and can subsequently be expanded in stages to eventually serve all designated healthcare specialty disciplines with a common vocabulary usable in any medium and become a central resource for industrial Suppliers of informatics products and services. These Suppliers can then create implemented information architectural components to be used by Acquiring Healthcare Enterprises to build into their individual enterprise information architectures such that there is conceptual, as well as technical, commonality across all enterprises. The initial version could then be included in the work of various national/international Standards Developer Organizations (SDOs) and be iteratively extended and evolved.

Aligning the Top-Level of SNOMED-CT with Basic Formal Ontology

William R. Hogan, MD, MS
UPMC & Dept. of Biomedical Informatics of University of Pittsburgh, Pittsburgh, PA
hoganwr@upmc.edu

At UPMC, we are evaluating approaches to the use of terminologies and ontologies to achieve semantic interoperability among applications. As one part of comparing and contrasting SNOMED-CT (SNCT) with a realist ontology approach, we aligned the top-level concepts of SNCT within Basic Formal Ontology (BFO). We report our findings here.

INTRODUCTION

At UPMC, we intend to achieve semantic interoperability among our information systems. We and our technology partner, dbMotion, are evaluating various approaches to the use of terminologies and/or ontologies in support of this mission. We therefore sought to better understand SNCT,[1] BFO[2] and realist ontology in general, and their unique views of the world. To compare and contrast these views, we conducted the exercise of aligning the top-level concepts of SNCT within BFO.

MATERIALS

We downloaded the 01/2008 version of SNCT and BFO in .obo format from the BFO web site.

METHODS

The January, 2008 version of SNCT has 19 top-level concepts. We reviewed each one and either:

1. Equated it with an existing BFO class.
2. Gave it an is_a relationship to a BFO class.
3. Split it into two or more classes that we either equated with a BFO class or added to BFO.
4. Did not add it to BFO at all because either (a) it was too vague to serve as a synonym of a BFO term, (b) it did not represent anything in reality, or (c) it mixed epistemology with ontology.

To understand as completely as possible what each top-level concept meant, we studied their children and read the SNOMED-CT User Guide. The net result of this study was a new .obo file containing BFO extended with synonyms and new classes.

Note that the descendants of the 19 top-level concepts of SNCT do not necessarily follow their top-level ancestor into BFO either for the same reasons in #4 above or because they represent a different kind of BFO entity than their ancestor.

RESULTS

We arranged the 19 top-level concepts of SNCT into BFO as follows (Table):

SNCT Concept	Placement in BFO
Body structure	is_a independent continuant
Clinical finding	None (epistemology)
Environment or geographic location	Split: environment = niche, geographic location is_a spatial region
Event	Equate with event
Linkage concept	None (nothing in reality)
Observable entity	Equate with dependent continuant
Organism	is_a object
Pharmaceutical / biologic product	Split into three classes, each of which is_a object
Physical force	is_a dependent continuant
Physical object	Equate with object
Procedure	is_a process
Qualifier value	Equate with dependent continuant
Record artifact	is_a independent continuant
Situation with explicit context	None (ambiguous)
Social context	None (ambiguous)
Special concept	None (nothing in reality)
Specimen	is_a object
Staging and scales	Split into two classes, each of which is_a dependent continuant
Substance	Equate with substance

Table – Destination in BFO.

DISCUSSION

This exercise demonstrates that most (14/19 or 74%) of the top-level concepts of SNCT can be fitted into the framework of BFO, but only after significant reorganization. Five concepts did not fit into BFO for various reasons. One of the most important of these concepts is clinical finding, which is intended to comprehend diseases and signs and symptoms of disease. However, a finding of disease (epistemology) is not the same thing as a disease (ontology). Thus, this discrepancy between SNCT and BFO is important to consider further. Future work is to align the next level of SNCT (345 concepts) with BFO.

Acknowledgements

We thank Barry Smith for suggesting this study.

References

1. Donnelly K. SNOMED-CT: The advanced terminology and coding system for eHealth. *Stud Health Technol Inform.* 2006;121:279-290.
2. Grenon P, Smith B, Goldberg L. Biodynamic ontology: applying BFO in the biomedical domain. *Stud Health Technol Inform.* 2004;102:20-38.

Defined Representation of the Clinical Care of the Newborn Infant by SNOMED CT®

Andrew G. James MBChB MBI FRACP FRCPC[1], Kent A. Spackman MD PhD[2].
Departments of [1]Paediatrics, University of Toronto, Toronto, Ontario, Canada, and
[2]Medical Informatics and Clinical Epidemiology, OHSU, Portland, Oregon, USA.

Abstract: Defined SNOMED concepts enable and enhance subsumption, and the computation of equivalence between different methods of expressing the same meaning. The SNOMED CT® July 2006 release was systematically examined using the CliniClue browser to determine the degree of representation for a set of commonly used terms that are relevant for the clinical care of the newborn infant. There was defined representation for 24.4% of terms drawn from the categories of diagnosis, intervention, drug or observation. There was primitive representation for 62.0% of the terms, partial representation for 10.2% of the terms, and no representation for 3.4% of the terms.

Background: SNOMED CT®, the Systematized Nomenclature of Medicine Clinical Terms® (SCT), is a comprehensive, concept-based, clinical terminology with a semantic model based on description logic that uniquely identifies and describes clinically relevant concepts. SCT has a polyhierarchical structure with multiple parent-child relationships together with relationships between concepts in different subtype hierarchies that define the meaning of a concept relative to other concepts. SNOMED concepts must be sufficiently defined to enable subsumption, the computation of equivalence between different methods of expressing the same meaning, and other computational processes.

SCT has evolved from the Systematized Nomenclature of Pathology into its current form over the past forty years. The content of the terminology has been determined largely by the voluntary contributions of many, diverse clinical groups. An unforeseen consequence of this opportunistic evolutionary process may be that some unique, clinically relevant concepts of highly specialized clinical domains do not have defined representation within SCT.

The purpose of this study was to determine the degree of representation within SCT for a set of commonly used terms that are relevant for the clinical care of the newborn infant.

Methods: The SNOMED CT® July 2006 release was systematically examined using the CliniClue browser [version 2006.2.8, November 2006] to determine if the 881 elements within five sets of terms that are relevant for the clinical care of the newborn infant are represented in SCT.

The term sets were extracted from the Clinical Information Management System used in the Neonatal Intensive Care Unit at the Hospital for Sick Children, Toronto, Ontario, Canada. The sets of terms were categorised as diagnosis [disorder], intervention [procedure], drug and observation [finding]. The representation for each element within SCT was classified as defined [SNOMED concept sufficiently defined], primitive, partial or no representation.

Results: Overall, there is defined representation for 24.4% of the terms. There is primitive representation for 62.0%, partial representation for 8.6%, and no representation for 6.4% of the terms.

The diagnosis and intervention categories have the highest defined representation. The drug and observation categories have the lowest defined representation.

Table. Representation within SNOMED CT®

	A	B	C	D
Diagnosis	33.2	57.6	6.4	2.8
Intervention	50.0	30.9	14.6	4.6
Drug	0	75.2	21.2	3.6
Observation	9.3	80.2	6.4	4.1
All terms	24.4	62.0	10.2	3.4

where A is defined representation, B is primitive representation, C is partial representation, and D is no representation (*expressed as %)*

Conclusion: SNOMED CT® provides defined, structured representation based on description logic for only 25% of this set of terms that are used for the clinical care of the newborn infant.

References

College of American Pathologists, SNOMED International. SNOMED Clinical Terms® User Guide. July 2006 Release. Northfield, IL, USA; 2006.

CliniClue, Clinical Terminology Services. CliniClue Browser [version 5.2, November 16, 2006]. Available at http://www.cliniclue.com/

Evaluation of two French SNOMED indexing systems with a parallel corpus

Suzanne Pereira[a, b, c], Ph.D, Philippe Massari[a], MD, Antoine Buemi[d], MD, Badisse Dahamna[a] , Elisabeth Serrot[c], Michel Joubert[b], PhD, Stéfan J. Darmoni[a], MD-PhD.
[a] CISMeF, LITIS EA 4108, Institute of Biomedical Research, University of Rouen, France
[b] LERTIM, Marseille Medical University, France
[c] Vidal, Issy les Moulineaux, France
[d] APHP Hospital, Paris, France
suzanne.pereira@chu-rouen.fr

Abstract
Background: In this paper, we developed F–MTI (French Multi-Terminologies Indexer) which is a generic automatic indexing tool able to index documentation in several health terminologies such as SNOMED 3.5 (Internationale Systematized Nomenclature of human and veterinary MEDicine). Objective: We compared F–MTI and Snocode (a Canadian commercial tool) on a corpus of 100 discharge summaries. Results: The results showed that Snocode and F-MTI indexing are as close as two manual indexing can be. They also provided close results in terms of diagnosis.

Keywords :
Abstracting and Indexing/methods; Systematised Nomenclature of Medicine; medical records; international classification of diseases

INTRODUCTION

France chose in 2006 the SNOMED 3.5, the most exhaustive medical terminology in French, for medical record indexing. SNOMED 3.5 is included at 91% in the SNOMED CT (Clinical Terms). It contains 150,000 concepts distributed among eleven axes. The huge number of codes and the complexity of this terminology accounts for the reluctance of the physicians to index deseases in medical records. A computer support for this time-consuming procedure is then urgently required. We developped F-MTI [1] (French Multi-Terminologies Indexer) that generates a document indexing in all the implemented terminologies (MeSH (Medical Subject Heading) , SNOMED 3.5, ICD10 (Classification of Deseases) and CCAM (French CPT). Then all the terminologies are projected in the terminology(ies) desired by the user with the help of the mappings (most of them are coming from the UMLS). The goal of this study is to compare the SNOMED indexing of F-MTI and Snocode3 (a Canadian commercial tool [2] used in several hospitals in England, Canada and France).

MATERIALS AND METHODS

A corpus of 100 patient discharge summaries manually indexed in SNOMED 3.5 is difficult to obtain due to the complexity of this terminology. Faced with these facts, we projected the SNOMED codes to ICD10 codes that can be manually indexed and which enables to compare the two tools in terms of diagnosis. The projection process of SNOMED into ICD10 was performed by the same mapping. The ICD10 manual Indexing of these documents was taken as the reference. First, the two sets of SNOMED codes performed by F-MTI and Snocode were compared without any reference with simple measures. Then the two sets of ICD10 codes resulting from the projection of the SNOMED codes into ICD10 codes, were compared using an ICD10 manual indexing reference.

RESULTS & DISCUSSION

The results showed a Hooper's measure of 32.9 comparing the two sets of SNOMED codes. With the help of a SNOMED-ICD10 mapping we could compare in terms of diagnosis these two sets with a manual ICD10 indexing. We obtained a precision of 6.1 for Snocode and 4.4 for F-MTI and respectively a recall 24.7 and 27.0. Snocode and F-MTI indexing are as close as two manual indexing can be. They also provided close results in terms of diagnosis.

CONCLUSION

This is encouraging for our project. With some improvements we hope that FMTI will integrate a French electronic patient record system.

References

1. Pereira S., Névéol A., Massari P., Joubert M. and Darmoni S.J. Construction of a semi-automated ICD-10 coding help system to optimize medical and economic coding. MIE2006:845-850
2. SNOCODE® MedSight Informatique Inc. http://www.medsight-info.com

Indications Modeling and Coordination Challenges with SNOMED CT® Implementation

Cheryl Bunch, Pharm.D., Catherine Lai, Pharm.D., Toni Morrison, BSc., Danielle Przychodzin, Pharm.D., Raecine Chaney, LPN
Thomson Healthcare, Greenwood Village, Colorado, USA
cheryl.bunch@thomson.com

INTRODUCTION

As a major provider of integrated healthcare decision support solutions, Thomson Healthcare (TH) incorporates internal extensions of SNOMED CT® content, particularly to represent indications (SNOMED CT® Clinical finding hierarchy). The XML-based authoring units that house TH clinical content are heavily linked to SNOMED CT®. TH product releases occur as often as once daily, therefore twice-yearly SNOMED CT® releases are insufficient to meet our needs for timeliness and currency of data. A more or less granular concept may be needed to capture the clinical condition within TH content. TH developed and implemented a pilot indications model to assist with standardizing the way new internal SNOMED CT® extension concepts are represented in Health Language's CyberLE® tool.

METHODS/RESULTS

The pilot model specifies that relationships to native SNOMED CT® concepts should fully define the new internal extension concept when possible, while avoiding comorbidities (finding + finding) and "qualifier of qualifier" groupings (as in SNOMED CT® role groups). In contrast to SNOMED CT®'s primitive (i.e., partially defined) status, if an internal TH extension concept cannot be fully defined but is required for TH content, the concept is added with no coordinating relationships and a request for addition is submitted to IHTSDO. When native SNOMED CT® concepts requested for addition become available, TH retires the internal version of that concept and re-links content to the native SNOMED CT® concept. If IHTSDO declines to add the requested concept to SNOMED CT®, internally-created extension concepts are retained.

TH has created and subsequently retired over 200 internal SNOMED CT® extension concepts because of replacement with native SNOMED CT® concepts. We will compare the replacement native SNOMED CT® concepts' defining attributes to the coordinating relationships on a sample of 21 duplicated TH-created extension concepts. The sample includes: 5 fully defined and 16 primitive native SNOMED CT® concepts. TH matched SNOMED CT® in selecting the same parent concept in 4 cases (80%) and 13 cases (81%) in the fully defined and primitive subsets, respectively. TH used 2-3 coordinating relationships per concept, compared to 2-7 defining attributes per SNOMED CT® concept. About half (33/64) of SNOMED CT® defining attributes added no further definition beyond that of the concept's parent(s) (redundancy). The TH model considered 2/16 (12.5%) of the primitive SNOMED CT concepts fully defined with one matching TH's relationships exactly. Examples of differences in granularity between TH and SNOMED CT® will be presented. Limitations of this approach to modeling indications include: SNOMED CT®'s attribute set is modified frequently; TH has created internal relationship types unavailable within SNOMED CT®'s attribute set; delivery to and use of coordinating relationships by TH's end-user customers are still in the investigative stages.

CONCLUSION

The continuous analysis and improvement of TH's standardized indications modeling within SNOMED CT® will facilitate usage of TH's integrated healthcare decision support solutions by end-user customers and third-party vendors. Ongoing process refinements include reevaluation of existing coordinating relationships to account for new, retired or revised defining attributes within SNOMED CT® and formalizing the pilot model.

References

1. College of American Pathologists. SNOMED CT® User Guide – January 2007 release. International Health Terminology Standards Development Organisation. http://www.ihtsdo.org/our-standards/technical-documents/ . Accessed January 10, 2008.
2. Nachimuthu SK, Lau LM. Practical issues in using SNOMED CT as a reference terminology. Medinfo. 2007;12(Pt 1):640-4.
3. Schulz S, Hanser S, Hahn U, Rogers J. The semantics of procedures and diseases in SNOMED CT. Methods Inf Med. 2006:45(4):354-8.

Inductive Inference of a SNOMED CT Subset for Intensive Care Services

Jon Patrick[1], Yefeng Wang[1], Bahram Vazirnezhad[1], Alan Rector[2], Sebastian Garde[2], Jeremy Rogers[2] , Robert Herkes[3], Angela Ryan[3].

[1] Univ. of Sydney, [2] Univ. of Manchester, [3] Royal Prince Alfred Hospital

1. Abstract

A corpus of 44 million words of patient progress notes has been drawn from the clinical information system of the Intensive Care Service (ICS) at the Royal Prince Alfred Hospital, Sydney, Australia. The corpus has been processed by a variety of natural language processing procedures including the computation of all SNOMED CT candidate codes. The false positive error rate is estimated to be about 30%. The false negative rate is unknown but is believed to be of the order of 10% based on inspection of some of the texts. There are 13,136,022 concept instances making up 30,000 unique concept types detected in the corpus. These instances have been processed by a tool which computes the transitive closure of the concept types over the SNOMED hierarchy thus inferring the complete subset of SNOMED CT that would necessary for an ICS. A subset of 2718 concepts gives a coverage of 96% of the corpus but only needing to use less than 1% of all of SNOMED. This will give significant advantage to clinical information systems for efficiently delivering SNOMED terminology to the presentation interface.

2. Introduction

The RPAH-ICU corpus contains patients' daily medical measurements, conditions, care taken from the ward round. Notes were written by doctors and nurses. The notes were from year 2002 to 2006 amounting to a total of 461,969 notes for 12,076 patients. Tables 1 and 2 give a profile of the lexical distributions in the corpus and Table 3 gives the profile of SNOMED CT concepts identified in the corpus.

Table 1. The sizes of the total corpus and the token types

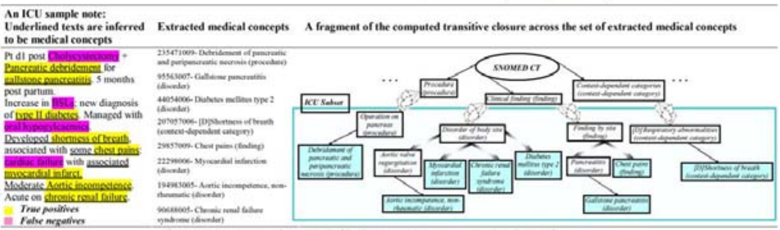
No. of token includes punctuation	44,072,299
No. of token type case sensitive	809,662
No. of token type case insensitive	703,198

Table 2. Frequencies of purely alphabetic tokens and their lexical validation.

Token Type	No. of token types	No. of tokens in corpus	%age
Alphabetic words	157,866	31,646,421	71.8%
Words in Moby English lexicon	32,081	28,095,490	63.7%
Words in SNOMED	22,421	29,008,594	65.8%
Words in UMLS (excludes SNOMED words)	25,956	27,893,156	63.3%
Words in SNOMED but not in Moby	5,005	1,985,391	4.5%
Words in either SNOMED or Moby	37,086	30,080,881	68.3%
Words in neither SNOMED nor Moby (Unknown words)	120,780	1,565,540	3.6%

Table 3. The frequencies of the SNOMED CT Categories.

CLASS ID	CLASS NAME	# of Concepts	%age
2	Clinical finding (finding)	3085935	23.5%
19	Substance (substance)	1799003	13.7%
1	Body structure (body structure)	1742270	13.3%
7	Observable entity (observable entity)	1609937	12.3%
12	Procedure (procedure)	1600392	12.2%
11	Physical object (physical object)	968856	7.4%
9	Pharmaceutical / biologic product (product)	853474	6.5%
15	Social context (social concept)	592940	4.5%
18	Staging and scales (staging scale)	474280	3.6%
3	Context-dependent categories (context-dependent category)	165741	1.3%
8	Organism (organism)	123683	0.9%
10	Physical force (physical force)	53439	0.4%
17	Specimen (specimen)	24204	0.2%
16	Special concept (special concept)	21525	0.2%
5	Events (event)	20343	0.2%



An ICU sample note: Underlined texts are inferred to be medical concepts

Pt d1 post Cholycystectomy + Pancreatic debridement for gallstone pancreatitis. 5 months post partum. Increase in BSLs, new diagnosis of type II diabetes. Managed with oral hypogylcaemics. Developed shortness of breath, associated with some chest pains: cardiac failure with associated myocardial infarct. Moderate Aortic incompetence. Acute on chronic renal failure.

🟨 *True positives*
🟪 *False negatives*

Extracted medical concepts

235471009- Debridement of pancreatic and peripancreatic necrosis (procedure)
95563007- Gallstone pancreatitis (disorder)
44054006- Diabetes mellitus type 2 (disorder)
207057006- [D]Shortness of breath (context-dependent category)
29857009- Chest pains (finding)
22298006- Myocardial infarction (disorder)
194983005- Aortic incompetence, non-rheumatic (disorder)
90688005- Chronic renal failure syndrome (disorder)

A fragment of the computed transitive closure across the set of extracted medical concepts

SNOMED CT
. . .
Procedure (procedure)
Clinical finding (finding)
Context-dependent categories (context-dependent category)
. . .
ICU Subset
Operation on pancreas (procedure)
Disorder of body site (disorder)
Finding by site (finding)
[D]Respiratory abnormalities (context-dependent category)
Debridement of pancreatic and peripancreatic necrosis (procedure)
Aortic valve regurgitation (disorder)
Myocardial infarction (disorder)
Chronic renal failure syndrome (disorder)
Diabetes mellitus type 2 (disorder)
Pancreatitis (disorder)
Chest pain (finding)
[D]Shortness of breath (context-dependent category)
Aortic incompetence, non-rheumatic (disorder)
Gallstone pancreatitis (disorder)

Figure 1. A sample ICU note (left), Extracted medical concepts (centre) and computing the transitive closure for the ICU subset (right)

3. Methods

A prototype strategy has been developed to asses the merit of the approach to the inductive inference of the SNOMED CT subset for ICUs from the texts of the patient notes. The steps are:

1. Identify all the SCT candidates in the clinical notes - this is the *extracted set*.
2. Construct a histogram of all the concept codes and separate them into SCT categories.
3. Import the code frequency tables of body structure, clinical finding and procedures into one table to be a fair sample of concepts.
4. From that extracted set, select the codes that were used at least 100 times (an arbitrary cut off point) - this is the *reduced set*.
5. Using appropriate software, compute the transitive closure across the set of extracted codes - this is the *closure set*.
6. Clean the closure set semi-automatically to remove anomalous concepts and to correct defective SNOMED modeling - this is the *clean closure set*.

Figure 1 shows an example clinical note transformed through steps 1 to 5. There are 4 false negatives: "Cholycystectomy" is "Cholecystectomy" in SCT, a spelling error in "hypogylcaemics", BSL is an unknown abbreviation in SCT, and cardiac failure is an incomplete expression in SCT used in a total of 27 different concepts.

4. Analysis

The *reduced set* comprises 2718 codes, and covers 6,177,077 of the 6,428,597 codable item instances summarised by the histogram reports, or 96.09% of all codable items. The remaining 3.91% of codable items (n=251,520) requires an additional 21,375 SNOMED codes.

Restricting the set to only the top 1000 codes by usage would cover 89.25% of the codable items in the RPAH-ICU corpus.

The list of 2718 codes was then run against a KRSS version of SNOMED circa Dec-06, using a segmenter to extract the minimal SCT fragment containing all 2718 codes, and a plugin to parse sct_concepts files for FSNs, taken from the 2008 release.

Note: A small number of the 2718 codes relate to SCT codes that were added to the international release after the date (Dec06) of the most recent KRSS version we had available on which to perform the segmentation.

5. Coverage of SNOMED CT

The subset uses 1 percent of SCT. The current SNOMED (Jul 2007 release) has 310,311 active concepts and 1,218,983 relationships, so 1,529,294 rows from two tables.

The subset covering 96% of everything codable in the RPAH_ICU corpus plus a skeleton of ancestors to wrap it in, contains only 7540 classes, 5600 subclass axioms and 1939 equivalent class axioms. This corresponds (more or less) to 15,079 rows from the full 1.5 million, or just less than 1% of the content.

6. Conclusions

It appears feasible to construct a useful subset for SCT using the clinical notes. This work is preliminary. The major project requires:

- a more reliable extraction from the source corpus with better orthographic corrections for better lexical verification.,
- identification of the transitive closure from the complete set of SCT categories,
- cleaning of the transitive closure of inappropriate concepts and poorly modeled concepts,
- implementation of the subset in an ICU clinical information system for testing its efficiency.

This research is sponsored by the Royal Prince Alfred Hospital - Intensive Care Service, Sydney, the Health Information Technology Research Laboratory, University of Sydney, Australia <www.it.usd.edu.au/~hitru>, and the Bio-Health Informatics Group, University of Manchester, UK <intranet.cs.man.ac.uk/bhig>

The University of Sydney

Representing and sharing knowledge using SNOMED
Proceedings of the 3rd international conference on Knowledge Representation in Medicine (KR-MED 2008)
R. Cornet, K.A. Spackman (Eds)

Inheritance of SNOMED CT Relations between Concepts by two Health Terminologies (SNOMED International and ICD-10)

Tayeb Merabti[1,2], Suzanne Pereira[1,2,3], Thierry Lecroq[1], MichelJoubert[2], Stefan Darmoni[1]
[1] CISMeF, Research Department, Rouen University Hospital,France &
TIBS, LITIS EA 4108, Institute of Biomedical Research, University of Rouen, France
[2] LERTIM, Medical University, Marseille, France
[3] Vidal, Issy les Moulineaux, France

Abstract

The situation of medical coding and medical economics is quite specific in France. Besides ICD-10, two specific terminologies are used: the International Nomenclature of Human and veterinary Medicine (SNOMED International) developed by the College of American Pathologists and CCAM[1]. This work aims at creating and optimizing inter and intra terminology relations between ICD-10 and SNOMED Int.

As 91% of SNOMED Int. preferred terms (PTs) and 87% of ICD-10 PTs are present into SNOMED CT, via the UMLS[2] Metathesaurus, we explore the automatic inheritance of SNOMED CT relations (Finding Site of, Associated Morphology, ...) by SNOMED Int. and ICD-10 terms.

Method

In a first step, we extracted all UMLS concepts linked by a SNOMED CT relation. For example, the two UMLS concepts C0000727, C0000726 are linked by the SNOMED CT relation "Finding Site of". In a second step, we mapped the SNOMED CT relations to two terminologies namely SNOMED Int. and ICD-10. As the terms of SNOMED Int and ICD-10 are also linked to the concepts found in step 1, we projected the relations found between the UMLS concepts to the ICD-10 and SNOMED Int. PTs. Finally, we obtained for each terminology a set of PTs pairs linked by the SNOMED CT relations.

Results

We found a total of 264,216 SNOMED International PTs pairs linked via a SNOMED CT relation and 6,417 ICD-10 PT pairs linked via a SNOMED CT relation. We also obtained 114,036 pairs of one ICD-10 PT and one SNOMED Int. PT linked by a SNOMED CT relation.

For example the ICD-10 term "Achondroplasia" (ICD-10 code: Q77.4) was linked according to the "Associated morphology" SNOMED CT relation to the term "Dysplasia, congenital" (SNOMED Int. Code: M-20020) and linked according to the "Finding Site of" SNOMED CT relation to the term "Bone"(SNMOED Int. code: T-11001).

Conclusion

We have several perspectives in mind. The first one is to apply the same methodology to another health terminology (the MeSH thesaurus) which is the current terminology used in CISMeF (a Web site dedicated to Catalog and Index Health Resources in French). The information retrieval algorithm using the relations that will be produced between MeSH terms could be very easily implemented in the CISMeF Web site, (e.g. the relation "Adams Stokes syndrome" "Finding site of" "heart conduction system" could lead to expand or limit the initial query "Adams Stokes syndrome").

Acknowledgements

This work is partially granted by the project Semantic Interoperability of terminologies in French Health Information Systems (French Acronym InterSTIS) project funded by the French National Research Agency (French acronym ANR).

Address for Correspondence

Tayeb Merabti, Medical Library-University Hospital,
1 Rue Germont, 76031 Rouen Cedex,
e-mail: tayeb.merabti@chu-rouen.fr

[1] The French equivalent to the US CPT4

[2] We exploited the UMLS 2007AB

LexValueSets: A Framework for Context-Driven Value Sets Extraction

Jyotishman Pathak, Guoqian Jiang, Sridhar O. Dwarkanath, James D. Buntrock, Christopher G. Chute
Division of Biomedical Informatics, Mayo Clinic College of Medicine, Rochester, MN, USA

INTRODUCTION

The main objective of modeling *value sets* is to specify a *concept domain* with certain *slots* or *attributes* of interest such that the *attribute-values* can be obtained from one or more terminologies of interest. Typically these values are extracted manually, thereby warranting the development of (semi-) automatic techniques. However, in general, this has been an unresolved issue in part due to the lack of (i) linkage to clinical context patterns that act as constraints in defining a concept domain, (ii) techniques for automatically analyzing membership of values to a particular concept domain, and (iii) approaches based on formal languages such as the Web Ontology Language (OWL). Toward this end, we propose a novel approach for context-driven (semi-) automatic value sets extraction and evaluation called LexValueSets.

METHODS

The crux of LexValueSets is to render the semantics of a concept domain using a formal model that takes into consideration various context patterns (e.g., location, time duration), specified typically by subject matter experts (SMEs), to drive the development of two complementary techniques for value sets extraction: *extensional* and *intensional*. The extensional approach comprises of an explicitly enumerated set of local terms, provided initially by SMEs, which correspond to an initial list of values for different slots of the concept domain, and are used for automatically extracting concepts from a particular terminology or a coding scheme. For example, given a concept domain pain in humans, the set of local terms for a slot location would comprise of hand, hip and other anatomical structures. The intensional technique, on the other hand, leverages the computable semantic definition of a concept domain to automatically identify relevant concepts for filling the slots. For example, the SNOMED CT concept “661005 jaw region structure” can be used to fill the *location* slot of pain since it is a finding_site for the SNOMED CT concept “274667000 jaw pain”. We implement a prototype by adopting the LexGrid terminology model (http://www.lexgrid.org) for both these approaches and provide preliminary evaluation based on SNOMED CT.

RESULTS

To evaluate our LexValueSets, we extracted three different values sets from the 20070731 version of SNOMED CT of UMLS (version 2007AC) for both the extensional and intensional techniques. For the extensional technique, the first value set (VS-EA) contains all the matching results (filtered only for the anatomical concepts of SNOMED CT) obtained directly from the lexical match between the local terms and SNOMED CT concepts, the second value set (VS-EB) contains additional child concepts of concepts contained in VS-EA, and the third value set (VS-EC) contains all concepts from VS-EB and additional concepts obtained through traversing the target concepts of all associations for each concept in VS-EB. On the other hand, for the intensional technique, we identified all the sub-concepts of the concept “22253000 pain” in SNOMED CT and extracted the target concepts of the association “363698007 finding site” as candidates for the value set (VS-IA) corresponding to the location slot. Similar to the extensional technique, additional value sets, VS-IB and VS-IC, were extracted by traversing the hierarchy.

	Intensional Approach		
	VS-IA (n=217)	VS-IB (n=24404)	VS-IC (n=26712)
Extensional Approach			
VS-EA (n=858)	23	811	829
VS-EB (n=17128)	136	17056	17099
VS-EC (n=25635)	215	24156	25606

Table 1: Overlap between extensional and intensional value sets

Table 1 shows that the number of overlapping concepts between VS-EA and VS-IA is 23 (accounting for about 3% of the concepts in VS-EA), whereas the number of overlapping concepts between VS-EA and VS-IB is 811, accounting for about 93% of the concepts in VS-EA. This result indicates that most concepts in VS-EA are more granular (i.e., closer to the leaf nodes in the SNOMED CT hierarchy) than those identified in VS-IA that are derived by the intensional approach. The number of overlapping concepts between VS-EC and VS-IC is 25606, accounting for about 99.9% of VS-EC and 95.8% of VS-IC. This result indicates that the coverage of the two value sets for both the approaches, once hierarchy traversal is employed, is almost same.

Semantic Interoprability of Legacy eHealth Systems using SNOMED

Mohammad H. Yarmand and Kamran Sartipi
Department of Computing and Software, McMaster University, Hamilton, ON, Canada

Abstract. *We introduce a framework for applying healthcare standards and clinical terminology systems to achieve semantic interoperability between distributed Electronic Medical Record (EMR) systems. We follow healthcare standards from HL7 [1] and Canada Health Infoway [2] Infostructure (EHRi) guidelines and documents in an integration project. This allows us to tackle the involved complexity and high technical requirements in order to provide guidelines for similar system integration projects. HL7 specifies the details of different healthcare scenarios by identifying the involved entities and required transactions and messages. Scenario information details and actual payload are then encoded into HL7 v3 message structure.*

Semantic interoperability

To achieve semantic interoperability, we map data fields of two healthcare systems onto the HL7 v3 clinical terms using three major Infoway documents: *Vocabulary Status Worksheet*, *Message Definition Worksheet* and *Scope & Package Tracking Framework*. The overall translation framework consists of three phases: *Interactions Extraction*, *Message Analysis*, and *Domain Analysis* to generate HL7 standard messages from typical healthcare scenarios. The legacy healthcare system provides healthcare scenarios and the framework generates the corresponding HL7 v3 messages that implement those scenarios.

Phase 1: interaction extraction. In this phase, HL7 standard interactions are developed through analysis of transactions. Given a scenario of legacy healthcare system, we divide it into smaller transactions required to complete a scenario. The legacy transactions are mapped onto the standard HL7 transactions which have similar semantics. Each transaction consists of a sequence of interactions to support outgoing and incoming communications.

Phase 2: message selection. In this phase, the elements of HL7 message structure (i.e., Transmission Wrapper, Trigger Event Control Act Wrapper and Message Payload) are created. Each interaction resulted form previous phase is assigned to a transmission wrapper schema. The transmission content is defined in this schema. The semantic of transaction content can be understood from the associated HL7 R-MIM.

Phase 3: domain analysis. In this phase, the final HL7 v3 message instance is generated from the message schema resulted form last phase. The HL7 domain that should be used for each field of schema is mentioned inside the schema. The clinical terminology system that should be used for each HL7 domain is defined in 'Vocabulary Status Worksheet'.

Parallel to the steps of the above phases, a mapping file should be generated that assigns data fields of legacy system to HL7 domains and the appropriate clinical terminology system concept defined by SNOMED. Using this mapping file and domains of HL7 schema extracted in the last step, another mapping file should be generated to map legacy system data fields to HL7 domain, HL7 message and the appropriate filed inside the message schema. Using this second mapping file, the pair <*legacy attribute, value*> can be translated by <*HL7 attribute mapped to legacy attribute, value*> inside the HL7 XML message.

Case study environment

As a case study, different algorithms (as proprietary services) of an existing research oriented Clinical Decision Support System (CDSS) have been provided for a Cardiac Rehab Center in another city (as client). The proposed framework has been applied on this integration project and the results are available. Our current research involves using Oracle's Health Transaction Base (HTB) [3] as the application development environment to develop and transfer HL7 v3 messages using Service Oriented Architecture (SOA).

References

[1] Health Level Seven official website. www.hl7.org.
[2] Canada Health Infoway. EHRS Blueprint, an interoperable EHR framework, April 2006.
[3] ORACLE. Oracle HTB datasheet, August 2005.

SNOMED Information Accessible to the Danish Community

Ulrich Andersen, CEO, MD MPA, Bente Maegaard, professor, Lina Henriksen, senior consultant, Anna Braasch, senior researcher, Lars Kayser, associate professor MD PhD
IHTSDO and University of Copenhagen, Denmark
uan@ihtsdo.org, {bente,lina,anna}@cst.dk, LK@sund.ku.dk

BACKGROUND AND AIM

Currently, citizens are increasingly taking responsibility for their own health through actively seeking information. The SNOMED CT database constitutes an excellent framework for the creation of an e-Health system with a natural language interface. This interface will translate citizens' natural language questions into structured database queries and return natural language answers. This poster presents the first steps of the development process with focus on a number of diabetes related questions.

QUESTION AND ANSWER

The innovative aspect of this approach is to develop a user friendly Question Answering (henceforth QA) system based on language technology methods and established database techniques (1).

The following example illustrates a QA session exploiting SNOMED terminology and relations: *Why do I feel increased thirst?* The answer to this question is a list of all diseases known by SNOMED and related to this symptom. A reduction of the list to the most likely diseases can be achieved by obtaining supplementary information from the user wrt. other experienced symptoms. The user might add: *I also have frequent urination, some weight loss and breathing difficulty.* The answer will be generated on the basis of pre-defined templates and will comprise the disease(s) showing all symptoms mentioned or as many of them as possible. For example, *You may suffer from a metabolic disorder or diabetes.*

QUESTION CLASSIFICATION

The SNOMED QA system must classify the natural language input wrt. the syntactic as well as the semantic dimensions. A syntactic classification of the question pattern concerns the grammatical structure and the type of question, such as *yes/no* and *wh*-type. The semantic dimension concerns mapping of the question to a predefined logical representation which will in turn be used as a basis for the generation of the database query which will retrieve the answer.

LOGICAL REPRESENTATION

The method selected for the logical language is based on the representation methods applied in AquaLog (2) and MOSES (3). These approaches take their starting point in relations, arguments and attributes. An example of a general template expressing arguments and a relation between them is: <?*wh*-> *exists(symptom, bodypart)*, which covers questions like the following. *Why do I have an ulcer on my foot?* The template will be instantiated with the question as follows: *<?why> exist(ulcer,foot).*

ONTOLOGY ANCHORING

One of the main challenges is to locate information relevant to the question within the SNOMED system. A feasible approach involves the combination of logical templates with specific query strategies. The query strategy invoked will point to one or more hierarchies, depending on the type, semantic content, explicitness and the complexity of the question. The question *Why do I have breathing difficulties?* could involve look-ups in the following hierarchies: OBSERVABLE ENTITY, BODY STRUCTURE and CLINICAL FINDING. The information retrieved from these hierarchies will reveal that *breathing difficulty* is a *disorder* in the *respiratory system*, and a list of *diseases* with this symptom can be generated.

PERSPECTIVES

It is expected that the QA system will be able to exploit SNOMED knowledge to provide advanced access to expert systems within the medical area.

References

1. Strzalkowski, T., Harabagiu, S. (eds.) *Advances in Open Domain Question Answering.* 2008.
2. Lopez Garcia, Vanessa, E. Motta, V. Uren. *AquaLog: An ontology-driven Question Answering System to interface the Semantic Web.* In: HLT Conference of the NAAcl, Companion Vol., 269-272, New York City 2006.
3. Pazienza M.T., A. Stellato, L. Henriksen, P. Paggio, F. Massimo Zanzotto. *Ontology Mapping to support ontology-based question answering:* In: Meaning workshop, 2005.

The Integration of SNOMED CT into the OpenMRS Electronic Medical Record System Framework

Thomas Meyer[1], **Chris Seebregts**[2], **Aurona Gerber**[1], **Arina Britz**[1], **Laurette Pretorius**[1,3], **Ronell Alberts**[1], **Deshen Moodley**[4]

[1] Knowledge Systems Group, Meraka Institute, CSIR, Pretoria, Gauteng, South Africa

[2] Informatics and Knowledge Management Directorate, Medical Research Council, Cape Town, South Africa

[3] School of Computing, University of South Africa, Pretoria, South Africa

[4] School of Computer Science, University of Kwazulu Natal, Durban, South Africa

Abstract

OpenMRS, an open-source framework used for the management of medical records, uses a basic concept dictionary to drive its data model. SNOMED CT is an ontology organising medical record content in order to provide a consistent mechanism to store, retrieve and use clinical data across specialties and sites of care. This poster proposes a project that aims to extend OpenMRS with the integration of SNOMED CT and related ontology technologies.

PROJECT DESCRIPTION

OpenMRS is an open-source application framework enabling the design and implementation of customizable medical records systems aimed primarily at medical informatics efforts in developing countries [1]. It is based upon an application developed by the Regenstrief Institute and Partners in Health based on experiences in Kenya, Haiti and Rwanda [2], and is currently implemented in countries such as Kenya, Rwanda, South Africa, Uganda, Tanzania, Zimbabwe, and Peru with scope to extend the adoption in multiple other locations throughout Africa [1]. It also claims nearly twelve million discrete observations collected for nearly 50,000 HIV patients with over 550,000 encounters in the AMPATH OpenMRS implementation in Kenya alone [1, 3]. It is implemented in Java and uses MySQL as database. The central data model is driven by a concept dictionary, which allows for the collection of coded, reusable data without requiring changes to the data model. The concept dictionary is a collection of coded, unique concepts used to generate forms and encode data that is captured within the system [4].

Generally, ontologies facilitate the structuring of information and data in a specific domain in such a way that systems can reason over it. Ontologies thus provide mechanisms that extend the representational and computational limits of traditional databases and other knowledge representation systems [5, 6]. Arguably, one of the most successful application area in this regard is the biomedical field, as witnessed, for example, by the widespread use of the medical ontology SNOMED CT [7] which addresses most areas of clinical information using a representation language that allows for computer processing [7].

The OpenMRS concept dictionary can be regarded as a crude ontology, and the extension thereof to use a formal ontology such as SNOMED CT with the associated technologies for the management and use of captured data, as well as for the generation of input forms, is the purpose of the project proposed by this poster. The project aims to investigate all aspects with regards to methodology, enhanced functionality and reasoning that can be the benefits of integration of SNOMED CT into OpenMRS.

Address for Correspondence

Tommie Meyer, Knowledge Systems Group,
Meraka Institute, CSIR, PO Box 395, Pretoria, 0001, South Africa,
tommie.meyer@meraka.org.za

References

[1] Openmrs.org. `http://openmrs.org/`, 2008.

[2] Regenstrief institute, inc. `http://www.regenstrief.org/`, 2008.

[3] Ampath. `http://medicine.iupui.edu/kenya/hiv.aids.html`, 2008.

[4] Openmrs concept dictionary. `http://openmrs.org/wiki/Concept_dictionary/`, 2008.

[5] T.R. Grüber. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5:199–220, 1993.

[6] Udo Hahn and Stefan Schulz. Ontological foundations for biomedical sciences. *Artificial Intelligence in Medicine*, 39(3):179–182, 2007.

[7] Snomed. `http://www.ihtsdo.org/`, 2008.

The Resource Impact of Supporting SNOMED CT® Updates in a Clinical Decision Support System

Catherine Lai, Pharm.D., Danielle Przychodzin, Pharm.D., Cheryl Bunch, Pharm.D., Toni Morrison, BSc, Raecine Chaney, LPN
Thomson Healthcare, Greenwood Village, Colorado, USA
catherine.lai@thomson.com

INTRODUCTION

Thomson Healthcare (TH) is a major provider of integrated healthcare decision support solutions. The XML-based authoring units housing TH's clinical content are heavily linked to SNOMED CT®. As product releases occur as often as once daily, it is imperative that the most current standards available are supported in the content; this includes supporting SNOMED CT® Updates. The purpose of this poster is to outline the current TH SNOMED CT® Update process and to identify the resources employed to support the July 2007 SNOMED CT® Update.

METHODS/RESULTS

TH editors use XML-based authoring tools to create indexed disease, lab and drug indications content using SNOMED CT®. Which SNOMED CT® codes are used depends upon the content type; however most are linked to codes in the Clinical finding and Procedure hierarchies. While the rates of change within SNOMED CT® vary between releases, the average change in the number of active concepts per release was 4093 between Jan 2002 and Jan 2005. Over 228,000 instances of SNOMED CT® references currently exist in TH content. When evaluating the resource need for supporting an update, scale is a major issue and the resource requirements vary based upon the scope and complexity of the changes made within SNOMED CT® between versions.

The primary goal of the update process is to harmonize current SNOMED CT® and extension data with the most recent version of SNOMED CT®. The main process phases include: Analysis, Issue Resolution and Implementation.

During Analysis the following are identified: changes to SNOMED CT® concepts referenced in TH content, duplications between new SNOMED CT® and TH created extension concepts, and changes to SNOMED CT® concepts with relationships to TH extension concepts. For the July 2007 SNOMED CT® update, 1 Clinical Terminology Specialist (CTS) completed the analysis over 172 hours with support from an Information Technology Specialist (ITS). Overall, 285 concepts were identified for review.

In the Resolution Phase, CTS identify and coordinate the concept resolutions, editors review the proposed resolutions and ITS implement the changes in a development environment. Resolutions may include: re-referencing content to an active concept, retiring a TH extension, or creating a TH extension. Extensive quality assurance is completed in multiple development environments over a 7 day period to ensure the desired outcomes are achieved. For the July 2007 update, the Resolution Phase utilized 4 full-time CTS, 2 full-time ITS and editors representing 1750 man-hours over 7.5 weeks.

In the Implementation Phase, the new version of SNOMED CT® and the changes to the TH extension concepts are deployed into the production environment. Authoring unit changes are also implemented at this time. Comprehensive quality assurance processes are executed. For the July 2007 update, implementation and quality assurance occurred over 3 days and 2 days, respectively. This represented 420 man-hours involving 10 editorial consultants, 4 full time CTS and 4 full-time ITS.

CONCLUSION

Supporting SNOMED CT® updates requires extensive resources ranging from CTS to ITS to editorial support. The July 2007 TH SNOMED CT® Update required over 2300 man-hours during a 3 month time period. As the TH processes improve, the anticipated the time required to support an update will decrease, however this will continue to depend upon the scale of changes SNOMED CT® implements.

References

1. Spackman KA. Rates of change in a large clinical terminology: three years experience with SNOMED Clinical Terms. AMIA 2005 Symposium Proceedings. 2005:714-18.
2. Nachimuthu SK, Lau LM. Practical issues in using SNOMED CT as a reference terminology. *Medinfo*. 2007;12(Pt 1):640-4.

Unification of Electronic Patient Data in a Commercial Health Information System through Multi-dimensional Semantic Annotation

Isabel Barth, MA, Sven Tiffe, PhD MS, Evgueni Loukipoudis, PhD MS
Agfa HealthCare, Trier, Germany
Isabel.Barth@agfa.com

INTRODUCTION

To facilitate reusability of electronic patient data in Orbis, a complex hospital information system, LexGrid[1] has been used as a registry for elements of electronic health records and for their semantic annotation.

Orbis allows its users to define an arbitrary amount of data structures in addition to the system's central data model by providing a visual tool for the definition of forms. Such forms correspond to data containers stored in a generic, entity-attribute-value (EAV) data model. They are composed of input fields that can be defined from a set of possible data formats, such as check boxes, free text fields or lists.

This results in a very flexible system for clinical documentation that empowers clinicians to define the same clinical concept in multiple ways. The classification of such customized data structures is a major challenge as fields are identified by unique identifiers based on text labels.

METHODS

The data is classified in two phases: In the first step, the structural system information, such as form containers and their elements, is projected to a proprietary hierarchic terminology in a LexGrid compliant format. Each element corresponds to an individual concept whose code is derived from the data's unique database identifiers. The context of the data entry forms is preserved by part-whole relations to their individual form elements. In a second step, these structural concepts are semantically annotated by links to terms from controlled clinical terminologies, such as SNOMED CT.

This procedure is supported by concept properties that contain additional information about the system information, such as descriptions and data types. During semantic annotation, this information can be used to ease the challenge of finding appropriate codes from clinical terminologies that represent the meaning of the corresponding system information.

Such fine-granular semantic annotation yields unification of clinical data and facilitates interoperability and retrievability.

For example, structurally unrelated heterogeneous information – such as data denoting a particular patient's height which may exist in the scope of different forms – become isomorphic on the clinical semantic layer through their common SNOMED parent concept (see figure 1).

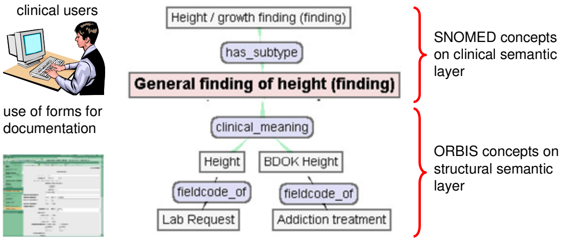

clinical users
use of forms for documentation
Height / growth finding (finding)
has_subtype
General finding of height (finding)
clinical_meaning
Height
BDOK Height
fieldcode_of
fieldcode_of
Lab Request
Addiction treatment
SNOMED concepts on clinical semantic layer
ORBIS concepts on structural semantic layer

Figure 1 - Unification of structural data in Orbis

OUTLOOK AND RESULTS

Using LexGrid's interterminological links, the resulting semantic network is extensible by further dimensions, such as Orbis patient data or higher semantic levels representing a Reference Information Model (RIM). The RIM elements will be composed of concepts from the intermediate clinical level. To ensure semantic integrity during concept selection, both the structural context provided by the Orbis terminology and the domain knowledge inherent in SNOMED relations must be considered.

Compared to our first approach[2], which required a semantic extension to the clinical database, the integration of structural system data into a semantic network of clinical terminologies enabled us to add semantics to a running hospital information system while avoiding changes to its core persistence structure. One of the major challenges is the practicability of semantic annotation in a real-life scenario which requires human expertise and performant tools to support the annotation process.

REFERENCES

[1] http://informatics.mayo.edu/LexGrid. Accessed: 01/21/2008.

[2] Barth, I Tiffe, S Dahlweid, M: Semantic Annotation of Patient Data in a Commercial Health Information System. Presented at MEDINFO 2007, Aug 20-24, Brisbane, AU.

Using SNOMED for Reference Hierarchies in the CPT Data Model

Marjorie Rallins, DPM, Eric Mays, PhD
American Medical Association, Chicago, IL
marjorie.rallins@ama-assn.org, emays@mays-systems.com

INTRODUCTION

CPT® is a HIPAA standard terminology widely used in the United States for reporting medical services and procedures, supporting administrative functions such as claims processing and medical guideline review. One characteristic of CPT is that the descriptions of procedures and services are provided at a level of generality required to make distinctions for the purposes of reporting. As electronic patient records receive greater adoption, it is desirable to record services and procedures at a detailed clinical level, requiring greater specificity than may be available in CPT.

One means to facilitate interoperation between CPT and more clinically specific terminologies is to provide a mapping. If clinical services and procedures are captured in SNOMED, for example, a mapping may be utilized to provide one or more relevant CPT codes to be selected for reporting purposes. This should be a familiar approach to those with knowledge of the SNOMED to ICD-9-CM maps. Indeed such a SNOMED to CPT map is presently under construction in a collaboration of the AMA and SNOMED Terminology Solutions, a division of the College of the American Pathologists.

In the CPT Data Model, now available as the CPT Developer's Toolkit, a description logic model of CPT is being developed which utilizes SNOMED concepts in the construction of the CPT reference hierarchies. This complementary approach to foster clinical specificity raises some interesting possibilities for recording of clinical data, enabling a flexible post-coordination approach in CPT.

CPT DATA MODEL

One of the goals of the CPT Data Model is to facilitate search and navigation. CPT has been primarily distributed in book format organized as chapter, section, sub-section, etc. with compactness of the printed form a key consideration. The CPT Data Model improves on the flat file CPT electronic distribution and printed materials by providing free standing descriptions for the headings, incorporating additional levels of organization, and several other improvements which are beyond the scope of this presentation. The hierarchical organization of the data model follows the book layout in order to provide access to those familiar with the book and also to maintain consistency with the history of the editorial process. Adopting a description logic model for CPT along with utilization of SNOMED for reference hierarchies enables alternative means of access and organization. This is especially relevant for guideline and utilization review where it is desirable to aggregate all procedures which use a certain device, for example.

REFERENCE HIERARCHIES

The CPT Data Model incorporates thirteen different reference hierarchies, such as Anatomic Site, Device, and Specimen which are generally defined as a union of one or more SNOMED sub-hierarchies. For example, the members of the Anatomic Site reference hierarchy are the sub-concepts of the Body Structure concept in SNOMED. Some reference hierarchies, specifically Complexity (e.g. High complexity decision making) and Patient Type (e.g. Established patient) have no SNOMED correlates and are defined as an enumeration of concepts in a CPT extension. Several reference hierarchies involve more than one set of SNOMED sub-concept hierarchies. The Approach reference hierarchy in CPT includes the SNOMED sub-hierarchies of Procedural Approach, Surgical Access Values, and Relative Sites. While a single more general concept could be chosen, there is significant value in constraining the reference hierarchies as specifically as possible.

The role relationships in the CPT Data Model align with the reference hierarchies and their ranges are constrained by a reference hierarchy. This approach has proven to be very useful during QA, and will have even greater value during the current editorial cycle as we surface these constraints in the Protégé OWL editing tool used for modeling by subject matter experts.

DISCUSSION

SNOMED has proven to be a rich source for defining the reference hierarchies in CPT, with the subsumption semantics having clear benefits for QA and the editorial process. We are providing the CPT specific additions to the IHTSDO for incorporation into future releases of SNOMED.

Representing and sharing knowledge using SNOMED
Proceedings of the 3rd international conference on Knowledge Representation in Medicine (KR-MED 2008)
R. Cornet, K.A. Spackman (Eds)

Using SNOMED to Normalize and Aggregate Drug References in the SafetyWorks Observational Pharmacovigilance Project

Gary H. Merrill, Ph.D., Patrick B. Ryan, M.Eng., Jeffery L. Painter, B.S.
GlaxoSmithKline, Research Triangle Park, North Carolina

The SafetyWorks project at GlaxoSmithKline developed an integrated set of methodologies to support the use of large observational data sources for monitoring and assessing drug safety. Here we focus on the SafetyWorks drug ontology, its construction and annotation, and its role in normalizing drug references across disparate data sources.

FDA “Guidance for Industry Good Pharmacovigilance Practices and Pharmacoepidemiologic Assessment” [1] characterizes pharmacovigilance as “all scientific and data gathering activities relating to the detection, assessment, and understanding of adverse events.” SafetyWorks is an integrated system for leveraging the use of observational data in such pharmacovigilance activities. Figure 1 illustrates the extraction of raw data from the GlaxoSmithKline Healthcare Information Factory (a repository of large disparate databases), the normalization and aggregation of the raw data by means of medical condition and drug ontologies, and the use of this normalized and aggregated data in observational screening and observational evaluation.

Observational screening applies an unmatched cohort design to provide a framework and context in which all relations amongst drugs and conditions can be explored. It should be considered as a hypothesis-generating step that may facilitate the identification and prioritization of drug-condition pairs warranting further evaluation. Observational evaluation is an analysis targeted at specific hypotheses, and it provides a robust estimate of the strength of drug/condition associations within the population of interest.

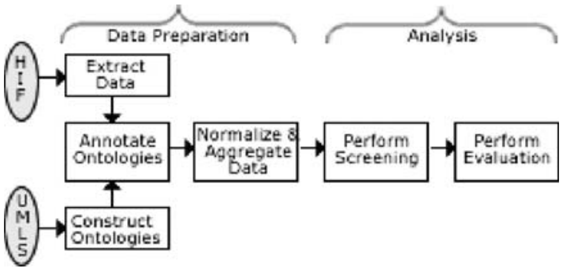


Figure 1: The SafetyWorks Process

The SafetyWorks drug ontology is constructed and annotated in a sequence of steps:

- The sub-hierarchy of SNOMED CT whose root is *Drug or medicament* is extracted from the UMLS Metathesaurus ([2]).
- RxNorm is employed to extend this hierarchy to include branded drugs by grafting branded drug nodes onto their generic forms in SNOMED CT.
- The extended ontology is then annotated with drug references from the set of observational data sources (electronic health records and insurance claims databases).
- The ontology is finally simplified by (a) pruning unnecessary ”forms” of drugs, (b) ensuring that no drug reference annotates both a node and an ancestor of that node, and (c) creating ”generic product” nodes to ensure that annotations are made only to the lowest level of the ontology.

The resulting drug ontology is an extended modified version of the SNOMED CT *Drug or medicament* hierarchy containing 16,100 categories; and it is employed as illustrated in Figure 1 to ”normalize” all drug references in each of the data sources to branded or generic drug categories. On the basis of that normalization, “aggregated drug eras” (periods of time during which a patient has continuously taken a particular drug) are created from the raw data, a similar process is used with our MedDRA-based medical conditions ontology to create “aggregated condition eras” (periods representing common episodes of care for the same medical condition), and observational screening and evaluation are applied to the resulting normalized and aggregated data.

References

[1] U.S. Food and Drug Administration, Guidance for industry: good pharmacovigilance practices and pharmacoepidemiologic assessment, March 2005, http://www.fda.gov/cder/guidance/6359OCC.htm.

[2] U.S. National Library of Medicine, Unified Medical Language System, http://www.nlm.nih.gov/research/umls/.