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)

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.

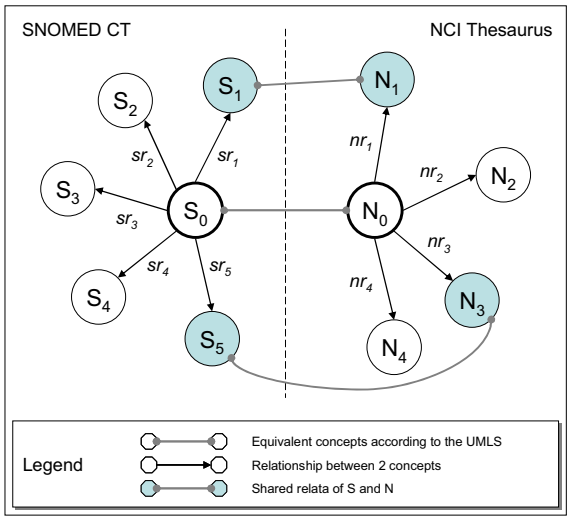

Figure 1. Graph formed by the related concept of one pair of equivalent concepts (S_0, N_0)

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™ (Interactive Tucana Query Language™), 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 equi-

valence 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	median
			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

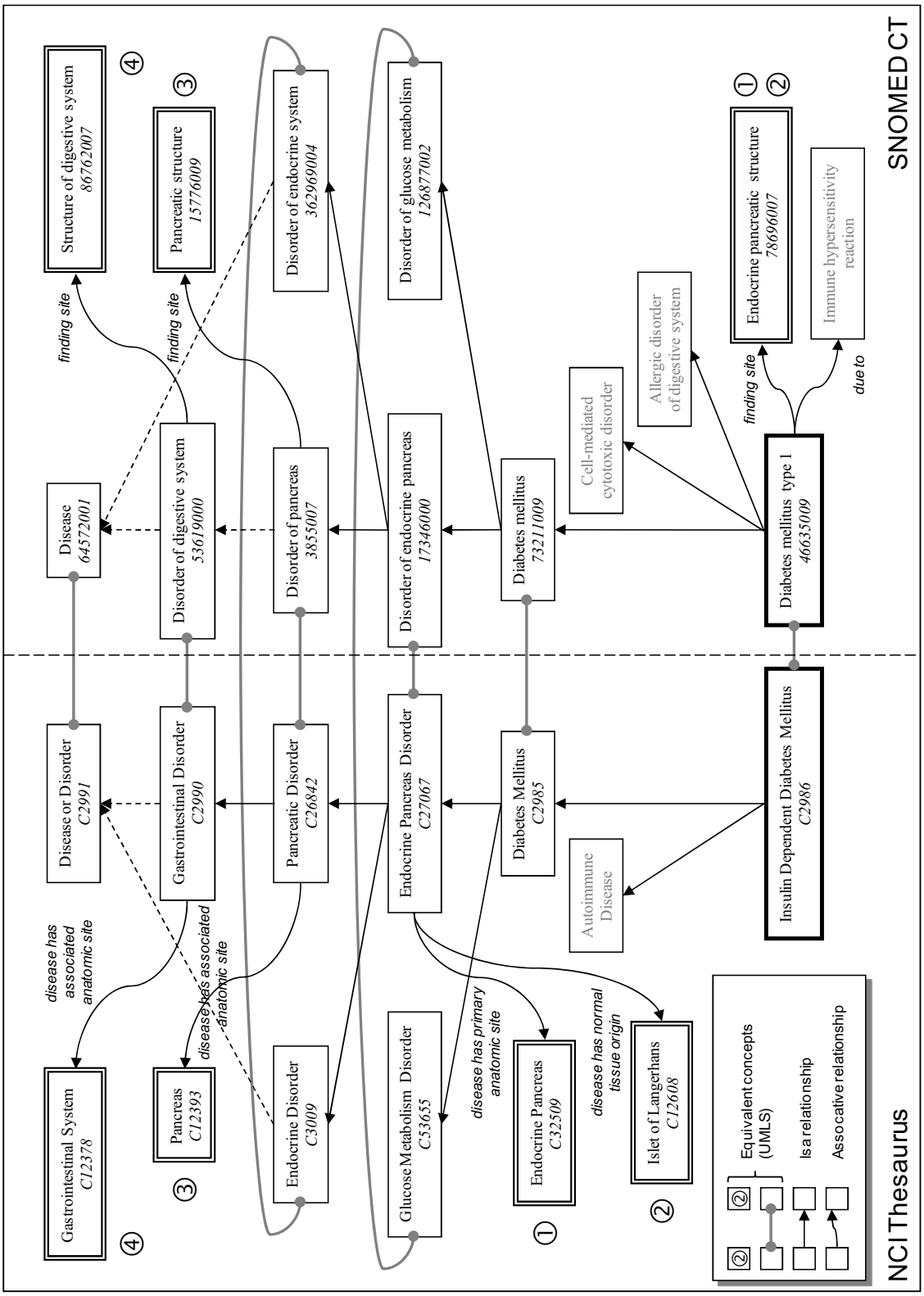

Figure 4. Representation of *Type 1 diabetes mellitus* in SNOMED CT and NCIt, showing shared relata for ancestors and associative relationships