Journal of Biomedical Informatics 42 (2009) 426–439

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Auditing associative relations across two knowledge sources

Lowell T. Vizenor [a], Olivier Bodenreider [b,*], Alexa T. McCray [c]

[a] Computer Task Group, Inc., Buffalo, NY, USA
[b] Lister Hill National Center for Biomedical Communications, National Library of Medicine, 8600 Rockville Pike – MS 3841 (Bldg 38A, Rm B1N28U), Bethesda, MD 20894, USA
[c] Harvard Medical School, Boston, MA, USA

ARTICLE INFO

Article history:
Received 16 June 2008
Available online 23 January 2009

Keywords:
Biomedical terminologies
Associative relationships
Auditing methods
Unified Medical Language System (UMLS)

ABSTRACT

Objectives: This paper proposes a novel semantic method for auditing associative relations in biomedical terminologies. We tested our methodology on two Unified Medical Language System (UMLS) knowledge sources.

Methods: We use the UMLS semantic groups as high-level representations of the domain and range of relationships in the Metathesaurus and in the Semantic Network. A mapping created between Metathesaurus relationships and Semantic Network relationships forms the basis for comparing the signatures of a given Metathesaurus relationship to the signatures of the semantic relationship to which it is mapped. The consistency of Metathesaurus relations is studied for each relationship.

Results: Of the 177 associative relationships in the Metathesaurus, 84 (48%) exhibit a high-degree of consistency with the corresponding Semantic Network relationships. Overall, 63% of the 1.8 M associative relations in the Metathesaurus are consistent with relations in the Semantic Network.

Conclusion: The semantics of associative relationships in biomedical terminologies should be defined explicitly by their developers. The Semantic Network would benefit from being extended with new relationships and with new relations for some existing relationships. The UMLS editing environment could take advantage of the correspondence established between relationships in the Metathesaurus and the Semantic Network. Finally, the auditing method also yielded useful information for refining the mapping of associative relationships between the two sources.

Published by Elsevier Inc.

1. Introduction

1.1. Objectives

The general framework of this study is the development of a methodology for the auditing of associative (or non-hierarchical) relations[1] in large biomedical terminologies for completeness and accuracy. Most research on terminology/ontology auditing focuses primarily on evaluating terminologies with respect to their hierarchical structure [1–8]. This is not surprising, since the backbone of most biomedical terminologies is the *isa relationship* [9,10] (and, to a lesser extent, the *part_of* relationship [11,12]). Still, some terminologies also contain associative relationships such as *treats* and *causes* that cut across the hierarchical structure of a given terminology [13]. What is more, relationships such as these may be found in relations expressing significant biomedical knowledge that cannot always be captured strictly in terms of hierarchical relations. So, while hierarchical relationships in terminologies warrant a great deal of interest, insufficient attention has been paid in the terminology literature to associative relations [14], perhaps because the methods used for auditing associative relations in terminologies are not as well understood as those used for auditing hierarchical relations.

This paper proposes a novel semantic method for auditing associative relations in biomedical terminologies. We tested our methodology on two Unified Medical Language System (UMLS) knowledge sources. Our motivation in undertaking this work in the context of the UMLS is to help achieve greater consistency between the Metathesaurus and the Semantic Network. We have done this by providing a framework for auditing associative relations in these two knowledge sources.

1.2. Overview of the UMLS Metathesaurus and Semantic Network

In this study, we use the Unified Medical Language System (UMLS) as a test bed for developing a methodology for auditing

* Corresponding author. Fax: +1 301 480 3035.
E-mail address: olivier@nlm.nih.gov (O. Bodenreider).

[1] Biomedical terminologies and ontologies can be represented as directed graphs in which nodes represent concepts (e.g., the organ *kidney* and the disease *nephroblastoma*). Throughout this paper, we use *relationship* to refer to the links among concepts in ontologies (e.g., *location_of*). In contrast, we use *relation* to refer to the association between two concepts linked by some relationship (e.g., "*kidney location_of nephroblastoma*"). In the literature, relationships are sometimes also called predicates, whereas relations also correspond to assertions, facts and subject-predicate-object triples.

1532-0464/$ - see front matter Published by Elsevier Inc.
doi:10.1016/j.jbi.2009.01.004

associative relations. The UMLS Metathesaurus contains some 1.5 million concepts derived from close to 150 biomedical and health related terminologies [15,16]. The Metathesaurus is not intended to represent a single consistent view of the world of biomedicine but rather to preserve the many views represented in its source vocabularies [17]. The UMLS Semantic Network, on the other hand, consists of 135 semantic types and 54 relationships and is intended to provide a consistent categorization of all concepts represented in the UMLS Metathesaurus [18]. The Semantic Network presents a high-level view of the world of biomedicine that is sufficiently general to categorize a wide range of terminologies in multiple domains. Two single-inheritance hierarchies, one for entities and another for events, make up the Semantic Network. The 135 semantic types are linked together through the *isa* relationship and form a hierarchy that allows semantic types to inherit properties from higher-level semantic types. In addition to the *isa* relationship, there is a set of 53 associative (or non-hierarchical) relationships in the Semantic Network, grouped into five major categories: 'physically related to', 'spatially related to', 'temporally related to', 'functionally related to' and 'conceptually related to'. The Semantic Network relations in which these relationships participate represent general, high-level biomedical knowledge, such as Body Part, Organ or Organ Component *location_of* Neoplastic Process.

In the UMLS, semantic types are used to categorize concepts in the Metathesaurus through categorization links assigned by the UMLS editors. That is, every Metathesaurus concept is assigned to at least one semantic type, independently of its hierarchical position in a source vocabulary. Fig. 1 shows the two-level structure of the UMLS. The rationale for this two-level structure is to provide a uniform semantics to the concepts regardless of the particular structure of the source vocabulary [19]. At the Metathesaurus level, there are a number of relations among concepts (derived from the individual source vocabularies), such as "*kidney location_of nephroblastoma*." However, unlike the categorization link between Metathesaurus concepts and semantic types, there is no direct link between the Metathesaurus relationships and Semantic Network relationships. One consequence of this is that it is difficult to provide a uniform semantics between the Semantic Network relationships and the Metathesaurus relationships. As illustrated in Fig. 1, one auditing method for the UMLS is to simply check the compatibility between a relationship asserted between two concepts in the Metathesaurus and the possible relationships defined in the Semantic Network between the semantic types of these two concepts. Intuitively, the Metathesaurus relationship is expected to be either equivalent to or more specific than the Semantic Network relationship. However, since no equivalence or subproperty associations are defined between relationships across the two levels of the UMLS, validation on a large scale is not easily accomplished.

Finally, the Semantic Network possesses an additional layer of structure in the form of fifteen high-level semantic groups, which are a coarse-grained set of semantic type groupings designed using the following principles: semantic validity, parsimony, completeness, exclusivity, naturalness and utility [20]. The semantic groups are useful in a number of applications including improved visualization [21] and (as we suggest in this paper) relation auditing.

1.3. Principles for auditing associative relations

1.3.1. Formal methods for auditing associative relations

In order to handle the size and complexity of terminologies, methods based on description logic have been developed to audit large biomedical terminologies—i.e., to verify and maintain (logical) consistency and semantic correctness of their contents [22–26]. For the most part, these studies have focused primarily on concept hierarchies. That said, there exist description logic-based tools such as Protégé-OWL that possess the capabilities to audit relations along the lines of the principles we lay out below. For this study, some thought was given to using a description logic-based approach to auditing Semantic Network relations, but we determined that the source materials used were not amenable to strict, logic-based approaches. The reason for this is that the UMLS contains a diverse range of biomedical terminologies and coding systems not all of which are suited for logic-based approaches [27], so our challenge was to develop a method for auditing sources that would approximate many of the features of the logic approaches.

1.3.2. Relationship signatures

Our auditing method takes advantage of the formal notion of a relationship signature defined in [28, pp. 478–480] as a key element of Sowa's conceptual graphs. For the purposes of this study, a relation can be thought of as a subject-predicate-object triple, where the predicate is a relationship such as *treats* that relates the subject of the relation to its object. For example, in the relation (**Pharmacologic Substance**, *treats*, **Pathologic Function**), *treats* is the relationship, **Pharmacologic Substance** is the subject and Pathologic Function is the object.

In order to identify inconsistencies in these relations, a relationship signature is introduced for each relationship that specifies what types of biomedical entities can be related to one another via a given Semantic Network relationship. In this paper, we take

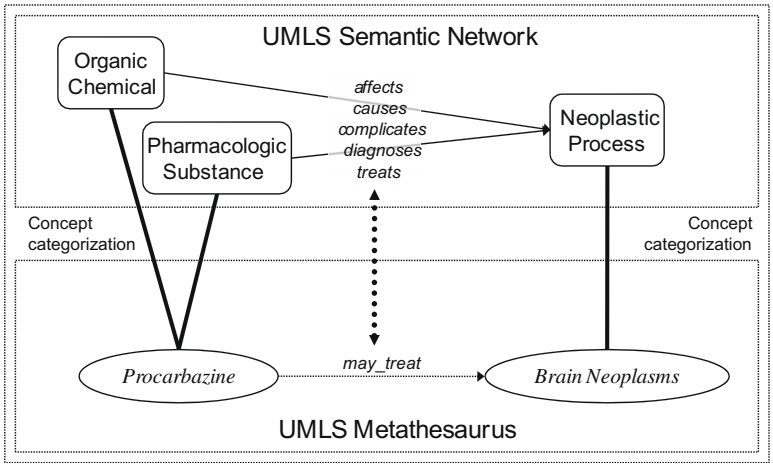

Fig. 1. Two-level structure of the UMLS. Each Metathesaurus concept is assigned to one or more semantic types from the Semantic Network. Relationships are inherited from the Semantic Network and indicate possible relationships between concepts.

advantage of the fact that every semantic type in the Semantic Network is a member of a semantic group and use these semantic groups to define the signatures of each Semantic Network relationship.[2] Relationships may have more than one semantic group signature.

The use of relationship signatures here is similar to the use of domain and range statements in formalisms such as RDFS (Resource Description Framework Schema) [30] and OWL (Web Ontology Language) [31]. For a given predicate (or what we call a relationship), it is possible in RDFS to declare the class of the subject (i.e., domain) and the class of the object (i.e., range) for any triple in which that property is a predicate. Nevertheless, these formalisms are too strong for our purposes. In RDFS/OWL, domain and range declarations are used to draw inferences about the values of the subject and object of a triple.

In contrast, we use relationship signatures as constraints. In other words, relationship signatures are used to simply identify whether or not a given Metathesaurus relationship is consistent with a given Semantic Network relationship. From the point of view of this audit, in order for a Metathesaurus relationship to be *consistent* with the corresponding Semantic Network relationship, it is *necessary* that there be a match between their signatures. Conversely, for a Metathesaurus relationship to be *inconsistent* with the corresponding Semantic Network relationship it is *sufficient* that there be no match between their signatures.

1.3.3. Relationship hierarchies

Just as it is possible to organize concepts into hierarchies, so too is it possible to organize relationships into hierarchies. In the case of a concept hierarchy, one concept, c_1, is a subclass of (i.e., is more specific than) another concept, c_2, only if every instance of c_1 is necessarily an instance of c_2. For example, in the Semantic Network, Human is a subclass of Mammal, which means that every instance of Human is necessarily an instance of Mammal. Relationship hierarchies can be defined in a similar fashion. For example, if we assert that *treats* is a subproperty of *affects* and c_1 *treats* c_2 then necessarily c_1 *affects* c_2.

When mapping Metathesaurus relationships to Semantic Network relationships, we established equivalence and subproperty associations between a given Metathesaurus relationship and the corresponding Semantic Network relationship. Because we use signatures based on mutually exclusive semantic groups to represent the domain and range of these relationships, we can simplify the conditions above and exploit them for auditing purposes. In practice, for a Metathesaurus relationship to be equivalent to or a subproperty of a Semantic Network relationship, it is necessary (but not sufficient) that the two relationships share at least one signature.

2. Background

2.1. Related work

There are a number of previous publications in the area of terminology/ontology auditing. Much of this research focuses on evaluating terminologies with respect to their hierarchical structure. Cimino [3,4] and Chen et al. [32] identify inconsistencies between the hierarchical relations in the UMLS Metathesaurus and the Semantic Network in order to audit Metathesaurus hierarchical relations. Bodenreider et al. [33], Ceuster et al. [2,34], Campbell et al. [35] and Wang et al. [8] audited the hierarchical relations in SNOMED CT. Auditing of cycles in of hierarchical relations in the UMLS is discussed in [36,37]. The focus of our study, however, is the auditing of associative (not hierarchical) relations in biomedical terminologies, which is intended to complement work on auditing hierarchical relations.

Less work has been done on terminology auditing from the perspective of associative relations. Campbell et al. [35] used lexical techniques between concepts with common substrings in SNOMED CT to identify potential missing associative (as well as hierarchical) relations. Wang et al. [8] and Min et al. [7] used a partition of a hierarchy of SNOMED and NCI Thesaurus, respectively, into areas of concepts with the same relationships to uncover missing and incorrect associative relations. Cohen et al. [38] audited the Gene hierarchy of NCI Thesaurus for missing associative relationships, using knowledge from the NCBI Entrez Gene database and the Biological Process hierarchy in the NCI Thesaurus. These research studies differ from our own insofar as we focus on identifying inconsistencies in mappings between the computed signatures of Metathesaurus relationships and Semantic Network relationships. Cimino [3], however, infers associative relations between semantic types of the UMLS Semantic Network from Metathesaurus relations between concepts participating in those semantic relationships.

More generally, this paper is a contribution to the study of relationships in terminologies [39,40] and extends previous work on the consistency of relations between the UMLS Metathesaurus and Semantic Network [41]. The methodology used for this audit was developed in part based on the fact that the source materials do not easily support a logic-based approach. That said, logic-based approaches to auditing terminologies/ontologies represent an important area of research. Schulz et al. [42] and Rogers et al. [43] used description logic techniques to audit the Read Codes. Cornet and Abu-Hanna [44] implemented DICE TS in Protégé Frames to audit the hierarchical relationships in DICE.

2.2. Mapping Metathesaurus relationships to the Semantic Network

In previous work [45], we explored a number of methods (both automated and manual) for establishing links (i.e., equivalent to or subproperty of) between Metathesaurus relationships and Semantic Network relationships. In the current paper, we take advantage of subsequent work done where the authors manually linked each (semantically significant) Metathesaurus relationship to a corresponding Semantic Network relationship. The total number of Metathesaurus (2008AA) relationships is 255, of which 177 were deemed semantically significant and were mapped to Semantic Network relationships. Those relationships that did not map to the Semantic Network exemplified three types of properties. Some indicated a lexical property, e.g., *noun_form_of, british_form_of*; others related in some way to the information model of the system from which they were derived, e.g., *patient demonstrates knowledge of nutrition outcome_of nausea*; and the remainder were relevant to vocabulary management; e.g., *sib_in_branch_of*, *classifies*. Table 1 shows the distribution of the full set of Metathesaurus relationships. Our auditing experiments were conducted using solely those Metathesaurus relationships that are semantically significant. Our mapping of these 177 relationships to Semantic Network relationships yielded the distribution according to the high-level relationship categories shown in Table 2.

In some cases, Metathesaurus relationships were lexically equivalent to existing Semantic Network relationships. For example, *ingredient_of*, *manifestation_of* and *tributary_of* exist in the Semantic Network, and they are Metathesaurus relationships, as well. Examining the use of Metathesaurus relationships reveals, however, that the same relationship name does not always indicate the same semantics. For example, the Metathesaurus relationship *contains* is actually used to mean—and was therefore mapped to—

[2] Other groupings of semantic types (e.g., [29]) could also support the definition of signatures.

Table 1
Distribution of Metathesaurus (2008AA) relationships. Auditing experiments were done using the 69% of Metathesaurus relationships that mapped to the UMLS Semantic Network.

Metathesaurus relationship type	Number of Metathesaurus relationships	Percentage (%)
Mapped to Semantic Network	177	69
Lexical property	12	5
Information model	26	10
Vocabulary management	40	16
Total	255	100

Table 2
Result of mapping Metathesaurus relationships to the Semantic Network high-level relationship categories.

High-level Semantic Network category mapped to	Number of Semantic Network relationships mapped to	Number of Metathesaurus relationships	Percentage (%)
conceptually related to	10	49	28
functionally related to	15	80	45
physically related to	7	14	8
spatially related to	2	24	14
temporally related to	2	10	5
Total	36	177	100

ingredient_of, rather than *contains* in the sense of the Semantic Network where it is defined as: "Holds or is the receptacle for fluids or other substances." The Semantic Network definition for *ingredient_of* is: "Is a component of, as in a constituent of a preparation", and this is the sense in which the Metathesaurus *contains* was used.

All Semantic Network relationships are explicitly defined in the Semantic Network distribution files. Our mapping of Metathesaurus relationships to the Semantic Network would have been considerably eased if the same had been true for the Metathesaurus terminologies.[3] In practice, our approach to mapping Metathesaurus relationships to Semantic Network relationships relies on the manual examination of a sample of Metathesaurus relations in which a given relationship participates, from which the domain and the range of the relationship are established. For example, the Metathesaurus relationship *gene_encodes_gene_product* is defined between some gene (e.g., *KLK15 Gene*) and some protein (e.g., *Kallikrein 11*). The Metathesaurus relationship is then manually associated with the corresponding high-level relationship category in the Semantic Network, based on domain and range information. In the example above, *gene_encodes_gene_product* is identified as a functional relation (*functionally_related_to*). Finally, whenever possible, we explore the relationship hierarchy in the Semantic Network to find a match for the Metathesaurus relationship. Among the subproperties of *functionally_related_to*, we identify *produces* as a close match, defined as "Brings forth, generates or creates. This includes yields, secretes, emits, biosynthesizes, generates, releases, discharges and creates." Because *gene_encodes_gene_product* is more specific than *produces*, we make it not equivalent to, but a subproperty of *produces*.

Fig. 2 shows the existing Semantic Network relationships. The 177 semantically significant Metathesaurus relationships mapped to a total of 36 of the 53 associative Semantic Network relationships. Indicated in parentheses after each relationship is the number of Metathesaurus relationships mapped to each Semantic Network relationship. As shown in Fig. 2, no Metathesaurus relationship corresponded to 17 Semantic Network relationships distributed among the five major categories of relationships. Examples of such Semantic Network relationships include *issue_in, interconnects, adjacent_to, complicates* and *carries_out*. Fig. 3 shows the overall distribution of the mappings. For each of ten Semantic Network relationships only one Metathesaurus relationship was mapped to it. For example, the Metathesaurus relationship *reformulation_of* mapped to the Semantic Network relationship *derivative_of*, and this was the only relationship that mapped to that particular Semantic Network relationship. By contrast, fully twenty-two Metathesaurus relationships mapped to the Semantic Network relationship *location_of*, including, for example, *disease_has_associated_anatomic_site*, *gene_found_in_organism* and *indirect_procedure_site_of*.

3. Methods

The method used for auditing Metathesaurus relations can be summarized as follows. All relations from both the Metathesaurus and the Semantic Network are transformed into signatures, an abstract representation of the kinds of entities involved with each relationship. More specifically, we use semantic groups to characterize entities in the domain and in the range of the relationships. Once the signatures have been established for all relationships, we compare the signature(s) of each Metathesaurus relationship to the signature(s) of the Semantic Network relationship mapped to. Fig. 4 illustrates the process. Shared signatures are indicative of consistent relationships, which is a necessary, but insufficient condition for the validity of the mapping between Metathesaurus and Semantic Network relationships. In contrast, discrepancies in the signatures can reveal inaccurate relations in the Metathesaurus, inaccurate mapping between Metathesaurus and Semantic Network relationships, wrong concept categorization, missing relations in the Semantic Network, or any combination thereof.

3.1. Creating signatures

As already noted, relations can be thought of as triples (e_d, r, e_r) in which e_d and e_r are entities and r is a relationship. In the Metathesaurus, concepts stand in relation to other concepts and relations are of the form (c_d, r, c_r), where c_d and c_r are concepts. In contrast, the entities related by Semantic Network relations are semantic types, with relations of the form (t_d, r, t_r). Metathesaurus concepts are categorized with semantic types from the Semantic Network and semantic types are partitioned into clusters called semantic groups. The signature of a relationship r is a pair of semantic groups (g_d, g_r), where g_d is the semantic group of the entity in the domain and g_r the semantic group of the entity in the range of the relationship. A given relationship may have more than one signature.

3.1.1. Creating signatures for Semantic Network relationships

The Semantic Network comprises 558 relations asserted between semantic types (SRSTR file), of which 135 are taxonomic relations (i.e., relations involving the relationship *isa*) and 423 are associative relations. 49 of the 53 Semantic Network associative relationships participate in these 423 relations. Relations asserted at a high-level are inherited along the subsumption hierarchy of the semantic types. For example, from the relation (Pharmacologic Substance, *treats*, Pathologic Function), additional relations involving the relationship *treats* are inferred among the descendants—direct or not—of Pharmacologic Substance and Pathologic Function. Such relations include (Antibiotic, *treats*, Disease or Syndrome), where Anti-

[3] National and International standards groups have recognized this problem, and they encourage explicit definitions of associative relationships. For example, the ANSI/NISO standard on controlled vocabularies states: *"The associative relationship is the most difficult one to define, yet it is important to make explicit the nature of the relationship between terms linked in this way and to avoid subjective judgments as much as possible; otherwise, RT [related term] references could be established inconsistently."* [13p.63].

associated_with	(none)
…..**conceptually_related_to**	(21)
.....property_of	(11)
.....conceptual_part_of	(5)
.....evaluation_of	(2)
.....measures	(1)
.....diagnoses	(3)
.....issue_in	(none)
.....derivative_of	(1)
.....developmental_form_of	(none)
.....degree_of	(1)
.....measurement_of	(none)
.....method_of	(2)
.....analyzes	(2)
.....assesses_effect_of	(none)
…..**physically_related_to**	(2)
.....part_of	(3)
.....contains	(1)
.....consists_of	(2)
.....connected_to	(none)
.....interconnects	(none)
.....branch_of	(1)
.....tributary_of	(1)
.....ingredient_of	(4)
…..**temporally_related_to**	(none)
.....co-occurs_with	(9)
.....precedes	(1)
…..**spatially_related_to**	(2)
.....location_of	(22)
.....adjacent_to	(none)
.....surrounds	(none)
.....traverses	(none)
…..**functionally_related_to**	(5)
.....manifestation_of	(4)
.....affects	(19)
.....manages	(none)
.....treats	(3)
.....disrupts	(7)
.....complicates	(none)
.....interacts_with	(none)
.....prevents	(1)
.....occurs_in	(13)
.....process_of	(1)
.....uses	(8)
.....indicates	(3)
.....result_of	(7)
.....brings_about	(2)
.....produces	(3)
.....causes	(3)
.....performs	(none)
.....carries_out	(none)
.....exhibits	(1)
.....practices	(none)

Fig. 2. Semantic Network relationships with number of Metathesaurus relationships (in parentheses) mapped to each relationship. A total of 177 Metathesaurus relationships mapped to 36 Semantic Network relationships.

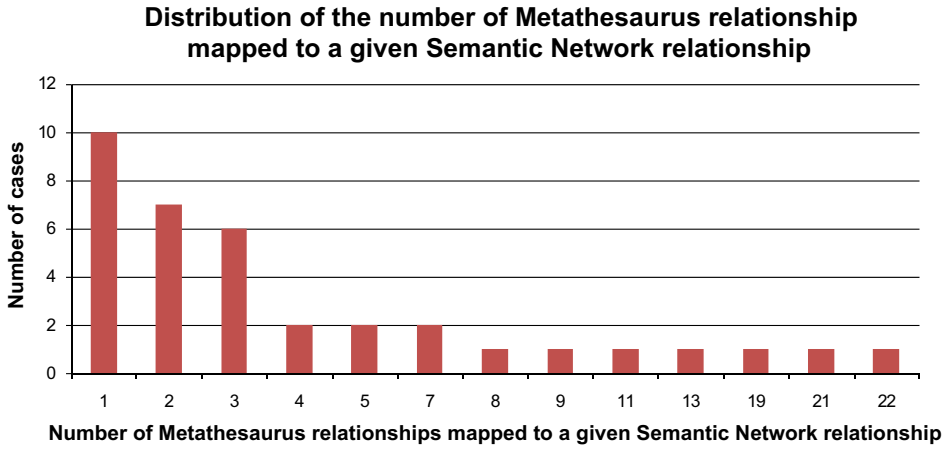

Fig. 3. Distribution of Metathesaurus relationships mapped to Semantic Network. The majority of Semantic Network relationships had only one, two or three Metathesaurus relationships mapped to them. One Semantic Network relationship (*location_of*) had 22 Metathesaurus relationships mapped to it.

biotic and Disease or Syndrome are descendants of Pharmacologic Substance and Pathologic Function, respectively. The fully inherited list of relations in the Semantic Network is provided as part of the UMLS distribution (SRSTRE2 files). There is a total of 6752 (asserted and inherited) relations between semantic types, of which 500 are taxonomic relations. Each one of the 135 semantic types is associated with one (and only one) of the 15 semantic groups. For example, Antibiotic belongs to the semantic group **Chemicals and Drugs**.[4]

In order to create the signature of a given Semantic Network relationship, we start by collecting all the relations in which this relationship participates. Each relation (t_d, r, t_r) is transformed into a signature r (g_d, g_r) by identifying the semantic groups g_d and g_r corresponding to the semantic types t_d and t_r, respectively. For example, the signature of the relationship *treats* created from the relation (Pharmacologic Substance, *treats*, Disease or Syndrome) is (**Chemicals and Drugs**, **Disorders**) because Pharmacologic Substance and Disease or Syndrome belong to the semantic groups **Chemicals and Drugs** and **Disorders**, respectively. Fig. 5 shows all the signatures for the Semantic Network relationship *treats*.

No semantic types are associated with 5 Semantic Network relationships (*functionally_related_to*, *physically_related_to*, *spatially_related_to*, *temporally_related_to* and *brings_about*). In order to compute the signature of these relationships, we assumed that their domain and

[4] *Chemicals and Drugs* is the official name of the semantic group representing the union – not intersection – of semantic types for chemicals and for drugs.

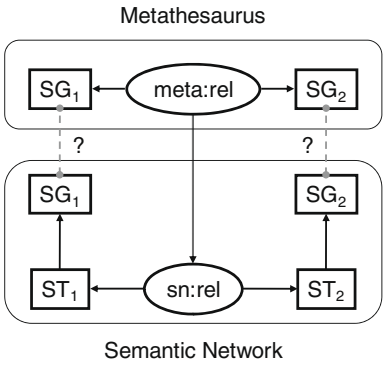

Fig. 4. Comparing signatures across two knowledge sources. Metathesaurus and Semantic Network relationship signatures are compared.

range would be the union of the domains and ranges of the relationships they subsume. For example, *brings_about* subsumes *produces* and *causes*. The relations involving *produces* include (Fully Formed Anatomical Structure, *produces*, Body Substance) and *causes* participates in the relation (Bacterium, *causes*, Pathologic Function). Therefore, although not explicitly represented in the Semantic Network, we assumed the existence of relations such as (Fully Formed Anatomical Structure, *brings_about*, Body Substance) and (Bacterium, *brings_about*, Pathologic Function) to create the following signatures for *brings_about*: (**Anatomy**, **Anatomy**) and (**Living Beings**, **Disorders**), respectively.

3.1.2. Creating signatures for Metathesaurus relationships

The method for creating signatures for Metathesaurus relationships is similar to that described for Semantic Network relationships. A minor difference is that Metathesaurus concepts are linked to semantic groups not directly, but through the semantic types. As a consequence, concepts first need to be linked to their semantic type(s), and each semantic type to its semantic group. While many concepts have more than one semantic type, only 1208 of the 1.5 M Metathesaurus concepts have more than one semantic group. In most cases, a given relation is transformed into one signature, but relations involving concepts with multiple semantic groups result in several signatures. As it is the case with Semantic Network relationships, in most cases, Metathesaurus relationships also have more than one signature. For each signature of a given relationship, we tally how many individual relations contributed to this signature, in order to determine, for example, whether one particular signature is most frequent for this relationship.

For each Metathesaurus and Semantic Network relation (e_1, r_d, e_2), there is also a reciprocal relation (e_2, r_i, e_1), where r_i is the inverse of r_d. For example, the Metathesaurus relation (*Lung*, *location_of*, *Radiation pneumonitis*) is mirrored by a relation (*Radiation pneumonitis*, *has_location*, *Lung*). In order to avoid double counting, we eliminated the inverse relations from the dataset. In practice, we selected the direct relation (e_1, r_d, e_2) as the one for which the Metathesaurus relationship r_d was mapped to a direct relationship from the Semantic Network. From the two relations above, we selected the former, because the Metathesaurus relationship *location_of* was mapped to the direct Semantic Network relationship *location_of*. Moreover, several copies of the same direct relation may be represented in the Metathesaurus when this relation is carried by multiple translations of a given source vocabulary, since translated terms are integrated as synonyms in the Metathesaurus and share the same concept unique identifier. We therefore eliminated from our UMLS dataset the various translations of MeSH, MedDRA and SNOMED CT, keeping the English version as the reference.

As shown in Fig. 6, of the 104,675 Metathesaurus relations involving the relationship *may_treat*, a majority holds between a

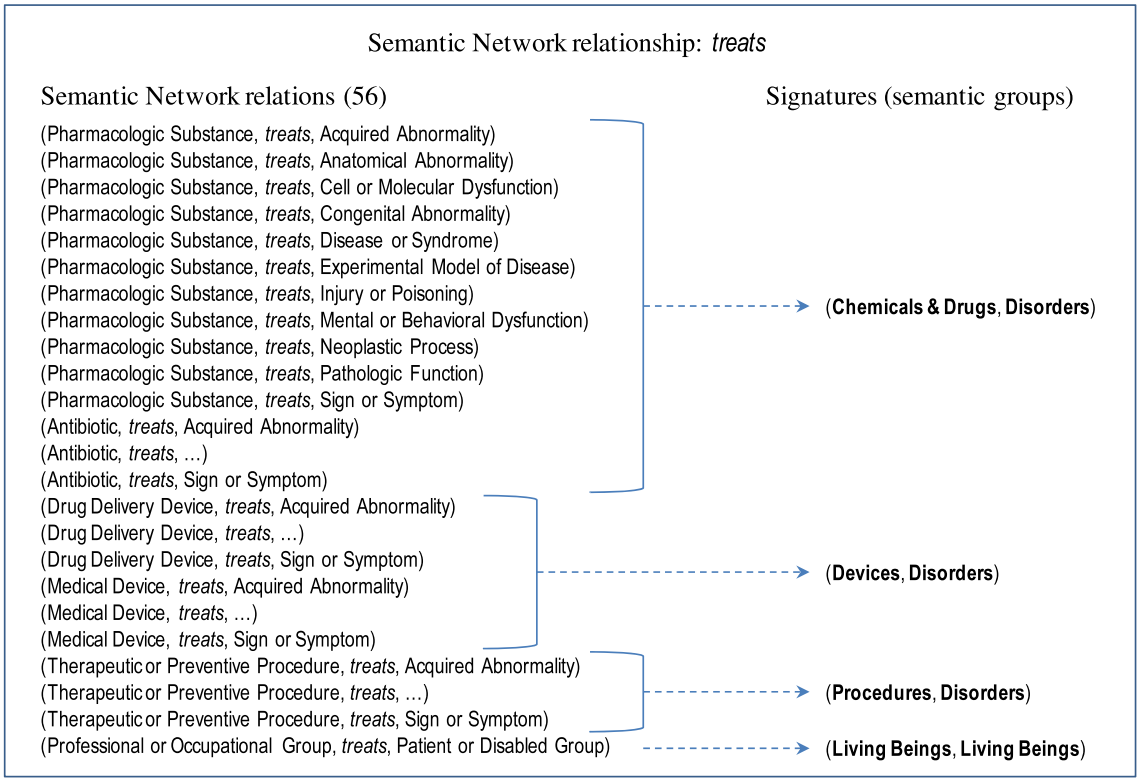

Fig. 5. All signatures for the Semantic Network relationship *treats*. Signatures involve chemical entities, devices and procedures with disorders, and also living beings with other living beings.

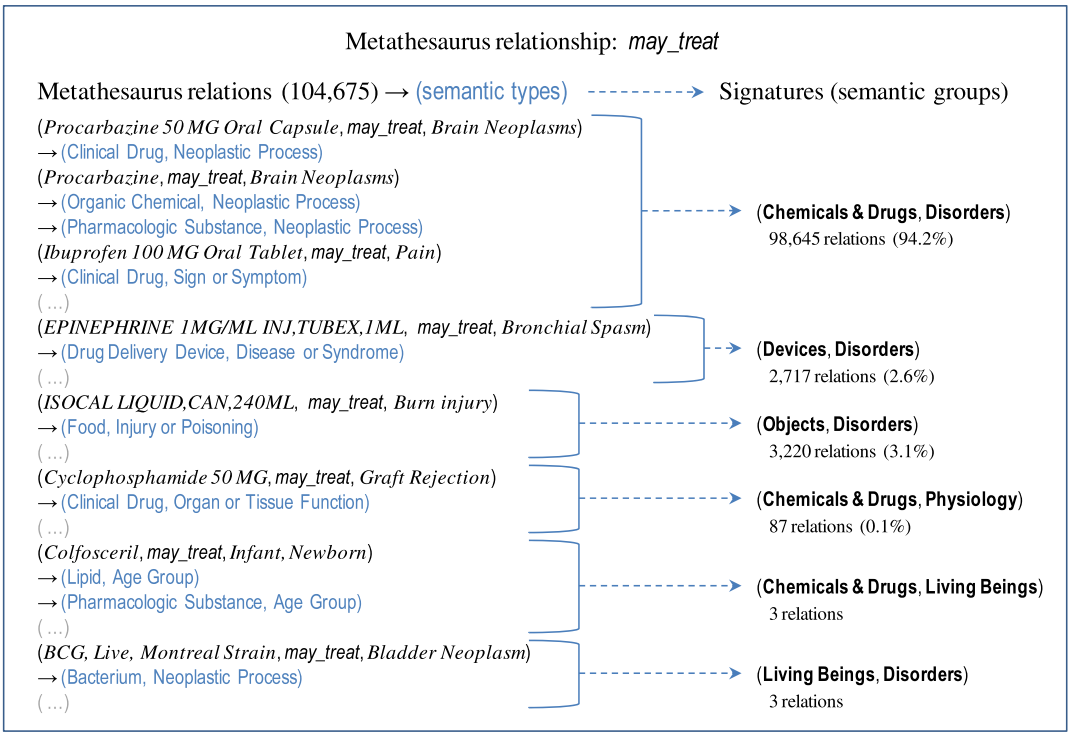

Fig. 6. Signatures for the Metathesaurus relationship *maytreat*. A majority holds between a chemical entity and a disorder.

chemical entity and a disorder, e.g., (*Procarbazine 50 MG Oral Capsule*, *may_treat*, *Brain Neoplasms*). The semantic types of the two concepts are Clinical Drug and Neoplastic Process, respectively. Based on this relation, the signature of the Metathesaurus relationship *may_treat* is (**Chemical and Drugs**, **Disorders**). Other signatures for the Metathesaurus relationship *may_treat* include (**Devices**, **Disorders**), e.g., from (*EPINEPHRINE 1MG/ML INJ,TUBEX,1ML*, *may_treat*, *Bronchial Spasm*), (**Objects**, **Disorders**), e.g., from (*ISOCAL LIQUID,CAN,240ML*, *may_treat*, *Burn injury*), (**Chemical and Drugs**, **Physiology**), e.g., from (*Cyclophosphamide 50 MG*, *may_treat*, *Graft Rejection*), (**Chemical and Drugs**, **Living Beings**), e.g., from (*Colfosceril*, *may_treat*, *Infant, Newborn*), and (**Living Beings**, **Disorders**), e.g., from (*BCG, Live, Montreal Strain*, *may_treat*, *Bladder Neoplasm*).

3.2. Comparing signatures

The mapping created between Metathesaurus and Semantic Network relationships resulted in associations between relationships across the two knowledge sources. The signatures of a given Metathesaurus relationship r_m are compared to the signatures of the Semantic Network relationship r_s to which this Metathesaurus relationship was mapped. For example, the Metathesaurus relationship *may_treat* was mapped to the Semantic Network relationship *treats*, allowing the 6 signatures of *may_treat* (Fig. 6) to be compared to the 4 signatures of *treats* (Fig. 5).

3.2.1. Consistent relationships

When a Metathesaurus relationship r_m shares at least one signature with the Semantic Network relationship r_s to which it is mapped, we consider that the semantics of the Metathesaurus relationship r_m is consistent with that of the Semantic Network relationship r_s. This condition is necessary, but not sufficient, for the mapping to be valid. From a quantitative perspective, we count not only how many signatures are shared between r_m and r_s, but also how many relations contributed to these shared signatures, relative to the total number of relations for this Metathesaurus relationship. We consider r_m highly consistent with r_s if at least 75% of the Metathesaurus relations involving r_m have a shared signature with r_s. For example, *may_treat* is mapped to *treats*, and, as shown in Fig. 7, these two relationships have two signatures in common: (**Chemical and Drugs**, **Disorders**) and (**Devices**, **Disorders**). Together, these two signatures represent 96.6% of all Metathesaurus relations involving *may_treat*. Therefore, *may_treat* is deemed consistent with *treats* (despite the fact that four of the six signatures observed for *may_treat* are not signatures of *treats*). The consistency between the two relationships helps confirm the validity of the mapping of *may_treat* to *treats*.

3.2.2. Inconsistent relationships

A Metathesaurus relationship r_m is inconsistent with the Semantic Network relationship r_s to which it is mapped when less than 75% of the Metathesaurus relations involving r_m have a shared signature with r_s. In such cases, we first consider the total number of relations in which relationship r_m participates, in order to prioritize the auditing effort. For example, the Metathesaurus relationship *has_time_modifier* shares no signatures with the Semantic Network relationship *has_property* to which it was mapped. While generally worrisome, this inconsistency will not immediately be the focus of our auditing effort, because *has_time_modifier* actually participates in only four of the 1.8 M associative relations in the Metathesaurus.

3.2.3. Dominant signature

Another characteristic used to direct our auditing effort is the existence of one dominant signature for a given Metathesaurus relationship. A signature is dominant for a given relationship if at least 75% of all relations in which this relationship participates have this signature. For example, among the six signatures for the Metathesaurus *may_treat* shown in Fig. 6, the dominant signature is (**Chemical and Drugs**, **Disorders**), corresponding to 94.2% of the relations involving *may_treat*. Metathesaurus relationships with one dominant signature are generally semantically homogeneous. In contrast, the existence of several large groups of relations with distinct signatures for a given Metathesaurus relationship may rather

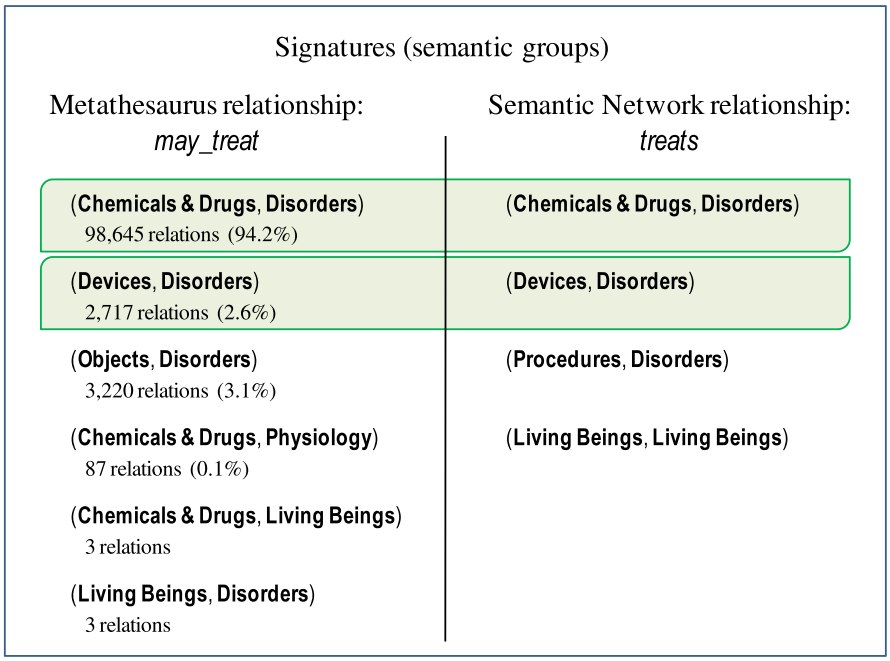

Fig. 7. Comparing the signatures of *maytreat* and *treats*. Two shared signatures correspond to 96.6% of the Metathesaurus relations involving *maytreat*: the two relationships are consistent.

be indicative of heterogeneous semantics for this relationship, especially if the various groups of relations correspond to different source vocabularies. We hypothesize that when one dominant signature captures a large proportion of the relations for a given Metathesaurus relationship but does not match the signature(s) of the Semantic Network relationship to which it was mapped, the mapping is inaccurate and needs to be revisited. For example, a majority of the relations for the Metathesaurus relationship *biological_process_has_initiator_chemical_or_drug* have the signature (**Physiology, Chemical and Drugs**), which does not match the signatures of the Semantic Network relationship *brought_about_by* to which it was mapped.

Finally, the mapping of Metathesaurus relationships to top-level relationships in the Semantic Network is considered with special attention. As noted before, the semantics of most top-level Semantic Network relationships is not asserted, but reconstructed from that of the descendants of the particular top-level relationship. Therefore, because mapping a given Metathesaurus relationship to a top-level Semantic Network relationship implies that there was no specific descendant of this Semantic Network relationship we could have mapped to, it is likely that the semantics of the Metathesaurus relationship is not covered by that of the top-level Semantic Network relationship. In this case, the Semantic Network should be, not only linked to, but potentially enriched with the corresponding relationship.

4. Results

In the following, we report the results of transforming associative Metathesaurus and Semantic Network relations into their signatures, and we report the consistency of the mappings according to several criteria.

4.1. Distribution of signatures

The transformation of the 177 semantically significant Metathesaurus relationships and the 53 associative Semantic Network relationships into their signatures resulted in the distribution shown in Fig. 8. The majority of the 177 Metathesaurus relationships have up to four signatures, while the Semantic Network relationships have on average five signatures. Additionally, however, as many as 13 Metathesaurus relationships have 20 or more signatures. Examples of these latter are *component_of, measures, interprets* and *related_to*. One Metathesaurus relationship, *associated_with*, has 91 signatures. One top-level Semantic Network relationship, *functionally_related_to*, has 140 signatures. This can be explained by the fact that this relationship subsumes many other relationships, and its signatures result from computing the union of the signatures of the relationships that it subsumes.

4.2. Mapping consistency

We evaluated the mapping of Metathesaurus relationships to Semantic Network relationships by comparing their semantic signatures. We investigated overall consistency in a variety of ways, including overall degree of consistency, degree of consistency according to high-level Semantic Network categories mapped to, according to the dominant signature of a Metathesaurus relationship, and, finally, according to the number of sources that contributed a particular Metathesaurus relationship.

4.2.1. Overall consistency

The consistency of the mapping is shown in Table 3. The table shows that 48% of the mappings are highly consistent (with at least 75% of their relations being consistent). 11% show some consistency, and in 41% of the cases there is no overlap at all in the signatures of the Metathesaurus relationship and the Semantic Network relationship to which it was mapped. In addition to assessing the consistency of Metathesaurus relationships, we also evaluated the consistency of the Metathesaurus relations in which these relationships participate. Overall, 63% of the 1.8 M associative relations in the Metathesaurus are consistent with relations in the Semantic Network.

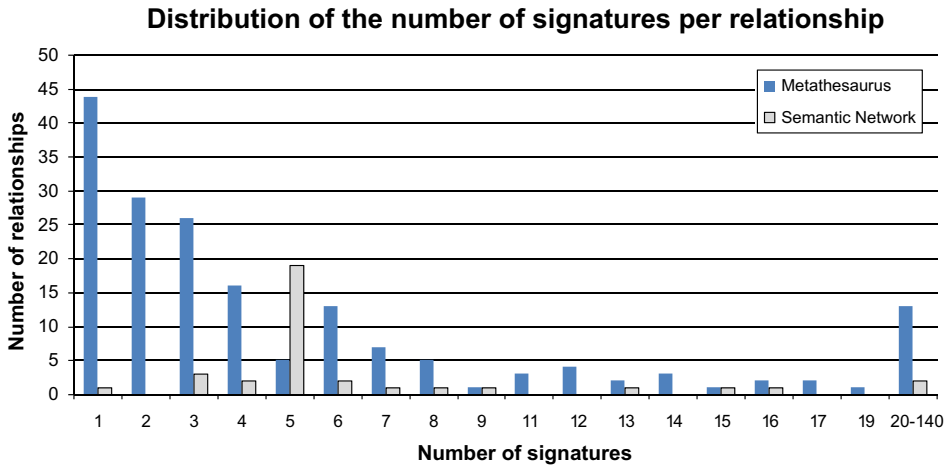

Fig. 8. Distribution of the number of signatures for Metathesaurus relationships compared with Semantic Network relationships. The majority of the 177 Metathesaurus relationships have no more than four signatures, while the Semantic Network relationships have on average five signatures.

Table 3
Overall consistency of mapping Metathesaurus relationships to the Semantic Network.

Consistency with Semantic Network	Number of Metathesaurus Relationships	Percentage (%)
High-consistency	84	48
Some consistency	20	11
No consistency	73	41
Total	177	100

4.2.2. Consistency of mappings to top-level Semantic Network relationships

In some cases, a Metathesaurus relationship was mapped directly to a top-level Semantic Network relationship because no suitable more specific relationship was available. 30 relationships were directly mapped to these high-level relationships as follows: 21 were mapped to *conceptually_related_to*, 5 to *functionally_related_to*; 2 to *physically_related_to*; and 2 to *spatially_related_to*. None was mapped directly to *temporally_related_to*. 20 (66%) of these relationships are not consistent with the Semantic Network relationship mapped to.

4.2.3. Consistency and dominant signature

One hundred and forty-seven (83%) of the 177 Metathesaurus relationships have a dominant signature. For the remaining 30 relationships no single signature was significantly more frequent than any of the other signatures. Of those that have a dominant signature, 76 (52%) are consistent with the Semantic Network relationship mapped to and 71 (48%) are not.

4.2.4. Consistency according to number of sources

Table 4 shows the consistency according to the number of sources in which a particular Metathesaurus relationship occurs.

Table 4
Consistency according to number of Metathesaurus sources. The majority of Metathesaurus relationships occur in only one source.

Number of sources	Number of relationships	High-consistency (%)	Some consistency (%)	No consistency (%)
1	161	48	9	43
>1	16	38	44	18

One hundred sixty-one (91%) of the 177 relationships occur in only one Metathesaurus source, while only 16 occur in multiple sources. The degree of consistency with Semantic Network relationships varies as shown in the table.

5. Discussion

5.1. Interpretation of findings

5.1.1. Variability in number of signatures

We hypothesized that if a relationship had a large number of signatures, this would likely be indicative of inconsistent mappings. The assumption is that the larger the number of signatures, the more likely it is that the relationship has imprecise or heterogeneous semantics. This hypothesis does not appear to have been borne out. Some of the very high-frequency relationships have a large number of signatures and yet the mapping was either highly or moderately consistent. For example, *clinically_associated_with* has 88 signatures and yet it has a mapping consistency rate of 75%. A low number of signatures alone is also not necessarily a good predictor of a consistent mapping. For example, both *allelic_variant_of* and *chemotherapy_regimen_has_component* have a small number of signatures and, yet, they have no consistency with the relationships mapped to.

5.1.2. Consistency of mappings

There are thirty-three Metathesaurus relationships that have greater than 10,000 relations represented in the Metathesaurus. These thirty-three relationships account for almost 88% of the total 1.8 million relations. Fig. 9 shows, at a glance, the consistency of the mappings for these high-frequency relationships. The left-hand side of the graph (in red) shows the number of inconsistent relations for the relationship; the right hand side (in green) shows the number of consistent relations, and on the far right the name of the Metathesaurus relationship is listed together with the percentage of consistent relations and the number of signatures for each of the relationships.

Overall, twenty-two (67%) of the Metathesaurus relationships that have greater than 10,000 relations are highly consistent ($\geqslant$75% consistency as indicated on the right hand side of Fig. 9) with the Semantic Network relationships to which they were mapped. Note that the highest frequency Metathesaurus

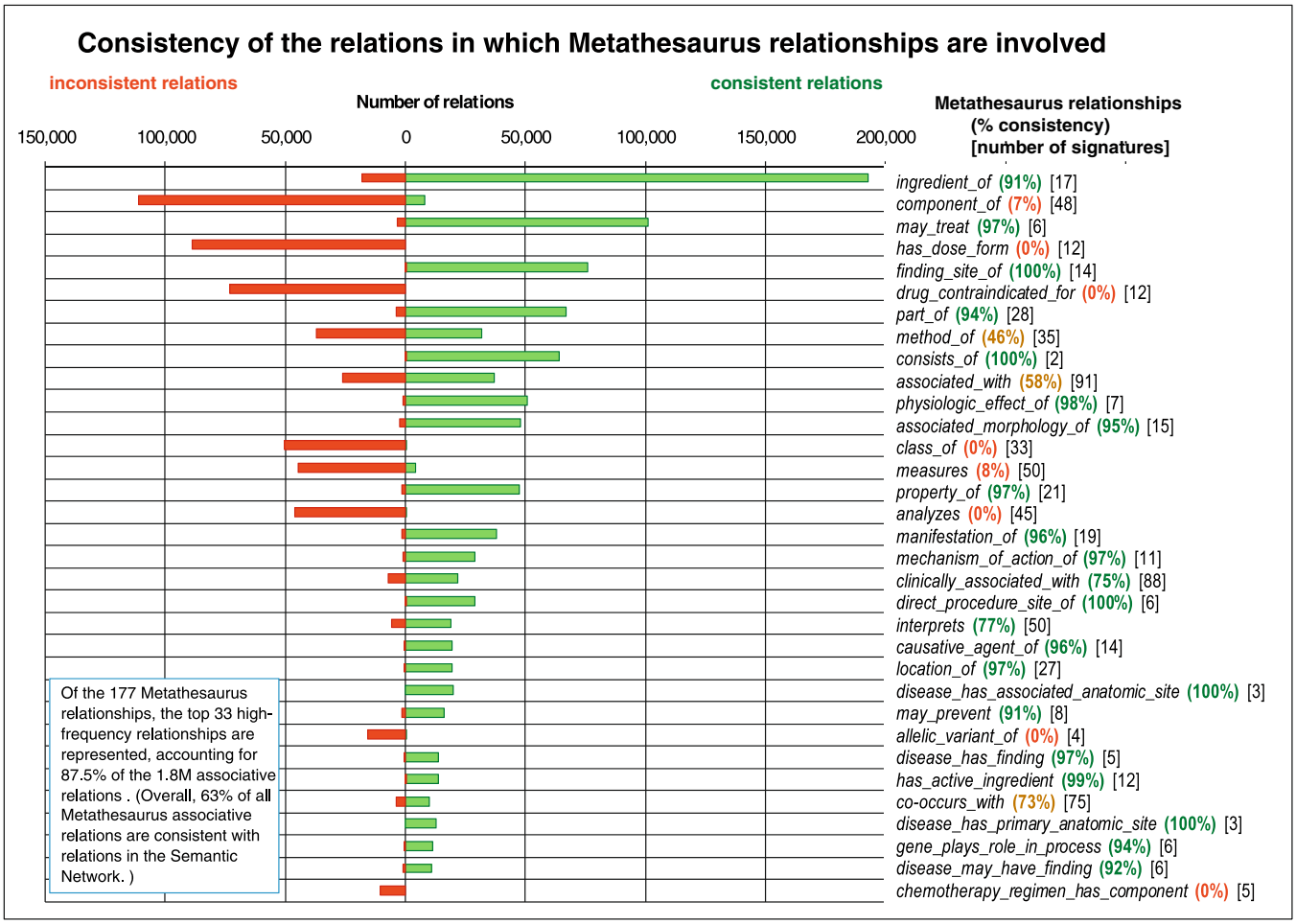

Fig. 9. Consistency of high-frequency Metathesaurus relationships (involved in at least 10,000 relations).

relationship is *ingredient_of*. It is represented by 210,740 relations and has a total of 17 signatures. For 91% of its relations, there is consistency with the Semantic Network relationship to which it has been mapped.

Of the eleven that are not highly consistent (<75% consistency as indicated on the right hand side of Fig. 9), six (18%) have no overlap with the signatures of the Semantic Network, two (6%) have a very small overlap, and three (9%) are moderately consistent. Because of their high-frequency, these eleven relationships are strong candidates for further investigation and potential modification.

5.1.3. Dominant signatures

For each of the 90 relationships that have a frequency of 1000 relations or more, we investigated whether its dominant signature matched the signatures of the Semantic Network relationship to which it was mapped. Fifty-seven (63%) of the ninety relationships have dominant signatures that match the signatures of the Semantic Network relationships to which they were mapped. Thirty-three (37%) are inconsistent.

The majority, 17 of the 33 inconsistent mappings, are mappings to semantic network relationships in the 'conceptually related to' high-level category, and another 11 are mappings to the 'function-

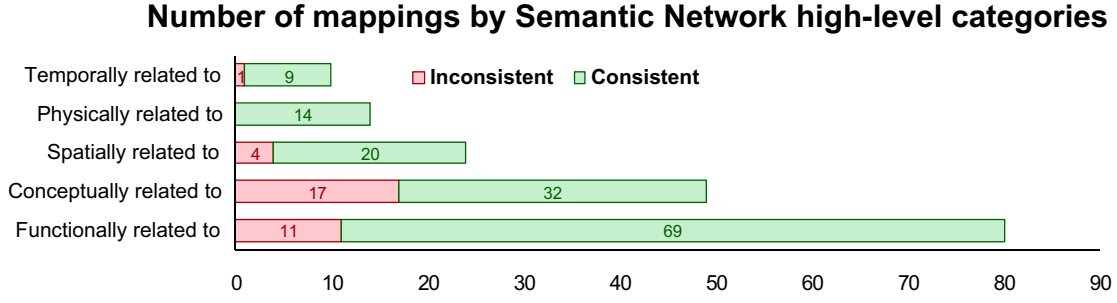

Fig. 10. Consistency of mappings of high-frequency relationships by high-level Semantic Network categories. Mappings to the 'conceptually related to' category are disproportionately worse than mappings to other high-level categories.

ally related to' category. The remainder were either mapped to a relationship in the 'spatially related to' or 'temporally related to' categories. This is in contrast with the overall mappings for all 177 relationships. (See Table 2).

Fig. 10 indicates that mappings to the 'conceptually related to' category are disproportionately worse than mappings to the other high-level categories. Eight of the seventeen inconsistent mappings in the 'conceptually related to category' are in the very high-frequency category and have already been discussed above. Another group of relationships in this category occur in the NCI thesaurus and they are of the general form such that a disease excludes a particular anatomic entity, either as its origin or as its anatomic site. An example is *disease_excludes_abnormal_cell*. The dominant signature of this relationship is (**Disorders**, **Anatomy**) which does not match the signatures of the conceptual relationship mapped to. It is not clear what the correct answer is in this case. While, on the one hand, an "exclusion" can be seen as a conceptual notion, it is not obvious that the relations in which the Semantic Network relationship participates should be modified to accommodate this relationship. For these cases, a clarification from the developers of the source vocabulary would be welcome.

5.1.4. Distribution across sources

Sixteen (9%) of the total 177 Metathesaurus relationships occur in more than one source. Table 5 shows the number of sources from which these relationships derive and the percentage of relations that are consistent with the Semantic Network. Six of the relationships that occur in more than one source are highly consistent; seven have some level of consistency, and three are not consistent at all. The semantics of relationships such as *ingredient_of*, *manifestation_of*, *part_of* and *location_of* seems consistent across vocabularies. In contrast, the semantics of relationships such as *component_of* and *contains* is not. While we hypothesized that a large number of sources would potentially lead to inconsistent mappings, it would appear that the number of sources alone does not predict whether a mapping will be successful or not.

5.2. Implications

As mentioned earlier, the consistency between Metathesaurus and Semantic Network associative relations assessed through the auditing process is only a necessary condition to the validity these relations. Therefore, the auditing process aims not at establishing semantic consistency, but rather at identifying inconsistencies, indicative of some semantic mismatch between the two knowledge sources. The auditing process has exposed a variety of errors, including some errors in the mapping process, as well as quality issues in the Metathesaurus. In addition, the investigation of inconsistencies indicated, on the one hand, some potential modifications to the Semantic Network and, on the other, to some necessary clarifications by the developers of a Metathesaurus source vocabulary.

Table 5
Relationships that occur in more than one source in the Metathesaurus. The number of sources alone does not predict whether a mapping is successful or not.

Relationship	Number of sources	Number of relations	Consistency (%)
ingredient_of	6	210,740	91
measures	6	48,854	8
evaluation_of	5	1771	66
analyzes	5	46,203	0
part_of	5	71,038	94
associated_with	5	63,500	58
has_dose_form	4	88,803	0
location_of	3	20,174	97
manifestation_of	3	39,599	96
component_of	3	119,177	7
method_of	3	68,994	46
result_of	2	4	75
conceptual_part_of	2	335	1
form_of	2	1245	0
contains	2	2203	8
property_of	2	48,795	97

5.2.1. Mapping issues

A majority of high-frequency Metathesaurus relationships are consistent with the Semantic Network relationships to which they were mapped, which helps confirm the validity of the mapping. When this is not the case, however, it is possible that we made an error in the original mapping. The mapping needs to be reevaluated in light of the auditing results of the associative relations.

For example, the Metathesaurus relationship *chemotherapy_regimen_has_component* has five signatures. Its dominant signature (**Procedures, Chemicals and Drugs**) does not match the signatures of the Semantic Network relationship, *conceptual_part_of*, to which it was mapped. An example relation is *busulfan/cyclophosphamide/etoposide chemotherapy_regimen_has_component Busulfan*. A better mapping might have been to the Semantic Network relationship *uses*, which does have the expected signature.

5.2.2. Quality issues in the Metathesaurus

The auditing approach proposed in this paper is also sensitive to inaccurate associative relations in the Metathesaurus and inaccurate concept categorization. In this case, it will falsely identify inconsistencies between Metathesaurus and Semantic Network relationships. Identifying such errors is indeed one of the expected benefits of auditing associative relations. Some quality issues in the Metathesaurus are illustrated in this section.

5.2.2.1. Inaccurate Metathesaurus relations. The Metathesaurus relationship *measures* was mapped to the Semantic Network relationship of the same name. It has 50 signatures with the majority of its relations derived from LOINC or a LOINC collaborative vocabulary. It does not have a dominant signature, and its most frequent signature (**Chemicals and Drugs**, **Physiology**) does not match the signatures of the Semantic Network relationship, *measures*. Some examples from LOINC are:

- *Chlorine measures Chlorine:Mass Concentration:Point in time:Water:Quantitative*
- *Surgical approach measures Surgical approach:Type:Point in time:Surgical procedure:Nominal*
- *Viscosity measures Viscosity:Viscosity:Point in time:Whole blood:Quantitative*

One issue here is that this relationship is used in LOINC to represent numerous distinct senses. Another, more acute problem is that the LOINC relationship is recorded "backwards" in the Metathesaurus.[5] For example, viscosity does not measure, but is rather measured by a viscosity measurement laboratory test. After correcting its direction, the Metathesaurus relation *Viscosity:Viscosity:Point in time:Whole blood:Quantitative measures Viscosity* becomes consistent with the Semantic Network relationship *measures* through the shared signature (**Procedures, Phenomena**).

5.2.2.2. Concept categorization. The Metathesaurus relationship *method_of* was mapped to the Semantic Network relationship of the same name. In almost half of the cases (46%), the mapping was consistent. Its most frequent signature (**Procedures, Proce-**

[5] Version 2008AA of the UMLS Metathesaurus asserts *Viscosity measures Viscosity:Viscosity:Point in time:Whole blood:Quantitative*. Previous versions of the Metathesaurus asserted this relation in the opposite direction (*Viscosity:Viscosity:Point in time:Whole blood:Quantitative measures Viscosity*).

dures) matches the signature of the Semantic Network relationship. In contrast, the signature (**Procedures, Physiology**) derived from LOINC relations does not. Two examples illustrate:

- *Serum Bactericidal Test method_of Almecillin:Susceptibility:Point in time:Isolate + Serum:Ordinal:SERUM BACTERICIDAL TITER*
- *Agar diffusion method_of Cefamandole:Susceptibility:Point in time:Isolate:Quantitative or Ordinal:Agar diffusion*

The definition of *method_of* in the Semantic Network ("The manner and sequence of events in performing an act or procedure") is consistent with another use in LOINC between a particular method (e.g., *Serum Bactericidal Test*) and the laboratory procedures in which this method is used. However, some laboratory entities in LOINC are typed as Clinical Attribute, rather than Laboratory Procedure. Since the semantic type Clinical Attribute is part of the semantic group **Physiology**, and not **Procedures**, the signature obtained from these relations does not match any signatures for *method_of* in the Semantic Network.

5.2.3. Potential additions to the Semantic Network

Unlike the Metathesaurus, the Semantic Network has not grown significantly during the past decade. On the one hand, the Semantic Network represents high-level, definitional knowledge and its size is purposely kept to a minimum. Therefore, fewer changes are expected over time. On the other hand, semantic types expected to support the categorization of Metathesaurus concepts and Semantic Network relationships should reflect salient information in the Metathesaurus, which prompted the addition of the semantic type Drug Delivery Device and relationships *tributary_of*, for example. Various changes have been suggested (e.g., for genomics [46]) and discussed at a workshop in 2005 [47]. However, the absence of clear use cases and the potential need for re-categorizing thousands of Metathesaurus concepts have precluded the implementation of such changes.

More fundamentally, the underlying question is whether the Semantic Network is a top-level ontology for the biomedical domain [48] and should provide a prescriptive organizational structure for Metathesaurus concepts and relations, or, as it is the case now, is should be used only as a loose reference. The former use would require a mapping between Metathesaurus and Semantic Network relationships and the addition of new Semantic Network relations to accommodate equivalent relations in the Metathesaurus. The role played by the Semantic Network in the Metathesaurus editing environment would also need to be modified if semantic consistency between the two structures were to be enforced. In fact, such a prescriptive role of the Semantic Network might be fundamentally incompatible with the original goal of the Metathesaurus to accommodate all relations from its source vocabularies. However, we believe that enriching the Semantic Network with new relations and taking greater advantage of the Semantic Network in the Metathesaurus editing environment would significantly benefit semantic consistency in the UMLS.

The auditing process revealed several cases where either the addition of a new Semantic Network relationship or additional relations for existing relationships might be considered. The Metathesaurus relationship *has_dose_form* was mapped to *conceptually_related_to*. It occurs in four sources and has twelve signatures. Its most frequent signature (**Chemicals and Drugs**, **Chemicals and Drugs**) does not match the signatures of the Semantic Network relationship to which it was mapped. An example is: *Mebendazole 100 MG Chewable Tablet*, *has_dose_form*, *Chewable Tablet*. Because the Semantic Network does not have a relationship of the appropriate specificity, we mapped to a top-level relationship, which itself does not have the relevant signature. Similarly, *drug_contraindicated_for* was mapped to *conceptually_related_to*. It occurs in one source and has twelve signatures. Its dominant signature (**Chemicals and Drugs**, **Disorder**) also does not match the signatures of the relationship to which it was mapped. An example is: *Fluphenazine drug_contraindicated_for Brain Damage, Chronic*. Both of these Metathesaurus relationships might well be candidates for addition to the Semantic Network. More generally, the eleven Metathesaurus relationships that have greater than 10,000 relations and are not highly consistent (shown in Fig. 9) should be examined for potential addition to the Semantic Network.

The 75 signatures of the Metathesaurus relationship *co-occurs_with* are consistent for 73% of their relations, and this is close to the threshold for being highly consistent. Nonetheless, it was worth considering why 27% of its relations are inconsistent with the Semantic Network relationship of the same name to which it was mapped. Its most frequent signature (**Disorders**, **Disorders**) does match the signatures of the Semantic Network relationship, so the mapping appears to have been reasonable. Almost 15% of the relations involve procedures, and these are not consistent with the current Semantic Network relationship. An example is, *Total excision of stomach NOS co-occurs_with Esophagojejunostomy*. These cases would argue for the addition of new relations to the existing Semantic Network relationship, *co-occurs_with*. Currently this relationship only allows two signatures (**Disorders**, **Disorders**) and (**Physiology, Physiology**).

5.2.4. Needed clarifications in a source vocabulary

The auditing process revealed unclear semantics of the Metathesaurus relationships. For example, the relationship *component_of* has the second most frequent number (119,177) of relations overall, and it has very low (7%) consistency with the Semantic Network relationship to which it was mapped. Notice that it has a very high-number (49) of signatures. This relationship was mapped to *conceptual_part_of* in the Semantic Network. It appears in three sources, with the majority (83%) of the relations and signatures (90%) derived from LOINC. There is no dominant signature, which would seem to indicate either that this relationship has a very broad semantics or that this single relationship represents numerous distinct senses. Some examples of its use in three vocabularies are shown below.

- *Blood component_of Blood in gastric contents measurement (SNOMED CT)*
- *Pharmaceutical Preparations component_of Urine drug screening (LOINC)*
- *Methotrexate component_of COMVP protocol (National Cancer Institute's Physician Data Query)*

The high-frequency relationship *class_of* occurs only in LOINC. Its most frequent signature (**Procedures, Physiology**) does not match the signatures of the relationship to which it was mapped, *conceptually_related_to*. Again, this is a case of potentially broad semantics inhering in a single relationship.

Some examples from LOINC are:

- *Ambulance claims attachment class_of Rationale for scheduled transport:Type:Point in time:EMS transport:Nominal*
- *Radiology studies class_of MRI of larynx*
- *Antimicrobial susceptibility class_of Acyclovir:Susceptibility:Point in time:Isolate:Quantitative or Ordinal*

The Metathesaurus relationship *analyzes* was mapped to the Semantic Network relationship of the same name. It occurs only in LOINC or a LOINC collaborative vocabulary. It has 45 signatures, the most frequent being (**Physiology, Anatomy**), but no dominant signatures.

Some examples from LOINC are:

- *Coding system.current:Type:Point in time:Race:Nominal analyzes Racial group*
- *Body surface area formula:Type:Point in time:Formula:Nominal:∗ analyzes Mathematical formula*
- *Bacteria identified^^^6:Presence or Identity:Point in time:Burn:Nominal:Culture analyzes Burn injury*

The Semantic Network relationship *analyzes* has only two signatures (**Procedures, Anatomy**) and (**Procedures Chemicals and Drugs**) and these are not compatible with the LOINC use of this relationship. In this case, there seems to be a mismatch in the meaning of the relationship itself. A definition of the relationship from the developers might assist in ensuring a better mapping.

6. Conclusions

The problems that were revealed by the auditing process described in this paper not only highlight some specific problems and errors in our mapping, but they also lead us to make a number of recommendations. First, and, perhaps, most helpful for ensuring consistency of mapping between terminologies, would be a recommendation that developers explicitly define not only the concepts in their terminologies, but also the relationships that link those concepts. Any terminology alignment effort would benefit enormously if all terminology developers would agree to this basic requirement. Second, just as we identified some problems in the Metathesaurus source vocabularies, we also identified some possible improvements to the Semantic Network. The Semantic Network would benefit from being extended with several new relationships and with new relations for some existing relationships. Finally, the UMLS editing environment could take advantage of the correspondence established between relationships in the Metathesaurus and the Semantic Network and could potentially validate new relations as they enter the system, rather than relying exclusively on a post-processing auditing step.

In this paper we developed a semantically-based method for auditing associative relations in biomedical terminologies. Importantly, these terminologies participate in the Unified Medical Language System (UMLS). This has the consequence that each of the terminologies has been enriched in a variety of ways. For our purposes, the enrichment of concepts by semantic types is the critical foundation on which we have built what we believe to be a novel auditing method. While our auditing was specifically directed to the results of a process that mapped associative relationships from a variety of sources to the UMLS Semantic Network, in principle, it could be applied to the mapping, or alignment, of any set of associative relationships to any other set. The only requirement would be that the participating terminologies have benefited from the semantic typing of their concepts. If that criterion has been met, then the auditing process can take advantage of our methodology for creating and subsequently comparing the semantic signatures of the relationships that have been mapped to each other.

Our auditing process revealed a certain level of consistency in our mapping, but it also uncovered a number of problems. This is exactly the role of an auditing process. Ideally, the process validates the work that has been done, but when it does not, it highlights areas for improvement. The auditing process will only be successful if it is seen as an iterative, rather than a one-time process. That is, once the auditing identifies the problems, attempts should be made to resolve them, and then the auditing cycle should begin again.

Acknowledgments

The authors thank the anonymous reviewers for their thoughtful and extensive comments. This research was supported in part by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM).

References

[1] Bodenreider O. Strength in numbers: exploring redundancy in hierarchical relations across biomedical terminologies. AMIA Annu Symp Proc 2003:101–5.
[2] Ceusters W, Smith B, Kumar A, Dhaen C. Ontology-based error detection in SNOMED-CT. Medinfo 2004;11(Pt 1):482–6.
[3] Cimino JJ. Auditing the Unified Medical Language System with semantic methods. J Am Med Inform Assoc 1998;5(1):41–51.
[4] Cimino JJ, Min H, Perl Y. Consistency across the hierarchies of the UMLS Semantic Network and Metathesaurus. J Biomed Inform 2003;36(6):450–61.
[5] Gu H, Perl Y, Elhanan G, Min H, Zhang L, Peng Y. Auditing concept categorizations in the UMLS. Artif Intell Med 2004;31(1):29–44.
[6] Guarino N. The role of identity conditions in ontology design. Spatial information theory. Cognitive and computational foundations of geographic information science (LNCS 1661). Berlin/Heidelberg: Springer; 1999. p. 221.
[7] Min H, Perl Y, Chen Y, Halper M, Geller J, Wang Y. Auditing as part of the terminology design life cycle. J Am Med Inform Assoc 2006;13(6):676–90.
[8] 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.
[9] 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.
[10] Stearns MQ, Price C, Spackman KA, Wang AY. SNOMED clinical terms: overview of the development process and project status. Proc AMIA Symp 2001:662–6.
[11] Rosse C, Mejino JLV. A reference ontology for biomedical informatics: the Foundational Model of Anatomy. J Biomed Inform 2003;36(6):478–500.
[12] Smith B, Rosse C. The role of foundational relations in the alignment of biomedical ontologies. Medinfo 2004;11(Pt 1):444–8.
[13] ANSI/NISO. Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies (ANSI/NISO Z39.19-2005). Bethesda, Maryland, USA: NISO Press; 2005.
[14] Smith B, Ceusters W, Klagges B, Kohler J, Kumar A, Lomax J, et al. Relations in biomedical ontologies. Genome Biol 2005;6(5):R46.
[15] Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004;32:D267–70 [Database issue].
[16] Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf Med 1993;32(4):281–91.
[17] UMLS Metathesaurus Fact Sheet. Available from: http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html.
[18] UMLS Semantic Network Fact Sheet. Available from: http://www.nlm.nih.gov/pubs/factsheets/umlssemn.html.
[19] McCray AT, Nelson SJ. The representation of meaning in the UMLS. Methods Inf Med 1995;34(1–2):193–201.
[20] McCray AT, Burgun A, Bodenreider O. Aggregating UMLS semantic types for reducing conceptual complexity. Medinfo 2001;10(Pt 1):216–20.
[21] Bodenreider O, McCray AT. Exploring semantic groups through visual approaches. J Biomed Inform 2003;36(6):414–32.
[22] Cornet R, Abu-Hanna A. Two DL-based methods for auditing medical terminological systems. AMIA Annu Symp Proc 2005:166–70.
[23] Hartel FW, de Coronado S, Dionne R, Fragoso G, Golbeck J. Modeling a description logic vocabulary for cancer research. J Biomed Inform 2005;38(2):114–29.
[24] Noy NF, de Coronado S, Solbrig H, Fragoso G, Hartel FW, Musen MA. Representing the NCI Thesaurus in OWL DL: Modeling tools help modeling languages (Technical Report SMI-2008-1308): Stanford Center for Biomedical Informatics Resarch; 2008.
[25] Rector A, Rogers J. Ontological and practical issues in using a description logic to represent medical concept systems: experience from GALEN. Reasoning Web 2006;4126:197–231.
[26] Stevens R, Aranguren ME, Wolstencroft K, Sattler U, Drummond N, Horridge M, et al. Using OWL to model biological knowledge. Int J Hum Comp Stud 2007;65(7):583–94.
[27] Rector AL. What's in a code? Towards a formal account of the relation of ontologies and coding systems. Medinfo 2007;12(Pt 1):730–4.
[28] Sowa JF. Knowledge representation: logical, philosophical, and computational foundations. Pacific Grove: Brooks Cole Publishing Co.; 2000.
[29] Perl Y, Chen Z, Halper M, Geller J, Zhang L, Peng Y. The cohesive metaschema: a higher-level abstraction of the UMLS Semantic Network. J Biomed Inform 2002;35(3):194–212.
[30] RDF Schema. Available from: http://www.w3.org/TR/rdf-schema/.
[31] OWL. Available from: http://www.w3.org/TR/owl-ref/.
[32] Chen Y, Gu HH, Perl Y, Geller J. Structural group-based auditing of missing hierarchical relationships in UMLS. J Biomed Inform 2009;42(3):452–67.

[33] Bodenreider O, Smith B, Kumar A, Burgun A. Investigating subsumption in SNOMED CT: an exploration into large description logic-based biomedical terminologies. Artif Intell Med 2007;39(3):183–95.
[34] Ceusters W, Smith B, Kumar A, Dhaen C. Mistakes in medical ontologies: where do they come from and how can they be detected? Stud Health Technol Inform 2004;102:145–63.
[35] Campbell KE, Tuttle MS, Spackman KA. A "lexically-suggested logical closure" metric for medical terminology maturity. Proc AMIA Symp 1998:785–9.
[36] Bodenreider O. Circular hierarchical relationships in the UMLS: Etiology, diagnosis, treatment, complications and prevention. Proc AMIA Symp 2001:57–61.
[37] Mougin F, Bodenreider O. Approaches to eliminating cycles in the UMLS Metathesaurus: Naïve vs. formal. AMIA Annu Symp Proc 2005:550–4.
[38] Cohen B, Oren M, Min H, Perl Y, Halper M. Automated comparative auditing of NCIT genomic roles using NCBI. J Biomed Inform 2008;41(6): 904–13.
[39] Bean CA, Green R, editors. Relationships among knowledge structures: vocabulary integration within a subject domain. Dordrecht, Boston: Kluwer Academic Publishers; 2001.
[40] Green R, Bean CA, Myaeng SH, editors. The semantics of relationships: an interdisciplinary perspective. Dordrecht, Boston: Kluwer Academic Publishers; 2002.
[41] McCray AT, Bodenreider O. A conceptual framework for the biomedical domain. In: Green R, Bean CA, Myaeng SH, editors. The semantics of relationships: an interdisciplinary perspective. Dordrecht, Boston: Kluwer Academic Publishers; 2002. p. 181–98.
[42] Schulz EB, Barrett JW, Price C. Semantic quality through semantic definition: refining the Read Codes through internal consistency. Proc AMIA Annu Fall Symp 1997:615–9.
[43] Rogers JE, Price C, Rector AL, Solomon WD, Smejko N. Validating clinical terminology structures: integration and cross-validation of Read Thesaurus and GALEN. Proc AMIA Symp 1998:845–9.
[44] Cornet R, Abu-Hanna A. Description logic-based methods for auditing frame-based medical terminological systems. Artif Intell Med 2005;34(3):201–17.
[45] Vizenor L, Bodenreider O, Peters L, McCray AT. Enhancing biomedical ontologies through alignment of semantic relationships: exploratory approaches. AMIA Annu Symp Proc 2006:804–8.
[46] Yu H, Friedman C, Rhzetsky A, Kra P. Representing genomic knowledge in the UMLS semantic network. Proc AMIA Symp 1999:181–5.
[47] The Future of the UMLS Semantic Network. Available from: http://mor.nlm.nih.gov/snw/.
[48] McCray AT. An upper-level ontology for the biomedical domain. Comp Funct Genomics 2003;4(1):80–4.