Journal of Biomedical Informatics 75 (2017) 129–137

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Evaluating the granularity balance of hierarchical relationships within large biomedical terminologies towards quality improvement

Lingyun Luo[a,*], Ling Tong[a], Xiaoxi Zhou[a], Jose L.V. Mejino Jr[b], Chunping Ouyang[a], Yongbin Liu[a]

[a] School of Computer Science, University of South China, Hengyang, Hunan 421001, China
[b] Structural Informatics Group, Department of Biological Structure, University of Washington School of Medicine, Seattle, WA 98195, USA

ARTICLE INFO

Keywords:
Biomedical terminology
Quality assurance
FMA
SNOMED CT
Granularity balance

ABSTRACT

Organizing the descendants of a concept under a particular semantic relationship may be rather arbitrarily carried out during the manual creation processes of large biomedical terminologies, resulting in imbalances in relationship granularity. This work aims to propose scalable models towards systematically evaluating the granularity balance of semantic relationships. We first utilize "*parallel concepts set (PCS)*" and two features (the length and the strength) of the paths between PCSs to design the general evaluation models, based on which we propose eight concrete evaluation models generated by two specific types of PCSs: single concept set and symmetric concepts set. We then apply those concrete models to the IS-A relationship in FMA and SNOMED CT's Body Structure subset, as well as to the Part-Of relationship in FMA. Moreover, without loss of generality, we conduct two additional rounds of applications on the Part-Of relationship after removing length redundancies and strength redundancies sequentially. At last, we perform automatic evaluation on the imbalances detected after the final round for identifying missing concepts, misaligned relations and inconsistencies. For the IS-A relationship, 34 missing concepts, 80 misalignments and 18 redundancies in FMA as well as 28 missing concepts, 114 misalignments and 1 redundancy in SNOMED CT were uncovered. In addition, 6,801 instances of imbalances for the Part-Of relationship in FMA were also identified, including 3,246 redundancies. After removing those redundancies from FMA, the total number of Part-Of imbalances was dramatically reduced to 327, including 51 missing concepts, 294 misaligned relations, and 36 inconsistencies. Manual curation performed by the FMA project leader confirmed the effectiveness of our method in identifying curation errors. In conclusion, the granularity balance of hierarchical semantic relationship is a valuable property to check for ontology quality assurance, and the scalable evaluation models proposed in this study are effective in fulfilling this task, especially in auditing relationships with sub-hierarchies, such as the seldom evaluated Part-Of relationship.

1. Introduction

Biomedical ontologies and controlled terminologies play a vital role in the field of biomedical informatics, including in image retrieval [1], information extraction [2], and data integration [3], which further demonstrates the importance of Ontology Quality Assurance (OQA) [4,5]. As the key component of biomedical terminologies, semantic relationships inevitably became the hot spot of auditing methods. Bodenreider investigated the use of adjectival modifiers for determining consistency in the UMLS [6]. Gu utilized relationship structures for detecting possible incorrect relationship assignments in FMA [7]. In our past studies, we leveraged lexical-structural information to audit the hierarchical relationships in FMA [8,9]. Zhang also performed lattice-based structural auditing of SNOMED CT [10]. Other than lexical and structural based approaches, Geller [11] and Wei [12] used abstraction networks [13] for quality assurance in the UMLS and SNOMED CT, respectively. We refer to Zhu's review article [14] for more detailed information of early studies in this field.

Under a particular hierarchical relationship, how to select and organize the descendants of a concept is a challenging problem. Usually, the descendants are at levels of semantic granularity finer than that of the concept [15,16]. To keep the relationship consistent, we believe that it is better for the descendants at the same level of hierarchy to stay at a same level of semantic granularity. In other words, the semantic distances between every pair of concepts at two particular levels of semantic granularity are supposed to be the same. However, in reality, although domain expert knowledge was involved during the creation processes, inconsistency and imbalance in relationship granularity may still be introduced into biomedical terminologies. By "granularity of a relationship *r*" in this work we mean the level of detail at which

* Corresponding author.
E-mail address: luoly@usc.edu.cn (L. Luo).

http://dx.doi.org/10.1016/j.jbi.2017.10.001
Received 13 May 2017; Received in revised form 15 September 2017; Accepted 2 October 2017
Available online 04 October 2017
1532-0464/

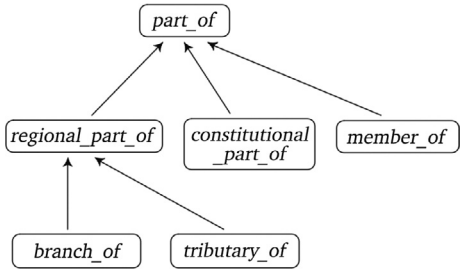

Fig. 1. The three-level hierarchy of the Part-Of relationship in FMA.

knowledge is presented in a biomedical terminology, represented by the semantic distances between *r*-related concepts. The longer the distances, the coarser the granularity, and vice versa.

Weng pointed out that granularity mismatch is one type of semantic mismatches across domain models that could be a bottleneck of semantic harmonization [17]. Richesson also concluded that different granularities between heterogeneous coding systems cause an issue for data interoperability [18], not to mention the imbalanced relationship granularity within a terminology. However, only a limited number of auditing methods on the granularity of relationships exist. Sun and Zhang [19] identified the granularity differences between two biomedical ontologies through rules to distinguish among different types of subclasses and classifications. He et al. [20,21] adopted the term "density" instead of granularity and defined a set of topological patterns to demonstrate different concept densities across pairs of terminologies, aiming to enrich SNOMED CT conceptual content and support semantic harmonization of SNOMED CT with other Unified Medical Language System (UMLS) terminologies.

These past works focused on granularity differences across terminologies using comparative studies, while we propose a novel systematic mechanism for auditing the intrinsic granularity balance of relationships within a targeted biomedical terminology in this paper.

As far as we know, currently there is no effective method for calculating the granularity of semantic relationships directly. Therefore, we choose to examine the granularity balance relatively by utilizing "*parallel concepts set (PCS)*." Two concepts are parallel if they share a similar level of conceptual knowledge. As a result, a PCS contains a number of concepts, which are parallel to each other. The assumption is: If the granularity of a relationship in a terminology is balanced, then all the paths along this relationship from one set of parallel concepts to another one are expected to be balanced in terms of "*semantic distance*," which is represented by the length and the strength of each path. The length of a path is the number of steps, and the strength of it is introduced to deal with cases when the semantic relationship has a hierarchy of subproperties, for example, the Part-Of relationship in FMA has a hierarchy of three levels, as shown in Fig. 1. Subproperties of each hierarchy are assigned a distinct semantic weight and the strength of a path is the aggregation of all the semantic weights of the subproperties along it. The granularity is balanced if the lengths of all the paths are the same, as well as the strengths, if needed. Fig. 2 demonstrates two examples with imbalanced length and imbalanced strength in SNOMED CT and FMA, respectively.

Based on the idea above, taking the length and the strength aspects into account, we first proposed two general models for evaluating the granularity balance of relationships. Then, we selected two kinds of PCSs for experiments: one contains a single concept in the set, as illustrated by the two nodes in Fig. 2(b), and the other one contains a pair of symmetric concepts, as illustrated by the two dashed rectangles in Fig. 2(a). As demonstrated in our previous work [8], symmetric concepts are bisimilar, thus suffice to be parallel concepts. Based on the two types of PCSs, we achieved four groups of specialized evaluation models. At last, we applied these models to the IS-A relationship in FMA and SNOMED CT's Body Structure subsets, as well as to the Part-Of relationship in FMA. Results show that there are 121 instances of imbalances in FMA and 124 of that in SNOMED CT for the IS-A relationship. On the other hand, 6801 instances of imbalances for the Part-Of relationship were also detected, including 2320 length redundancies and 926 strength redundancies. After removing those redundancies and rerunning the algorithms, the total number of imbalances for the Part-Of relationship dramatically reduced by 95%. The final results were automatically classified into missing concepts, misaligned relations, and inconsistencies using algorithms. In addition, manual curation performed by the FMA project leader confirmed that all of the inconsistencies and missing concepts, as well as most of the inappropriate relation assignments identified by our study are correct.

2. Background

2.1. The Foundational Model of Anatomy (FMA)

The FMA is both a theory of human anatomy and an ontology artifact [22]. In particular, it is a theory of the canonical, phenotypic structure of the human organism at all biologically salient levels of granularity. As a theory of canonical anatomy, it ranges over those categories of entities which are idealizations of an organism body and

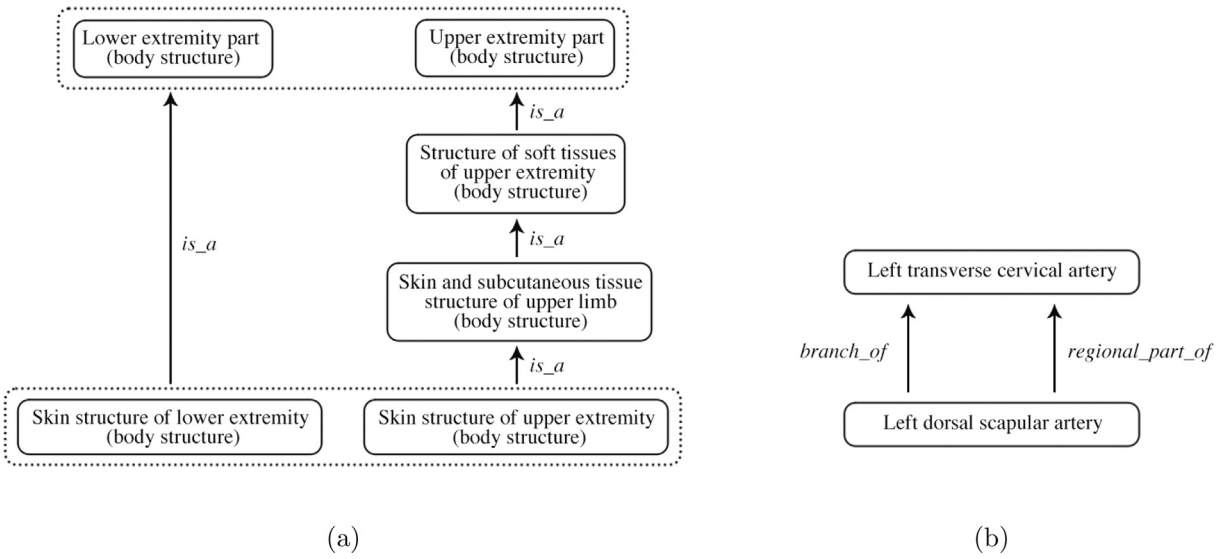

Fig. 2. (a) An example with imbalanced length in SNOMED CT. (b) An example with imbalanced strength in FMA.

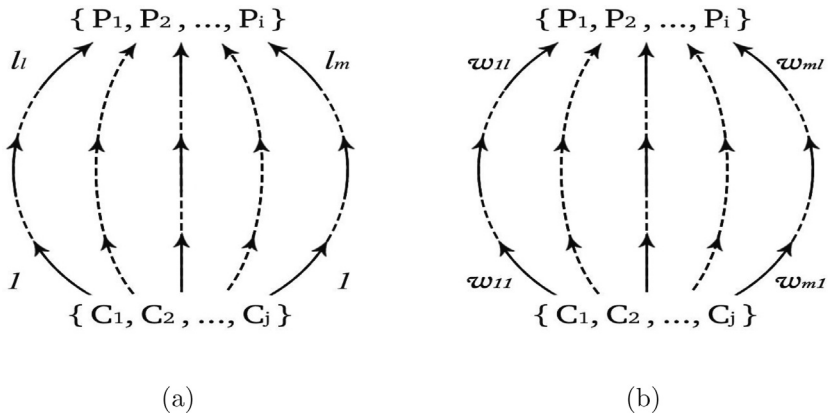

Fig. 3. The general models for evaluating the semantic distances between two PCSs: (a) The l_1: l_2:…: l_m length model. (b) The strength model (all paths have length l).

its typical component parts. As a computational artifact, it is a formal representation of this theory, suitable for machine manipulation. The model underlying the FMA is a frame-based representation with more than 100,000 concepts including macroscopic, microscopic and subcellular canonical anatomy.

For our analysis, we used the version 4.4.0 of FMA created in June, 2016. It is a version distributed as an RDF/XML-based serialization that enables it to be stored in an RDF data store and made available to be queried via SPARQL over internet protocol.

2.2. SNOMED CT

SNOMED CT is the largest clinical terminology in the world [23], currently maintained by SNOMED International [24]. It provides standards for encoding clinical content, ranging from clinical findings, procedures, to diseases and diagnoses in electronic health records (EHRs), aiming to enhance interoperability. Concepts in SNOMED CT are arranged into 19 subhierarchies including Body Structure. In this study, we used the US Edition of SNOMED CT released in March 2016, and focused on its Body Structure subset, which contains almost 30,000 concepts.

3. Methods

3.1. Evaluation models

For any hierarchical relationship, we examined the relative granularity balance by checking the properties of paths between nodes at different levels of conceptual knowledge. If the relationship has a balanced granularity, all the paths from a source node to a target node are expected to be balanced in terms of length and strength. As demonstrated in Fig. 2(b), the source node and the target node are both single concepts, and the paths between them are imbalanced in strength. In Fig. 2(a), the source node and the target node are both sets of *symmetric concepts*, and the paths between them are imbalanced in length. A pair of concepts is called *symmetric* if the concept names are the same except for the possible difference in a single occurrence of modifiers used [8]. For instance, ("*Lower extremity part*," "*Upper extremity part*") is a symmetric concept pair concerning the modifier pair *Upper* and *Lower*.

As noticed, two different types of end nodes were encountered in the examples of Fig. 2. In this study, to give a scalable mechanism for evaluating relationship granularity, we generalized the nodes to "*PCSs*," which are comprised of concepts sharing the same level of conceptual knowledge, such as the symmetric concepts in Fig. 2(a). Leveraging PCSs, we first introduced the formal evaluation models and then specialized them using two concrete types of PCSs.

3.1.1. Formal evaluation models

We audited the granularity balance in a relative way by evaluating the semantic distances between PCSs. For two PCSs $\{P_1,P_2,...,P_i\}$ and $\{C_1,C_2,...,C_j\}$ and a semantic relationship r, if there exist m ($m \geqslant 2$) distinct r-paths (paths with only the relationship r along them) from $\{C_1,C_2,...,C_j\}$ to $\{P_1,P_2,...,P_i\}$, we define the length of the k-th path as the number of steps along it, denoted as $l_k (1 \leqslant k \leqslant m)$. If there exists $k_1 \neq k_2$ such that $l_{k_1} \neq l_{k_2}$, which means, the semantic distances are imbalanced in length, r is determined as imbalanced in length. Furthermore, if all the r-paths have the same length, denoted as l, and r has a hierarchy of subproperties, we define the strength of the k-th path as $\sum_{h=1}^{l} w_{kh} (1 \leqslant k \leqslant m$ and $1 \leqslant h \leqslant l)$, where w_{kh} represents the semantic weight of the h-th step along the k-th path. The strengths of the m paths are expected to be the same if the relationship r is balanced in granularity.

Fig. 3 illustrates the general models for evaluating the semantic distances from $\{C_1,C_2,...,C_j\}$ to $\{P_1,P_2,...,P_i\}$ in terms of length and strength, respectively. The left model is denoted as an l_1: l_2:…: l_m model, which evaluates the granularity balance of r by length, and the right model evaluates the granularity balance of r by strength. Please note that the paths may have intersections and that the concepts in the end nodes (PCSs) may appear in the paths too.

In the remainder of the paper, we specialized the models in Fig. 3 and focused on the 1:n ($n \geqslant 1$) models instead of the generalized l_1: l_2:…: l_m models due to the fact that most of the m:k ($m, k \geqslant 1$) models have intermediate PCSs in the middle of the two paths, which will essentially divide them into 1:n models.

3.1.2. Specialized evaluation models

Two types of PCSs were selected to evaluate the granularity balance of semantic relationships in both FMA and SNOMED CT: (1) Type I: single concept set containing only one concept and (2) Type II: symmetric concepts set containing a pair of symmetric concepts. As demonstrated in our previous work [8], symmetric concepts are bisimilar, thus suffice to be parallel concepts. The four combinations of the two types generate four cases of evaluation models, which are demonstrated in Fig. 4. For each case, there is a 1:l model on the left for evaluating the relationship by length, and on the right, there is a 1:1 model for evaluating the relationship by strength, where r and r' represent two subproperties of the same relationship at different levels, if they exist.

Note that for each case, the labels along the paths in the left graph represent the number of steps, while the labels along the paths in the right graph represent two subproperties at different levels of the same relationship. Besides, the two paths in each model of Case (II) should aim to different destination concepts, otherwise, the case will be covered by Case (I). For the same reason, the two paths in the models of Case (III) should depart from different source concepts. Also, the paths in the models of Case (IV) do not converge because otherwise the scenarios are covered by the former three cases.

In Cases (II), (III) and (IV), we use the pair (u,v) to stand for a pair of symmetric modifiers such as *(left, right)*. Putting the symmetric

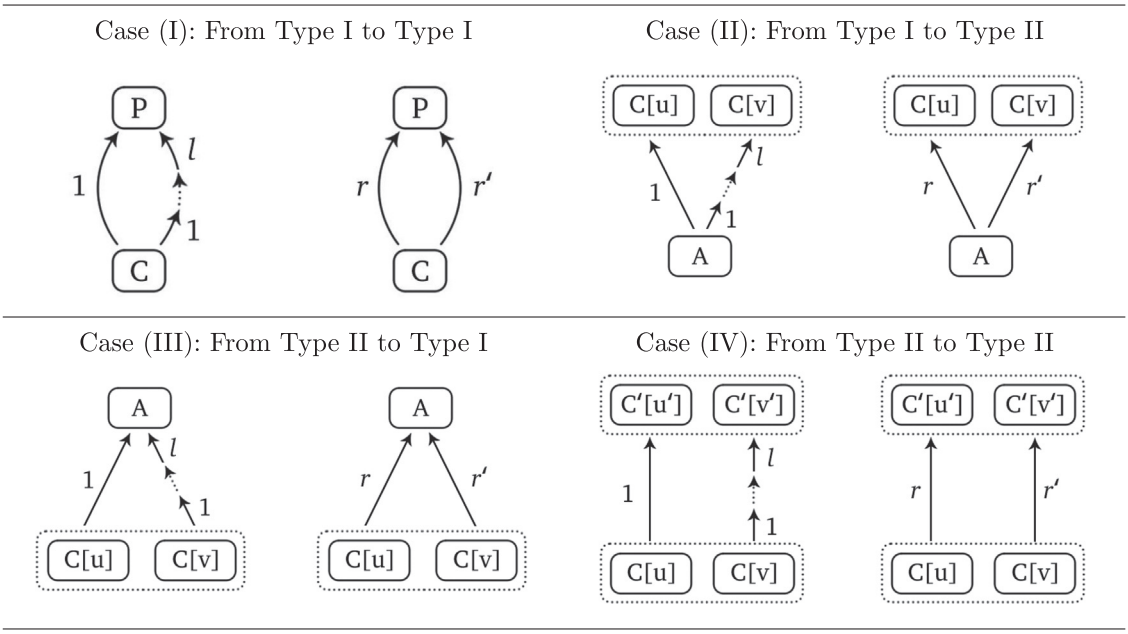

Fig. 4. The four cases of specialized evaluation models. For each case, there is a 1:*l* model on the left for evaluating the relationship by length, and on the right, there is a 1:1 model for evaluating the relationship by strength. Dashed rectangles represent PCSs.

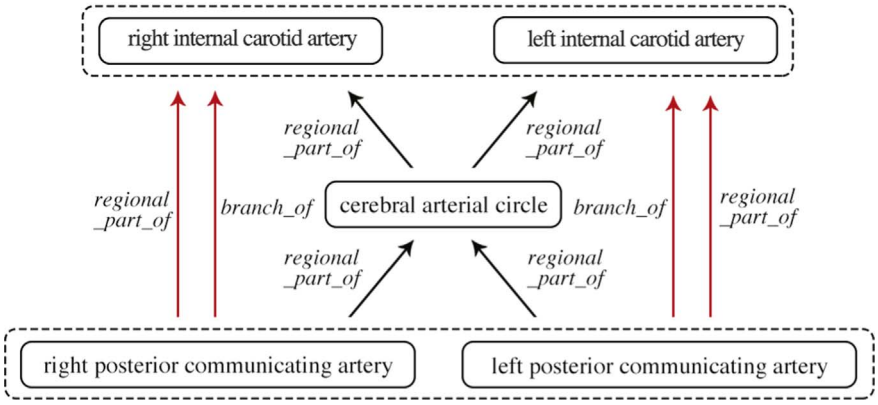

Fig. 5. A mixed-case instance for the Part-Of relationship in FMA. The relations indicated by red arrows are redundant.

modifiers into the same context C results in a symmetric concept pair $(C[u],C[v])$, such as *(left heart, right heart)*. Each symmetric concept pair $(C[u],C[v])$ forms a PCS.

Recall the examples in Fig. 2, we can discover that Fig. 2(a) is an instance of Case (IV) 1:3 length model, and Fig. 2(b) is an instance of Case (I) strength model.

For simplicity, in the remainder of the paper, we call the model that evaluates the granularity balance of relationships by length *the length model*, and correspondingly for *the strength model*. For a semantic relationship r with a hierarchy of subproperties such as Part-Of in FMA, when evaluating its granularity balance using length models, we consider all of the subproperties as r since every subproperty r' implies r.

3.2. Redundancy removal

One observation of the models in Fig. 4 is that all Case (I) instances indicate redundancies based on the following two facts.

1. The transitive feature of a relationship r states that
   If (A,r,B) and (B,r,C), then (A,r,C).
   Since both IS-A and Part-Of are transitive, the direct path with length 1 in the 1:*l* length model of Case (I) (see Fig. 4) is redundant.
2. Suppose r' is a subproperty of r, then

   (C,r',P) implies (C,r,P),

   which indicates that the (C,r,P) path is redundant in the strength model of Case (I). Fig. 2(b) illustrates an instance of this case: Since *branch_of* is a subproperty of *regional_part_of* (see Fig. 1), the edge *regional_part_of* is redundant.

In the remainder of the paper, redundancies implied by the first fact are named length redundancies and those implied by the second fact are named strength redundancies.

As argued in FEDRR [25], "*The principle of parsimony in ontological modelling is a direct consequence of closed-world assumption (CWA). It refers to the fact that relations implied by the transitive property of a relationship, …, must not be explicitly stated…. Detecting redundant relations is an important task for OQA,*" it is necessary to remove redundancies from ontologies for quality improvement.

Another observation is that there may be instances with mixed cases, so the results in the four cases are not necessarily disjoint if only concept nodes are concerned. An instance is depicted in Fig. 5: It contains 1:2 length imbalances from all the four cases in Fig. 4, as well as Case (I) strength imbalances.

Based on the above two observations, both length redundancies and strength redundancies in Case (I) may introduce imbalanced instances into the other three cases. For instance, if the four redundant Part-Of

relations indicated by red arrows in Fig. 5 were removed, there would be no granularity imbalance in this graph any more. Hence, in order to analyse the influence of redundancies on the total results of imbalances, without loss of generality, we select the complex Part-Of relationship in FMA as experiment subject and conduct two additional rounds of granularity balance checking after removing length redundancies and further strength redundancies.

3.3. Implementation

We checked the granularity balance for the IS-A relationship in FMA and SNOMED CT's Body Structure subset, as well as for the Part-Of relationship in FMA. Since there is no subproperty for IS-A, only length models were needed to evaluate its granularity balance in both FMA and SNOMED CT.

3.3.1. Preprocessing of data

The OWL version v4.4.0 of FMA downloaded from its official site was stored in Virtuoso [26] and queried using SPARQL language [27], and the US Edition of SNOMED CT released in March 2016 was stored in MySQL for computation. Due to the large volume and complex structures of the two terminologies, only the 1:2 and the 1:3 patterns for length models were considered in our experiments. Several datasets were prepared beforehand:

1. For the IS-A relationship: we computed a set $\mathbf{IR_1}$ which stores concept pairs related by one step of IS-A, a set $\mathbf{IR_2}$ which stores concept pairs indirectly related by two steps of IS-A's, and a set $\mathbf{IR_3}$ which stores concept pairs indirectly related by three steps of IS-A's.
2. For the Part-Of relationship: we computed a set $\mathbf{PR_1}$ which stores concept pairs related by one step of Part-Of, a set $\mathbf{PR_2}$ which stores concept pairs indirectly related by two steps of Part-Of's, and a set $\mathbf{PR_3}$ which stores concept pairs indirectly related by three steps of Part-Of's.
3. To retrieve Type II PCSs, i.e., symmetric concepts sets, we computed a set **P** which stores 11 modifier pairs: In order to retrieve all the modifier pairs, we first obtained all the class names in FMA and used the Stanford Parser [28] to obtain all the Noun-Phrase (NP) chunks without prepositions. For all the modifiers in those 21,616 NP chunks, any two of them that share a common context were selected out to form a modifier pair. Finally, we ranked all the modifier pairs by the number of common contexts the two members share. It turned out that there are 23 modifier pairs whose two members appeared spontaneously in at least 100 distinct contexts, generating 7867 PCSs. To reduce computational complexity, we only chose the first 11 pairs to form the set **P**: *(left, right), (anterior, posterior), (lateral, medial), (first, second), (second, third), (first, third), (superior, inferior), (lateral, anterior), (fourth, third), (fourth, second), (upper, lower)*, which generate 6168 PCSs (78.4% out of 7867) in FMA. Without loss of generality, we restricted the symmetric concepts observed in this study to those generated by **P** only.

The imbalance-checking algorithms were designed based on the cases in Fig. 4 and quite routine, thus omitted in this paper.

3.3.2. Whole procedure

The three rounds of computations are briefly described as follows:

- Apply the imbalance-checking algorithms to the IS-A relationship in FMA and SNOMED CT, as well as to the Part-Of relationship in FMA. Output the original results. The two additional rounds of computation are conducted on the Part-Of relationship only. It will be the same procedure for the IS-A relationship.
- Round 2: Remove length redundancies (including the 1:2 and the 1:3 patterns) detected by Round 1 from FMA. Re-calculate $\mathbf{PR_1}$,$\mathbf{PR_2}$ and $\mathbf{PR_3}$. Rerun the algorithms. Update results.
- Round 3: Further remove strength redundancies in Case (I) from the updated results. Re-calculate $\mathbf{PR_1}$,$\mathbf{PR_2}$ and $\mathbf{PR_3}$. Rerun the algorithms. Output the final results.

4. Results

After applying the preprocessing procedure to FMA, we found 104,226 concept pairs directly related by one IS-A in $\mathbf{IR_1}$, 104,064 concept pairs related by two-step IS-A's in $\mathbf{IR_2}$, and 103,702 concept pairs related by three-step IS-A's in $\mathbf{IR_3}$. Those numbers for $\mathbf{PR_1}$,$\mathbf{PR_2}$ and $\mathbf{PR_3}$ in FMA are 61,273, 82,681 and 123,893, respectively. On the other hand, the numbers of elements in $\mathbf{IR_1}$,$\mathbf{IR_2}$ and $\mathbf{IR_3}$ for IS-A in SNOMED CT's Body Structure subhierarchy are 38,976, 64,062, and 106,769, respectively. They are all presented in Tables 1 and 2.

Table 1
Results for IS-A in FMA and SNOMED CT's body structure subset.

Terminologies	Pair counts			Case I		Case II		Case III		Case IV		Total	
	IR_1	IR_2	IR_3	1:2	1:3	1:2	1:3	1:2	1:3	1:2	1:3	1:2	1:3
FMA	104,226	104,064	103,702	15	3	2	–	61	11	29	–	107 (88%)	14 (12%)
SNOMED CT	38,976	64,062	106,769	1	–	5	1	75	11	24	7	105 (85%)	19 (15%)

Table 2
Comparative results for the three rounds of examination on the Part-Of relationship in FMA.

Models	Pair counts			Case I		Case II		Case III		Case IV		Total
	PR_1	PR_2	PR_3	1:2	1:3	1:2	1:3	1:2	1:3	1:2	1:3	
Round 1	61,273	82,681	123,893	1807	513	107	123	1000	208	1048	129	(Rnd1: 4935)
Round 2	59,513	75,781	106,036	–	–	19	15	222	106	156	17	(Rnd2: 535)
Round 3	58,620	71,676	95,283	–	–	19	15	123	22	95	17	(Rnd3: 291)
Strength	Regional-Part-Of vs. branch-of			850/830/–		3/3/1		319/315/5		502/483/5		
	Regional-Part-Of vs. tributary-of			72/59/–		5/5/5		64/31/7		36/36/6		(Rnd1: 1866)
	Part-Of vs. regional-Part-Of			–		–		3/3/3		4/4/4		(Rnd2: 1777)
	Part-Of vs. constitutional-Part-Of			4/4/–		–		–		4/4/–		(Rnd3: 36)

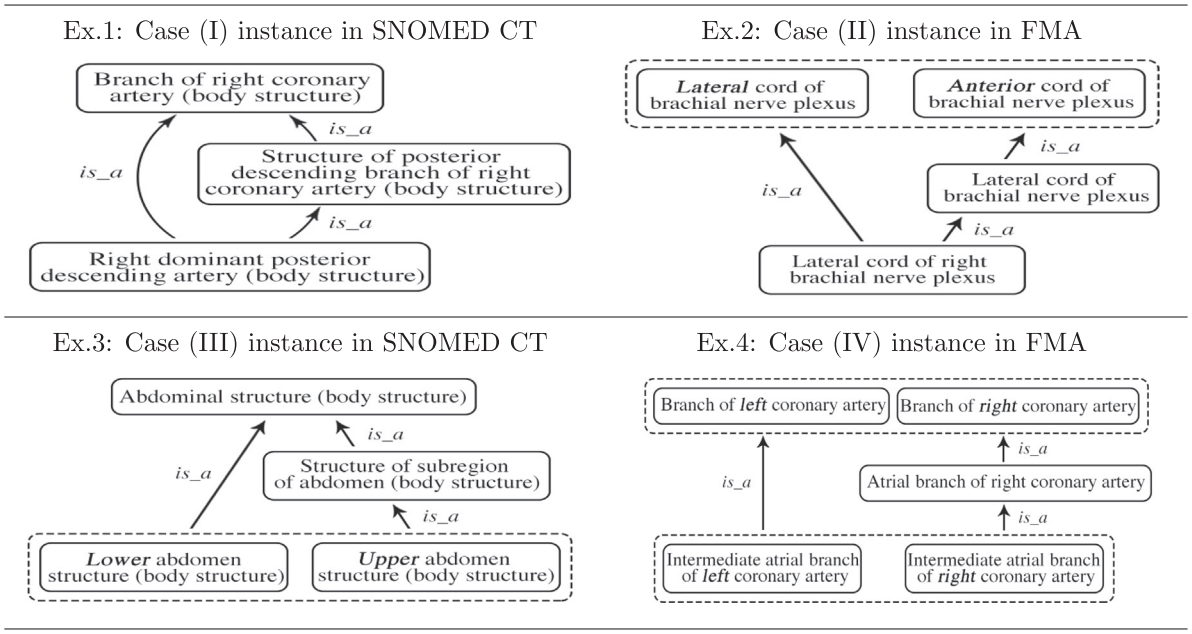

Fig. 6. Length-imbalanced instances for the IS-A relationship: the left two examples are in SNOMED CT, and the right two examples are in FMA.

4.1. Examination on the IS-A relationship in both FMA and SNOMED CT's Body Structure sub-hierarchy

For the IS-A relationship, Table 1 illustrates the numbers of length imbalances of the four cases in FMA and SNOMED CT's Body Structure subset separately after Round 1. For each case, the length imbalances are divided into two patterns: the 1:2 pattern and the 1:3 pattern. The number of instances for each pattern is presented individually. As shown in Table 1, more than 85% of the instances take the 1:2 pattern, in both FMA and SNOMED CT, and more than 50% of them are Case (III) instances. On the other hand, considering the vast number of concept pairs in $\mathbf{IR_1}$,$\mathbf{IR_2}$ and $\mathbf{IR_3}$, it is remarkable that there are only about 100 length imbalances in both terminologies, which may be owed to the frequent quality assurance attempts on the IS-A relationship in the past.

Fig. 6 demonstrates four examples with imbalanced length for the IS-A relationship in the two terminologies. To save space, only one example is presented for each case. Note that there are two duplicate nodes *Lateral cord of brachial nerve plexus* in Ex.2, and the 1:2 length imbalance is actually caused by the triple *(Lateral cord of brachial nerve plexus, is_a, Anterior cord of brachial nerve plexus)*, which is not correct.

Since all Case (I) instances indicate redundancies, Table 1 shows that there are 18 redundancies in FMA and 1 redundancy in SNOMED CT's Body Structure subset concerning the IS-A relationship. We manually removed them from the two terminologies and found that instances of the other three cases remain unaffected, which indicates that these redundancies do not mix with the other cases.

4.2. Three rounds of examinations on the Part-Of relationship in FMA

Table 2 presents the comparative results for the three rounds of evaluation on the Part-Of relationship in FMA using the length models and the strength models of the four cases (see Fig. 4). Round 1 presents the original results after the first round of computation. Round 2 and Round 3 present the computation results after removing length redundancies and further strength redundancies, respectively. The upper part of Table 2 illustrates the distribution of the length-imbalanced instances on the 1:2 and the 1:3 patterns from Case (I) to Case (IV), separated by horizontal lines for each round. The bottom part of Table 2 demonstrates the distribution of the strength-imbalanced instances on different pairs of subproperties of the Part-Of relationship in the four cases, separated by '/' for each round. The last column demonstrates the total numbers of imbalances after each computation round.

Results of Round 1 show that for length imbalances, 3962 out of the total 4935 (80%) instances are 1:2 length imbalances. Among all the four cases, Case (I) turns out to be the most common one (almost 50%), and Case (II) has the least number of instances. For strength imbalances, although the Part-Of relationship in FMA has six subproperties (including itself) on three levels (see Fig. 1), the strength imbalances only take place in the following four pairs: *regional_part_of vs. branch_of, regional_part_of vs. tributary_of, part_of vs. regional_part_of* and *part_of vs. constitutional_part_of,* and nearly 90% of them occur between *regional_part_of* and *branch_of*. Besides, more than a half of all the instances with imbalanced strengths are Case (I) imbalances, while Case (II) occupies only a negligible portion.

As Table 2 shows, the number of length imbalances is reduced by 89% (from (4935 − 535)/4935) after Round 2 and by another 5% (from (535 − 291)/4935) after Round 3. Also, the number of strength imbalances is reduced by 5% (from (1866 − 1777)/1866) after Round 2 and by another 93% (from (1777 − 36)/1866) after Round 3. These results illustrate that around 90% of the length imbalances are introduced by length redundancies, and the same for strength imbalances. Other types of anomalies contributing to the remaining imbalances are analysed in the following section.

5. Evaluation of results

We first perform automatic evaluation for the IS-A and the Part-Of results, and then conduct manual curation on them. As shown by Table 1, there are 121 instances of imbalances in FMA and 124 of that in SNOMED CT for the IS-A relationship. Besides, Table 2 shows that there remain 291 length imbalances and 36 strength imbalances for the Part-Of relationship after eliminating all the redundancies at the end of Round 3. To validate these potential anomalies, we analyse their abnormal types using automatic evaluation algorithms based on their different features.

Inconsistencies. The 36 strength imbalances from Case (II) to Case (IV) (see Fig. 4) indicate inconsistencies. An instance of Case (III) is shown in Ex.1 of Fig. 7: The two triples are *(Right hemidiaphragm, regional_part_of, Diaphragm)* and *(Left hemidiaphragm, part_of, Diaphragm)*.

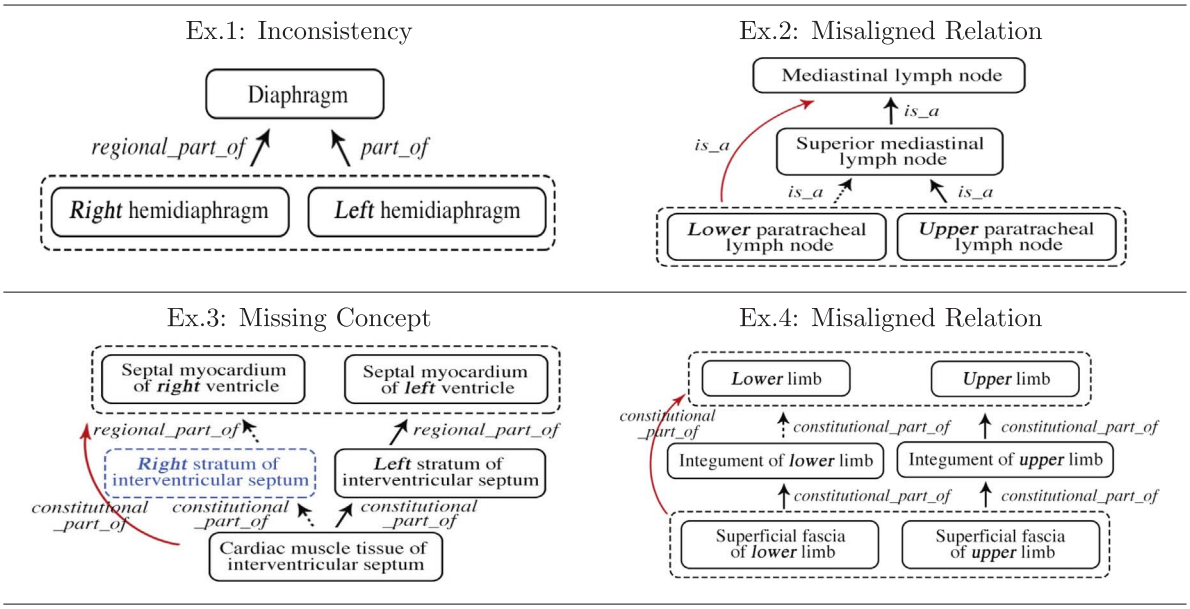

Fig. 7. Instances of abnormal types for 1:2 length imbalances. The concept in blue indicates a missing concept in FMA. Dashed arrows represent missing relations. Red arrows represent direct paths to be further removed. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Although *regional_part_of* implies *part_of*, the reverse does not hold true, which means, *(Left hemidiaphragm, regional_part_of, Diaphragm)*does not necessarily exist, thus introduces inconsistency. A suggestion for remediation is to change the relation *part_of* to *regional_part_of*, which was verified by an FMA expert.

Missing concepts, and misaligned relations. Length imbalances from Case (II) to Case (IV) suggest the possibility of introducing new intermediate concepts and corresponding relations to the terminology, or deleting existing concepts and relations if the granularity is supposed to be coarse. Usually, the former suggestion is more feasible. As a result, for each intermediate concept node in the 1:2 and 1:3 length imbalances, we automatically check if its symmetric concept exists in the terminology using algorithms.

1. If the targeted intermediate concept name does not contain any related modifier, the symmetric concept of it is supposed to be itself. An example is shown in Ex.2 of Fig. 7. In this example, the modifier pair of interest is *(upper, lower)*. Since the concept *Superior mediastinal lymph node* does not contain any element from the pair, it takes itself as its symmetric concept in this context. To keep the granularity balanced, there supposes to be a relation *is_a* from *Lower paratracheal lymph node* to *Superior mediastinal lymph node*. Therefore, instances in this group can be viewed as misaligned relations.
2. If the symmetric concept for the intermediate concept does not exist in the terminology, the instance will be classified as missing concept. As illustrated by Ex.3 of Fig. 7, there may need to be a new concept named *Right stratum of interventricular septum* and the two corresponding relations.
3. If the symmetric concept for the intermediate concept exists in the terminology but the relations around it do not keep the granularity balanced, it is a case indicating misaligned relations, too. For instance, in Ex.4 of Fig. 7, although *Integument of lower limb* exists in FMA, the relation *constitutional_part_of* does not exist from *Integument of lower limb* to *Lower limb*, thus introduces a misalignment.

Note that adding the suggested new concepts and relations to the terminology will make the direct paths (represented as red arrows in Fig. 7) redundant again. As thus, redundancy removal needs to be conducted at last.

Using our classification algorithms, for the Part-Of relationship, we detected 39 missing concepts and 198 misalignments in the 237 1:2 length imbalances. In addition, for the 54 1:3 length imbalances, since each instance has two intermediate nodes, the total number of intermediate concepts is 108, among which 12 missing concepts and 96 misalignments were identified by our algorithms, shown in Table 3. For the IS-A relationship, 34 missing concepts and 80 misalignments in FMA as well as 28 missing concepts and 114 misalignments in SNOMED CT were uncovered, as also presented in Table 3.

Table 3
Automatic evaluation on the results.

	FMA (IS-A)		FMA (Part-Of)		SNOMED CT (IS-A)	
	1:2	1:3	1:2	1:3	1:2	1:3
Missing concepts	33	1	39	12	22	6
Misalignments	59	21	198	96	82	32

All the results for FMA along with suggested remediation have been reported to the FMA project leader, and will be used to implement the corrections in FMA. As a result of manual curation, he confirmed that all of the inconsistencies, missing concepts and most of the inappropriate relation assignments identified by our methods are correct. Those misaligned relations incorrectly identified are caused by asymmetry in some body parts. For example, the mediastinum is divided into the superior and inferior mediastinum, of which the latter is larger, and the inferior mediastinum is further divided into the anterior, middle and posterior mediastinum. Hence, although the triple *(content of anterior mediastinum, regional_part_of, content of inferior mediastinum)* exists in FMA, there should not be a *regional_part_of* from *content of anterior mediastinum* to *content of superior mediastinum*. As the effectiveness of our models in identifying curation errors was confirmed by the FMA expert, the results for SNOMED CT will also be reported to SNOMED International to help with quality improvement.

6. Discussions

In this study, we proposed a systematic mechanism for evaluating the granularity balance of hierarchical semantic relationships within large biomedical terminologies for quality assurance. Applications of the evaluation models were conducted on the IS-A relationship in both

FMA and SNOMED CT, as well as on the Part-Of relationship in FMA. A considerable number of imbalanced instances was uncovered. Further analysis shows that above 90% of the imbalanced instances for the Part-Of relationship in FMA are redundancies or are caused by them, and the remaining instances are either inconsistencies, missing concepts or misaligned relations. The results have been reported to an ontology expert and will be used to implement the corrections in further versions of FMA. In all, the granularity balance of hierarchical semantic relationship is a valuable property to check for ontology quality assurance, and the scalable evaluation models proposed in this study are effective in fulfilling this task, especially in auditing relationships with sub-hierarchies, such as the seldom evaluated Part-Of relationship.

The approach for examining the granularity balance of hierarchical semantic relationships proposed in this paper relies on the discovery of PCSs. A single concept is surely parallel to itself, but how to find more parallel concepts? It depends on the unique features of each terminology and the knowledge field it describes. For instance, based on the largely bilaterally symmetric features of the human anatomy, we found that symmetric concepts in FMA and SNOMED CT's Body Structure subsets suffice to be parallel concepts proposed in this study. To discover more parallel concepts, human efforts may be involved when necessary. Another question that arises is: Do we really need PCSs with more than two elements. Since granularity imbalances are discovered when the paths between PCSs are imbalanced, and each path has only one source and one destination, essentially the granularity imbalances happen between pairs of paths. As a result, PCSs with two elements are enough for uncovering granularity imbalances.

There are two reasons for not combining the length model and the strength model together: One is that they discover different aspects of granularity imbalance, including redundancies, inconsistencies and revision suggestions, as demonstrated in the results. The other reason is that the semantic weight assigned to a relationship may prevent the discovery of length imbalances. For instance, suppose there is a subproperty r_1 with semantic weight k and another subproperty r_2 with semantic weight 1, then the semantic distance of one r_1 will be equal to that of k r_2's, which prevents the detection of 1:k length imbalances.

The two types of models introduced in this work can be used separately. For instance, only the length models were applied to SNOMED CT in our experiments. As a matter of fact, the length models can be applied to any hierarchical relationship and the strength models can be applied to any relationship with a sub-hierarchy, indicating the scalability of our methods.

Comparison with prior work. Our work is close to that of He [20,21] in that we both leveraged semantic structures between two end nodes (concepts or PCSs) to investigate potential problems in biomedical terminologies, but differs from it in three aspects: Firstly, their study focuses on differentiation across terminologies, while we pay attention to the intrinsic imbalances of semantic relationships inside terminologies. Secondly, the set of topological patterns proposed in [21] only illustrates cases of length imbalances but no strength imbalances, because they only consider the IS-A relationship. Lastly, the end nodes for structurally congruent concepts are single concepts in [20,21] while they are PCSs in this study, which shows that our models are more general.

In our previous work [8], we proposed a principled ontology auditing approach based on structural bisimilarity between symmetric concepts, and provided exhaustive evaluation on the IS-A relationship in FMA. Although symmetric concepts are also leveraged in this study to form PCSs, the underlying theory basis is totally distinct from that in [8]. Besides, if we were to view the Matches in [8] from the perspective of granularity balance, they would be 0:1 or 1:1 length models.

Agrawal [29] defined positional similarity sets as *sets of lexically similar concepts that differ from each other by exactly one word at the exact same position of their fully specified names*, which is similar to the Type II PCSs (symmetric concepts sets) in our study. The differences are: (1) Unlike symmetric concepts, the lexically similar concepts in a positional similarity set may not suffice to be parallel concepts. (2) The modifier pairs in our symmetric concepts were extracted from the existing terminology with certain conditions, while the differing word in a positional similarity set can be any word.

As a by-product, this study detected a significant number of redundancies for the Part-Of relationship in FMA. As a matter of fact, many studies in the literature focused on auditing redundancies in large biomedical terminologies, such as UMLS [30,31] and GO [32]. Gu [30] checked multiple relationships between a given pair of concepts to identify erroneous modeling in the UMLS, which differs from our study in that the redundancies we examined were caused by the same relationship other than different ones. Mougin also checked multiply-related concepts in the UMLS [31] as well as redundant relations in the Gene Ontology (GO) [32]. The relations with granularity differences identified in the UMLS [31] and the redundancies generated by the same relationships in GO [32] presented in Mougin's studies can both be viewed as specific cases covered by our models. Moreover, we are not aware of studies that take strength redundancies into consideration. To the contrary, our study discovered both length redundancies and strength redundancies from a new aspect: the granularity balance of semantic relationships, and automatically analysed the influence of redundancies on granularity imbalances as well.

Limitations. Although the models presented in this study are general and comprehensive, taking the large sizes of terminologies and the complex structures of the Part-Of relationship into account, we made the following compromises in our experiments to reduce computational complexity: (1) Only the 11 most frequent modifier pairs were chosen to generate Type II PCSs for examination in FMA. There are in fact 92 modifier pairs whose two members appeared in at least 50 distinct contexts and more if the threshold was set lower. Hence, it will require much more computational effort to retrieve all of the Type II PCSs from FMA. (2) For the 1:l ($l > 1$) patterns in the length models, we only focused on the 1:2 and the 1:3 patterns because the search space would expand exponentially with the increase of the number l. Although our results demonstrated that most of the instances follow the 1:2 pattern, to obtain exhaustive results for other patterns, big data approaches such as MapReduce cloud computing may be needed to improve algorithm efficiency [33,34].

7. Conclusions

Our methodology is innovative in three main aspects: Firstly, we defined parallel concepts and utilized them to design the general evaluation models, which provide a scalable approach easily adoptable by other semantic relationships for studying their granularity balance. Secondly, the Part-Of relationship in FMA was less frequently audited than the IS-A relationship in the literature due to its complex sub-hierarchies, while the comprehensive analysis of it performed in this study will make it easier to handle for future auditing tasks. Lastly, our study sheds light on an aspect of biomedical terminologies seldom studied: the granularity balance of semantic relationships. We not only discovered redundancies from this new aspect, but also presented an automatic mechanism to analyse the influence of them on imbalances and to categorize the final results into inconsistencies, missing concepts and misaligned relations.

Acknowledgements

This research is supported by the following grants: National Science Foundation of China (Nos. 61502221, 61402220), Scientific Research Fund of Hunan Provincial Education Department for Excellent Talents (Nos. 14B153, 16C1378), and the double first class construct program of USC (No. 2017SYL16).

References

[1] Y. Chen, X.F. Ren, G.Q. Zhang, R. Xu, Ontology-guided organ detection to retrieve web images of disease manifestation: towards the construction of a consumer-based health image library, J. Am. Med. Inf. Assoc. 20 (6) (2013) 1076–1081 (PMID: 23792805).

[2] L. Cui, S.S. Sahoo, S.D. Lhatoo, G. Garg, P. Rai, A. Bozorgi, G.Q. Zhang, Complex epilepsy phenotype extraction from narrative clinical discharge summaries, J. Biomed. Inf. (51) (2014) 272–279 (PMID: 24973735).

[3] S. Mate, F. Köpcke, D. Toddenroth, et al., Ontology-based data integration between clinical and research systems, PloS One 10 (1) (2015) e0116656 (PMID: 25588043).

[4] H. Min, Y. Perl, Y. Chen, M. Halper, J. Geller, Y. Wang, Auditing as part of the terminology design life cycle, J. Am. Med. Inf. Assoc. 13 (6) (2006) 676–690 (Nov-Dec, PMID: 16929044).

[5] O. Bodenreider, Quality Assurance in Biomedical Terminologies and Ontologies, Lister Hill National Center for Biomedical Communications, National Library of Medicine, Bethesda, 2010.

[6] O. Bodenreider, A. Burgun, T.C. Rindflesch, Assessing the consistency of a biomedical terminology through lexical knowledge, Int. J. Med. Inf. 67 (1–3) (2002) 85–95 (Dec 4, PMID: 12460634).

[7] H.H. Gu, D. Wei, J.L. Mejino Jr, G. Elhanan, Relationship auditing of the FMA ontology, J. Biomed. Inf. 42 (3) (2009) 550–557 (Jun, PMID: 19475727).

[8] L. Luo, J.L.V. Mejino, G.Q. Zhang, An analysis of FMA using structural self-bisimilarity, J. Biomed. Inf. 46 (3) (2013) 497–505 (PMID: 23557711).

[9] G.Q. Zhang, L. Luo, C. Ogbuji, C. Joslyn, J. Mejino, S.S. Sahoo, An analysis of multi-type relational interactions in fma using graph motifs with disjointness constraints, in: AMIA Annual Symposium Proceedings of 2012; 2012, pp. 1060–1069. PMID: 23304382.

[10] G.Q. Zhang, O. Bodenreider, Large-scale, exhaustive lattice-based structural auditing of SNOMED CT, in: AMIA Annual Symposium Proceedings of 2010, pp. 922–926. PMID: 21347113.

[11] J. Geller, H. Gu, Y. Perl, et al., Semantic refinement and error correction in large terminological knowledge bases, Data Knowl. Eng. 45 (1) (2003) 1–32.

[12] D. Wei, O. Bodenreider, Using the abstraction network in complement to description logics for quality assurance in biomedical terminologies - a case study in SNOMED CT, Stud. Health Technol. Inf. 160 (Pt 2) (2010) 1070–1074 (PMID: 20841848).

[13] Y. Wang, M. Halper, H. Min, Y. Perl, Y. Chen, K.A. Spackman, Structural methodologies for auditing SNOMED, J. Biomed. Inf. 40 (5) (2007) 561–581.

[14] X. Zhu, J.W. Fan, D.M. Baorto, C. Weng, J.J. Cimino, A review of auditing methods applied to the content of controlled biomedical terminologies, J. Biomed. Inf. 42 (3) (2009) 413–425, http://dx.doi.org/10.1016/j.jbi.2009.03.003 (PMID: 19285571).

[15] A. Rector, J. Rogers, T. Bittner, Granularity, scale and collectivity: when size does and does not matter, J. Biomed. Inf. 39 (3) (2006) 333–349 (PMID: 16515892).

[16] A. Kumar, B. Smith, D.D. Novotny, Biomedical informatics and granularity, Comp. Funct. Genom. 5 (6–7) (2004) 501–508.

[17] C. Weng, J.H. Gennari, D.B. Fridsma, User-centered semantic harmonization: a case study, J. Biomed. Inf. 40 (3) (2007) 353–364 (PMID: 17452021).

[18] R.L. Richesson, K.W. Fung, J.P. Krischer, Heterogeneous but standard coding systems for adverse events: issues in achieving interoperability between apples and oranges, Contemp. Clin. Trials. 29 (5) (2008) 635–645 (PMID: 18406213).

[19] P. Sun, S. Zhang, Identifying granularity differences between large biomedical ontologies through rules, in: AMIA Annual Symposium Proceedings of 2010, 2010, pp. 927–31 (PMID: 21347114).

[20] Z. He, J. Geller, G. Elhanan, Categorizing the relationships between structurally congruent concepts from pairs of terminologies for semantic harmonization, AMIA Sum. Transl. Sci. Proc. 2014 (2014) 48–53 (PMID: 25717400).

[21] Z. He, J. Geller, Y. Chen, A comparative analysis of the density of the SNOMED CT conceptual content for semantic harmonization, Artif. Intell. Med. 64 (1) (2015) 29–40 (PMID: 25890688).

[22] C. Rosse, J.L.V. Mejino, A reference ontology for biomedical informatics: the foundational model of anatomy, J. Biomed. Inf. 36 (2003) 478–500 (PMID: 14759820).

[23] K. Donnelly, SNOMED-CT: the advanced terminology and coding system for eHealth, Stud. Health Technol. Inf. 121 (2006) 279–290 (PMID: 17095826).

[24] SNOMED International Homepage < http://www.snomed.org > (last date accessed: August 16, 2017).

[25] G. Xing, G.Q. Zhang, L. Cui, FEDRR: fast, exhaustive detection of redundant hierarchical relations for quality improvement of large biomedical ontologies, BioData Min. 9 (2016) 31 (PMID: 27777627).

[26] Virtuoso < http://virtuoso.openlinksw.com > (last date accessed: August 16, 2017).

[27] https://www.w3.org/TR/sparql11-query/ (last date accessed: August 16, 2017).

[28] The Stanford Parser < https://nlp.stanford.edu/software/lex-parser.shtml > (last date accessed: August 16, 2017).

[29] A. Agrawal, Y. Perl, C. Ochs, G. Elhanan, Algorithmic detection of inconsistent modeling among SNOMED CT concepts by combining lexical and structural indicators, in: 2015 IEEE International Conference on Bioinformatics and Biomediccine (BIBM) 2015 Nov 9, pp. 476–483.

[30] H. Gu, G. Elhanan, M. Halper, Z. He, Questionable relationship triples in the UMLs, 2012 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), IEEE, New York, 2012, pp. 713–716.

[31] F. Mougin, N. Grabar, Auditing the multiply-related concepts within the UMLs, J. Am. Med. Inf. Assoc. 21 (e2) (2014) 185–193.

[32] F. Mougin, Identifying redundant and missing relations in the gene ontology, Stud. Health Technol. Inf. 210 (2015) 195–199 (PMID: 25991129).

[33] G.Q. Zhang, W. Zhu, M. Sun, S. Tao, O. Bodenreider, L. Cui, MaPLE: a MapReduce pipeline for lattice-based evaluation and its application to SNOMED CT, in: IEEE International Conference on Big Data 2014 (IEEE BigData 2014), pp. 754–759. PMID: 25705725.

[34] L. Cui, S. Tao, G.Q. Zhang, Biomedical ontology quality assurance using a big data approach, ACM Trans. Knowl. Discov. Data (TKDD) 10 (4) (2016) 41.