



# HHS Public Access

Author manuscript

*Proceedings (IEEE Int Conf Bioinformatics Biomed)*. Author manuscript; available in PMC  
2018 January 26.

Published in final edited form as:

*Proceedings (IEEE Int Conf Bioinformatics Biomed)*. 2017 November ; 2017: 1262–1269. doi:10.1109/  
BIBM.2017.8217840.

## Auditing the Assignments of Top-Level Semantic Types in the UMLS Semantic Network to UMLS Concepts

**Zhe He,**

School of Information, Florida State University, Tallahassee, FL, zhe.he@cci.fsu.edu

**Yehoshua Perl,**

Department of Computer Science, New Jersey Institute of Tehnology, Newark, NJ, perl@njit.edu

**Gai Elhanan,**

Department of Computer Science, New Jersey Institute of Tehnology, Newark, NJ,  
gelhanan@gmail.com

**Yan Chen,**

Department of Computer Information Systems, BMCC, CUNY, New York, NJ,  
ychen@bmcc.cuny.edu

**James Geller,** and

Department of Computer Science, New Jersey Institute of Tehnology, Newark, NJ,  
james.geller@njit.edu

**Jiang Bian**

Department of Health Outcomes and Policy, University of Florida, Gainesville, FL,  
bianjiang@ufl.edu

### Abstract

The Unified Medical Language System (UMLS) is an important terminological system. By the policy of its curators, each concept of the UMLS should be assigned the most specific Semantic Types (STs) in the UMLS Semantic Network (SN). Hence, the Semantic Types of most UMLS concepts are assigned at or near the bottom (leaves) of the UMLS Semantic Network. While most ST assignments are correct, some errors do occur. Therefore, Quality Assurance efforts of UMLS curators for ST assignments should concentrate on automatically detected sets of UMLS concepts with higher error rates than random sets. In this paper, we investigate the assignments of top-level semantic types in the UMLS semantic network to concepts, identify potential erroneous assignments, define four categories of errors, and thus provide assistance to curators of the UMLS to avoid these assignments errors. Human experts analyzed samples of concepts assigned 10 of the top-level semantic types and categorized the erroneous ST assignments into these four logical categories. Two thirds of the concepts assigned these 10 top-level semantic types are erroneous. Our results demonstrate that reviewing top-level semantic type assignments to concepts provides an effective way for UMLS quality assurance, comparing to reviewing a random selection of semantic type assignments.

## Keywords

Controlled Vocabulary; Semantic Network; Semantic Type; Quality Assurance

---

## I. Introduction

The Unified Medical Language System (UMLS) [1–5] designed by the National Library of Medicine (NLM), integrates more than 190 biomedical source terminologies. Its Metathesaurus (META) contains 3.4 million concepts in the 2016AB release<sup>1</sup>. The UMLS Semantic Network (SN) [6] is composed of 127 semantic types (STs) that provide a consistent semantic categorization of all the concepts in the META and are lined by important semantic relations. Each UMLS concept is then assigned one or more semantic types. Categorizing UMLS concepts using the semantic network has supported the integration of new UMLS sources over the years [7].

As one of the design principles for the UMLS semantic network, when assigning a semantic type to a concept, one should assign *a proper ST that is most specific* [7]. Thus, most UMLS concepts are assigned STs that are leaf nodes in the two SN trees (i.e., **Entity**<sup>2</sup> and **Event**), or are close to the bottoms of the trees. Nevertheless, some META concepts are assigned top-level STs, i.e., the two root nodes, **Entity** and **Event**, or their immediate children, e.g., **Conceptual Entity** and **Activity**. The following questions arise from these observations: 1) Which kinds of concepts are so general that they should be assigned top level STs? and 2) Are those assignments to top level STs justified in all cases? However, the rule that the most *specific* ST should be assigned to each concept was not strictly followed in some such assignments that could lead to inconsistency in the UMLS META. For example, the concept *Tropospheric Ozone* is assigned the ST **Chemical**, but should be assigned **Inorganic Chemical**, a child of **Chemical**.

In particular, we are concerned with the two root STs of the UMLS SN, **Entity** and **Event**, and their immediate children: **Physical Object** and **Conceptual Entity** which are the two children of **Entity**, and **Activity** and **Phenomenon or Process** which are the two children of **Event** (Fig. 1). We also investigate four more STs that are annotated in the UMLS documentation as “*few concepts are assigned this general type*”, which are **Chemical**, **Group**, **Anatomical Structure**, and **Biologic Function**. Each of these STs is a root of a subtree in the UMLS SN, and according to the above design principle of McCray and Hole [7], most relevant concepts will be assigned a descendant of such an ST rather than any of these four STs. Thus, the question that “which concepts are so general that they should be assigned top-level STs?” is of special interest.

The main purpose for the inclusion of these top-level STs in the UMLS SN was not for the categorization of META concepts, but for organizing the lower-level STs of the SN in a systematic way. Thus, the choice was made to structure the UMLS SN as two trees. The UMLS SN is similar in spirit to an upper-level ontology (Sowa [8]; Noy [9]), such as the

---

<sup>1</sup>[https://www.nlm.nih.gov/pubs/techbull/nd16/nd16\\_umls\\_2016ab\\_release.html](https://www.nlm.nih.gov/pubs/techbull/nd16/nd16_umls_2016ab_release.html)

<sup>2</sup>Concepts and STs are denoted in italics and bold typesets, respectively

Basic Formal Ontology (BFO) [10] and the BioTop ontology [11]. Sowa's *continuants* are organized in the **Entity** tree of the SN, while Sowa's *occurents* are located in the **Event** tree. Furthermore, the **Entity** tree is divided into physical objects under the **ST Physical Object** and abstract objects under the **ST Conceptual Entity**. This mirrors Sowa's distinction of categories into *physical* and *abstract*, which are further subcategorized as *continuants* and *occurents*. Such recursion is possible because Sowa does not limit himself to a tree structure.

In this paper, we investigate the prevalence of errors in the assignments of top-level STs. Further, we classify discovered errors into four categories, according to the nature of each error. This classification systematizes the process of auditing all concepts of the top-level STs. We hypothesize that assignments of top-level STs have a high concentration of errors. We note that auditing of ST assignments is not done just for its own sake. Rather, wrong ST assignments often indicate misconceptions about the true nature of a concept, which are typically indicative of other structural modeling errors at or near a misclassified concept [12]. For a state-of-the-art review of auditing methods for terminologies in general and for the UMLS in particular see [13].

## II. Background

### A. Applications that Use the UMLS Semantic Types

The UMLS semantic types have been widely used in information extraction [14], clinical annotation [15], ontology learning [16], and knowledge representation [17]. Albright et al. created annotated clinical narratives with syntactic and semantic labels using UMLS semantic types [15]. Zhang et al. used UMLS semantic types to extract new types of information in clinical notes including problems, medications, and laboratory information [14]. Weng et al. developed a semantic representation for clinical trial eligibility criteria using UMLS semantic types and concepts to support the electronic patient eligibility determination [17]. It is natural that these applications will benefit from more accurate semantic type assignments to UMLS concepts, which is the goal of this work.

### B. Quality Assurance of the UMLS Semantic Types

The Quality Assurance (QA) process of the UMLS is different from the QA of a specific source terminology. This difference stems from the UMLS being not a terminology but a compendium of over 190 terminologies' knowledge organized in a unified system. These UMLS source terminologies sometimes contradict one another. Hence, the UMLS will, by definition, contain contradictions, due to the UMLS commitment to maintain the knowledge from each source terminology as is. As a result, QA of the UMLS cannot resolve some contradictions, even though elimination of contradictions is a fundamental task of QA of a terminology, e.g. finding prohibited IS-A cycles in a hierarchy [18, 19].

However, there exist *some* resolvable contradictions, since they are the result of the integration process of the source terminologies by the UMLS team. An example is an IS-A cycle of three concepts that is the result of UMLS editors mistakenly making two concepts of different meaning from different sources the same concept in the UMLS, while they

should be created as two distinct UMLS concepts (see [20]). Separating such a UMLS concept into two different concepts, each conforming to the meaning in its source terminology, breaks the erroneous IS-A cycle.

Therefore, QA of the UMLS should concentrate on errors that can be corrected in the UMLS framework. An ST assignment of a concept is a UMLS artifact, intended to capture the semantics of this concept [7]. Hence, correcting erroneous ST assignments is under the control of the UMLS editors.

### C. Top-Level Semantic Type Assignments in the UMLS

In an effort to identify general patterns of assigning STs to concepts, we will now review examples of concepts that are assigned high level STs and will analyze possible motivations for these assignments. By definition, the *extent* of an ST is the set of UMLS concepts assigned that ST. One example of a high-level ST with a very large extent (93,627 concepts) is **Disease or Syndrome**. This ST has two children, **Mental or Behavioral Dysfunction** and **Neoplastic Process**. Each of these children categorizes a specific family of diseases. However, the coverage of diseases by these two STs is not exhaustive. For example, the concept *type 2 diabetes* should not be assigned one of the two children, but **Disease or Syndrome** itself.

A different situation exists regarding **Anatomical Abnormality**, which has two children, **Acquired Abnormality** and **Congenital Abnormality**. These two children constitute an exhaustive sub-categorization of **Anatomical Abnormality**. That is, any concept that represents an anatomical abnormality must either be assigned **Acquired Abnormality** or **Congenital Abnormality**, describing the origin of the abnormality. These two STs define mutually exclusive categories, where no concept may be assigned both of them simultaneously. The question is how to categorize an abnormality that may be either acquired or congenital (but not both). According to the UMLS usage notes of **Anatomical Abnormality**, users should “use this type if the abnormality in question can be either an acquired or congenital abnormality.” In other words, the way the NLM handles such a concept is to ascend from these two leaves to their “lowest common ancestor” (LCA). For **Acquired Abnormality** and **Congenital Abnormality**, their parent **Anatomical Abnormality** is used.

Each concept of the UMLS is assigned one or more of the 127 semantic types of the Semantic Network. While most ST assignments are correct, it is important to expose the concepts with erroneous ST assignment, since those are often indicators of modeling errors for these concepts (see e.g. [12, 21, 22]). However, a broad QA effort for detecting such errors is very expensive and will yield relatively few errors. The challenge is to design automated techniques that can identify the sets of concepts with a high likelihood of erroneous ST assignments. These sets are then presented to UMLS curators for QA. The current work presents such a technique.

### III. Method

The categorization principles reviewed above will guide our analysis of the UMLS assignments of top-level STs. Some UMLS concepts are very general, which naturally should be assigned top-level STs. For example, the concepts *Entity* and *Observable entity* are both correctly assigned the ST **Entity**. The concepts *Event* and *Reportable Event* are assigned the ST **Event**. The concept *Physical Object* is assigned the ST **Physical Object**.

As discussed above, another reason why a concept may be legitimately assigned a top-level ST is because this ST is the lowest common ancestor of two other STs,  $ST_1$  and  $ST_2$ , and the concept either has the semantics of  $ST_1$  or the semantics of  $ST_2$ , but not both. That is, the concept can be described by the disjunction of these two STs (the OR logical operator). For example, the concept *Products or substances for personal consumption* consists of two separate subterms, combined by a disjunction. Each of these subterms should be a separate concept in reality; however, this would require a change to the source terminology, which violates the UMLS principle to maintain the knowledge from each source terminology as is. The first putative concept, *Product*, should be assigned the ST **Manufactured Object**. The second, *substances*, should be assigned **Substance**. But this is not a concept that is both a product and a substance (conjunction; the logical AND operation). Either of the two STs above would only cover one subterm, but omit the other one. The solution used by the UMLS curators for such a disjunctive concept is to find the lowest common ancestor  $ST_3$  of  $ST_1$  and  $ST_2$  and assign  $ST_3$  to such a concept. For this example, the lowest common ancestor of **Manufactured Object** and **Substance** is their parent, the top-level ST **Physical Object**.

We note that when a concept has the semantics of both  $ST_1$  and  $ST_2$ , both semantic types are assigned to that concept. Such a concept is thus denoted as a conjunctive concept. These cases are common in the UMLS, especially for chemicals. Our previous research on combinations of STs is embodied in the Refined Semantic Network [23–32]. Further, for a study of disjunctive and conjunctive concepts in terminologies see work by Mendonca et al. [33].

We will now present our categorization for erroneous assignments of top-level STs to concepts. The first kind of error is that the concept is more specific than what is expressed by the ST assignment. Such a concept should be assigned a more specific ST, which is typically a descendant of the assigned ST. For example, the concepts *Gifts*, *Financial*, which is currently assigned **Entity**, should be assigned its child **Conceptual Entity** (Fig.1) by the nature of “gifts,” according to its definition, “A gift is the transfer of something without the expectation of receiving something in return” (Wikipedia). We call this kind of error a “more specific ST classification needed” error (abbreviated SPC).

To help with detecting such concepts, one can try to locate similar concepts that are indeed assigned the proper more specific ST, under the assumption that modeling of concepts in a terminology should be consistent. Hence, similar concepts should typically be assigned the same ST. By reviewing the ST(s) assigned to similar concepts, one can gain support for declaring an ST assignment an error. This comparison assumes, of course, that the

assignments of the similar concepts are correct. For example, *Unknown Terms* is assigned **Entity**. But a similar concept, *Term (lexical)*, is assigned **Idea or Concept**, which is a grandchild of **Entity** (Fig. 1). Thus, the former ST, **Idea or Concept**, should also be assigned to *Unknown Terms*.

The second kind of error is called “lowest common ancestor” error, coded as LCA. At the end of the Background section, we already discussed the need for using LCAs as one reason why concepts may be validly assigned a top-level ST. This was illustrated by *Products or substances for personal consumption*, which should be assigned **Physical Object** as the LCA (in this case, parent) of the STs **Manufactured Object** and **Substance**. However, in the UMLS, this concept is not assigned **Physical Object**, but **Entity**, the parent of **Physical Object**, which is a *common ancestor*, but not the *lowest common ancestor* of these two STs (Fig.1).

The third kind of error can be identified by a contradictory configuration of concepts and semantic types that Geller et al. have referred to as semantic inversion (SI) [34]. In this kind of error, one concept is assigned a top-level ST, ST<sub>1</sub>, while its parent concept is assigned a descendant ST<sub>2</sub> of ST<sub>1</sub>. This configuration of two concepts connected by a parent relationship is considered contradictory, since their assigned STs are connected in the inverse order by the hierarchical IS-A relationship in the Semantic Network. Normally, it is expected that a (more general) concept is assigned the same ST or a more general ST than its child concept.

To illustrate this kind of error, let us consider the concept *Triangular* assigned **Conceptual Entity** and its parent *Shapes* which is assigned **Spatial Entity**, a grandchild of **Conceptual Entity** (Fig. 2). This example presents a case of semantic inversion (coded as SI). This contradictory configuration can be resolved by changing the assignment of *Triangular* to **Spatial Concept**. This configuration has been well investigated by Geller et al. [34] and Bodenreider et al. [18].

The fourth kind of error involves a ST assignment that is “in the wrong subtree of the Semantic Network.” To illustrate this kind of error, let us consider the assignment of **Conceptual Entity** to the concept *Contract dispute*. This assignment is more general than the assignment of *Contract agreement*, assigned **Intellectual Product**. However, a *Contract dispute* is a dispute about a contract, and *dispute* is assigned **Social Behavior**. Hence *Contract dispute* should also be assigned **Social Behavior**. **Intellectual Product** is in the **Entity tree**, while **Social Behavior** is in the **Event tree** of the SN. We use the abbreviation MC (miscategorization) for these cases.

Based on these four categories, we performed an audit of ST assignments for the six top-level STs located in the top two level of the SN and four other STs with the usage note “few concepts will be assigned this broad type.” As noted above, most top level-STs have small extents, and we audited all concepts in those extents. For the larger extents, such as those of **Conceptual Entity**, **Activity**, **Phenomenon or Process**, and **Biologic Function**, we audited a random sample of 50 concepts. The audit was conducted by the authors ZH, YC and GE. All the authors are experts in terminologies and have conducted many previous audits. For

**Chemical, Anatomical Structure, and Biologic Function**, co-authors GE, who is a medical doctor, and YC, who was trained in sports medicine, performed the audit. The inter-rater reliability between the two auditors is high. Conflicts were resolved after discussion. The other seven STs analyzed in this study are not medical categories, thus no medical knowledge was required for the audit. The samples of these seven STs were first reviewed by ZH, and then ZH's determinations were reviewed by one of the other two auditors (GE and YC), who are more experienced in terminology auditing. Due to significant seniority level between ZH and the other two auditors, the final auditing results are from the two senior auditors. The inter-rater reliability for these seven STs was not computed. The auditors used the Neighborhood Auditing Tool (NAT) [35]. The NAT, a powerful tool for auditing the UMLS, displays the STs of the audited focus concept and its neighboring concepts (e.g., parents, children, siblings), as well as other assorted items of relevant information (e.g., definition, lateral relationships, etc.) about the focus concept. To illustrate how the neighboring concepts help to confirm an ST assignment error, let us consider the concept *Civilization*, which is assigned **Activity**. Its parent *Anthropological Culture* and its children *Arab World* and *Western World* are all assigned **Idea or Concept** which should be assigned to *Civilizations* as well. Note that the search function of NAT allows an auditor to view similar concepts based on the search term. For example, when the auditor search for the term *Term*, its similar terms such as *Unknown terms* can be also retrieved. When the auditors reviewed the samples, they were also asked to report new possible categories of ST assignment errors. No new categories were reported.

#### IV. Results

We performed an audit of top-level ST assignments in the 2013AA release of the UMLS and recently reviewed if the ST assignment errors remained in the 2016AB release. Aggregate information for each top-level ST and each kind of error is shown in Table I. The percentages of concepts with erroneous ST assignments are high, ranging from 27.7% for **Physical Object** to 70.4% for **Chemical** and 74% for **Conceptual Entity**. To validate our hypothesis that top-level STs have more errors than non-top-level STs. We also audited a control sample of 50 randomly chosen concepts that are assigned non-top-level STs. No erroneous ST assignments were found in the control sample. This supports the claim that assignments of top level STs have a high concentration of errors.

The average percentage of errors for the entire sample, 52.1%, is very high. Table I also shows the total numbers of occurrences for each ST and the percentages for each of the four different kinds of errors.

The majority of the errors (45.4%) are cases where the correct assignment should be more specific (SPC), i.e., to a descendant of the assigned top-level ST. In 24.4% of the cases, the top-level ST assignment is a miscategorization (MC), i.e., it cannot be repaired by changing the assignment to a descendant ST. Only 4 out of 328 erroneous cases were deemed to be assigned the lowest common ancestor (LCA) of the semantic types of the two components of the concept. The suggested resolution of this kind of error is similar to SPC or MC, i.e., assigning a more specific semantic type in the same sub-hierarchy, or a different semantic type in a different sub-hierarchy. Table II shows samples for each kind of error. The numbers

of samples shown in Table II for SPC, SI, MC, and LCA are seven, four, three, and two, respectively, based on the percentage of different kinds of errors found.

We have submitted the auditing report of this study to the UMLS editors through the customer service website of National Library of Medicine<sup>3</sup>. Due to the fact that NLM does not have a mechanism to give feedback, we only checked the new release of the UMLS to see if the suggested changes have been implemented. However, most of the errors have not been corrected so far (in 2016AB release of the UMLS). Among all the concepts in the auditing report, ST assignments of 5.0% of the concepts were corrected based on our suggestions; ST assignments of 4.3% of the concepts were changed; 1.0% of the concepts were deleted. It is our hope that some or all of the errors will be corrected in a future release of the UMLS.

## V. Discussion

There is a desire among terminology editors to automate as much as possible the QA of terminologies or terminological systems such as the UMLS. However, the detection of terminological errors and inconsistencies requires sophisticated analysis and deep knowledge of both the domain and terminologies. It is a challenge to automate such a process, except for cases when well-defined rules are violated. As examples of such errors and their QA see studies of “an algorithm for detecting redundant ST assignment for UMLS concepts” [36] and “redundant ST assignments for organic chemicals” [32]. The error of Semantic Inversion [34] in this paper, also falls into this category. Semantic inversion is considered contradictory. However, that is inevitable due to the nature of the UMLS. If source terminologies mix-up super- and subtypes, then this will be reflected in the UMLS, and a subtype may have a more generic semantic type than the supertype. The UMLS editors can algorithmically detect all the semantic inversion cases and correct them. However, for most errors it is not clear how to fully automate the QA process.

Therefore, we improve the QA process by automating the identification of concepts with high likelihoods of errors. The review of these concepts will still be done manually by editors who are domain experts and familiar with terminologies. The automation of the detection of candidates for review can be done algorithmically if the characterization of such sets of concepts follows clear, objective rules. The advantage of reviewing only such concepts is the high yield of errors discovered relative to the effort invested and relative to the number of concepts actually reviewed by a human expert. An example of such a method is reviewing concepts assigned multiple STs, with the special provision that the given combination of STs is assigned only to a few other concepts [25, 27, 37]. The method used in the current research, focusing on concepts assigned top-level STs, also falls into this category.

The large majority of errors found in this study are of the kind where a concept should be assigned a descendant of the currently assigned ST (SPC). Such an error shows a granularity misconception of the editors, either about the concept or about the nature of the ST. Errors

---

<sup>3</sup><http://apps.nlm.nih.gov/mainweb/siebel/nlm/index.cfm>



involving lowest common ancestors (LCA) are rare. Note that for this error kind, correcting also involves replacing the assigned ancestor ST by a more specific one, namely the lowest common ancestor.

Furthermore, for most of the 10 STs investigated in this study, a usage note by the designers of the UMLS SN [38] states “*a few concepts are assigned this broad type.*” Nevertheless, six of the 10 STs are assigned to more than 100 concepts, with error rates between 40% and 74%. The impression arises that one or a few editors of the UMLS were not aware of the vision of the creators of the Semantic Network, resulting in a high proportion of errors.

For three kinds of errors, all except miscategorization (MC), we speculate that there might be a tendency of some UMLS editors to “err up.” By this we mean that when an editor is choosing between two options, an ST and its ancestor ST, the tendency seems to be to “err on the safe side” by picking the more general ST. By assigning a too general ST, the editor does not pick a “totally wrong” categorization, just maybe one that is unnecessarily general. For example, when a concept should be assigned **Natural Phenomenon or Process**, then an assignment of **Human-caused Phenomenon or Process** will be totally wrong, since those two sibling STs are exclusive. On the other hand, an assignment of the parent **Phenomenon or Process** is too general, but not entirely wrong. While this tendency is psychologically understandable, it stands in contrast to the basic requirement of the UMLS SN [7] of assigning the most specific ST possible. In Table II, one can see the impact of this tendency, whether real or imagined, for all kinds of errors, except for miscategorization. According to Table I, about 75% of the observed errors could be explained by such a behavior. We note that some of the erroneous ST assignments of the fourth kind (MC) seem to occur due to categorizations done based on natural language processing of the term, ignoring the concept’s parents, children and neighbors. This phenomenon is demonstrated in the example of *civilization* at the end of the Method section.

Some limitations should be noted in this work. First, seven non-chemical STs were reviewed by a single senior auditor GE or YC (and a junior auditor ZH). As the audit was guided by the predefined categories of errors and conducted in the NAT tool, it is not likely that there would be many conflicting decisions. Second, the impact of the erroneous top-level ST assignments on other STs was not measured. In future work, we will analyze ST assignments along the hierarchy of the concepts assigned top-level ST, which may indicate other structural modeling errors at or near a misclassified concept.

## VI. Conclusions

We investigated the semantic type assignments of ten top-level semantic types of the UMLS Semantic Network. Many erroneous semantic type assignments were reported. Four kinds of errors were identified and described. In most cases, these errors can be easily corrected by reassigning a more specific semantic type to the concept in question. Within the given limited sample, a tendency of UMLS editors to “err up” by assigning a concept a semantic type that is an ancestor of the proper semantic type was noticed.

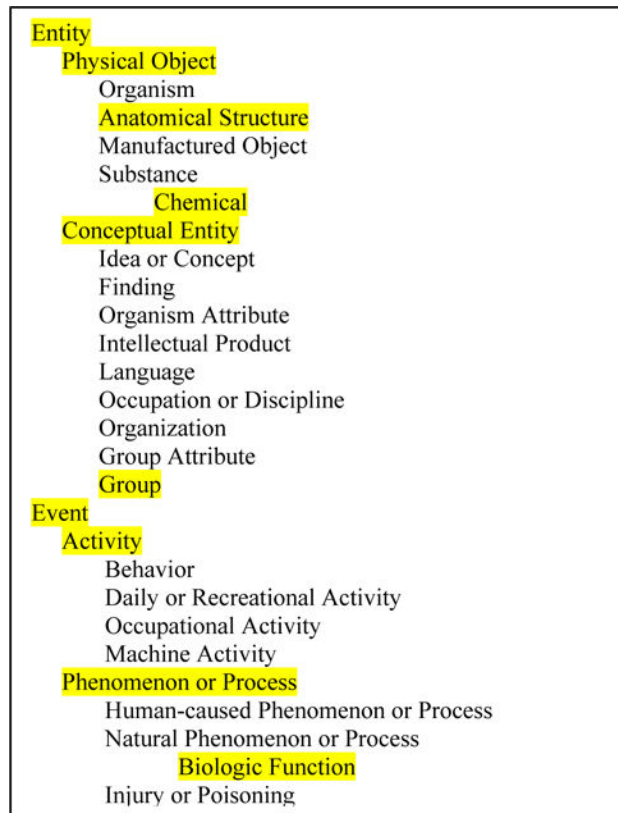
## Acknowledgments

This work was partially supported by the United States National Library of Medicine of National Institutes of Health under Award Number R01LM008445-01A2. The content is solely the responsibilities of the authors and does not necessarily represents the official views of the National Institutes and Health. This work was partially supported by the Leir Charitable Foundations through the School of Management at NJIT.

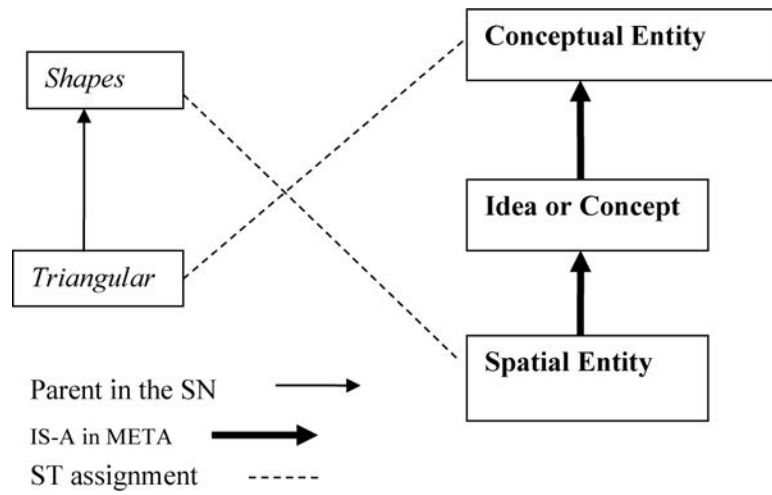
## References

1. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* Jan 1.2004 32:D267–70. [PubMed: 14681409]
2. Humphreys BL, Lindberg DA, Schoolman HM, Barnett GO. The Unified Medical Language System: an informatics research collaboration. *J Am Med Inform Assoc.* Jan-Feb;1998 5:1–11. [PubMed: 9452981]
3. Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. *Methods Inf Med.* Aug.1993 32:281–91. [PubMed: 8412823]
4. Humphreys BL, Lindberg DA. The UMLS project: making the conceptual connection between users and the information they need. *Bull Med Libr Assoc.* Apr.1993 81:170–7. [PubMed: 8472002]
5. Humphreys BL, Lindberg DA, Hole WT. Assessing and enhancing the value of the UMLS Knowledge Sources. *Proc Annu Symp Comput Appl Med Care.* 1991:78–82. [PubMed: 1807711]
6. McCray AT, Nelson SJ. The representation of meaning in the UMLS. *Methods Inf Med.* Mar.1995 34:193–201. [PubMed: 9082131]
7. McCray, AT., Hole, WT. The scope and structure of the first version of the UMLS Semantic Network. presented at the Proc 14th Annu Symp Comput Appl Med Care; Los Alamitos, CA. 1990.
8. Sowa, JF. Distinction, Combinations and Constraints. presented at the Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing; Montreal, Canada. 1995.
9. Noy NF, Hafner CD. The state of the art in ontology design: A survey and comparative review. *AI Magazine.* 1997:53–74.
10. Grenon P, Smith B, Goldberg L. Biodynamic ontology: applying BFO in the biomedical domain. *Stud Health Technol Inform.* 2004; 102:20–38. [PubMed: 15853262]
11. Stenzhorn H, Beisswanger E, Schulz S. Towards a top-domain ontology for linking biomedical ontologies. *Stud Health Technol Inform.* 2007; 129:1225–9. [PubMed: 17911910]
12. Chen Y, Gu HH, Perl Y, Geller J. Structural group-based auditing of missing hierarchical relationships in UMLS. *J Biomed Inform.* Jun.2009 42:452–67. [PubMed: 18824248]
13. Zhu X, Fan JW, Baorto DM, Weng C, Cimino JJ. A review of auditing methods applied to the content of controlled biomedical terminologies. *J Biomed Inform.* Jun.2009 42:413–25. [PubMed: 19285571]
14. Zhang R, Pakhomov S, Melton GB. Longitudinal analysis of new information types in clinical notes. *AMIA Jt Summits Transl Sci Proc.* 2014; 2014:232–7. [PubMed: 25717418]
15. Albright D, Lanfranchi A, et al. Towards comprehensive syntactic and semantic annotations of the clinical narrative. *J Am Med Inform Assoc.* Sep-Oct;2013 20:922–30. [PubMed: 23355458]
16. Hoxha J, Jiang G, Weng C. Automated learning of domain taxonomies from text using background knowledge. *J Biomed Inform.* Oct.2016 63:295–306. [PubMed: 27597572]
17. Weng C, Wu X, Luo Z, et al. EliXR: an approach to eligibility criteria extraction and representation. *J Am Med Inform Assoc.* Dec; 2011 18(Suppl 1):i116–24. [PubMed: 21807647]
18. Bodenreider O. Circular hierarchical relationships in the UMLS: etiology, diagnosis, treatment, complications and prevention. *Proc AMIA Symp.* 2001:57–61. [PubMed: 11825155]
19. Mougín F, Bodenreider O. Approaches to eliminating cycles in the UMLS Metathesaurus: naive vs. formal. *AMIA Annu Symp Proc.* 2005:550–4. [PubMed: 16779100]
20. Halper M, Morrey CP, Chen Y, Elhanan G, Hripscak G, Perl Y. Auditing hierarchical cycles to locate other inconsistencies in the UMLS. *AMIA Annu Symp Proc.* 2011; 2011:529–36. [PubMed: 22195107]

21. Gu HH, Hripcsak G, Chen Y, Morrey CP, Elhanan G, Cimino J, et al. Evaluation of a UMLS Auditing Process of Semantic Type Assignments. *AMIA Annu Symp Proc.* 2007;294–8. [PubMed: 18693845]
22. Gu HH, Elhanan G, Perl Y, Hripcsak G, Cimino JJ, Xu J, et al. A study of terminology auditors' performance for UMLS semantic type assignments. *J Biomed Inform.* Dec.2012 45:1042–8. [PubMed: 22687822]
23. Gu HH, Perl Y, Geller J, Halper M, Liu LM, Cimino JJ. Representing the UMLS as an object-oriented database: modeling issues and advantages. *J Am Med Inform Assoc.* 2000; 7:66–80. [PubMed: 10641964]
24. Geller J, Gu HH, Perl Y, Halper M. Semantic refinement and error correction in large terminological knowledge bases. *Data and Knowledge Engineering.* 2003; 45:1–32.
25. He Z, Morrey CP, Perl Y, Elhanan G, Chen L, Chen Y, et al. Sculpting the UMLS Refined Semantic Network. *Online J Public Health Inform.* 2014; 6:e181. [PubMed: 25422719]
26. Chen L, Morrey CP, Gu HH, et al. Modeling multi-typed structurally viewed chemicals with the UMLS Refined Semantic Network. *J Am Med Inform Assoc.* Jan-Feb;2009 16:116–31. [PubMed: 18952946]
27. Gu HH, Perl Y, Elhanan G, Min H, Zhang L, Peng Y. Auditing concept categorizations in the UMLS. *Artif Intell Med.* May.2004 31:29–44. [PubMed: 15182845]
28. Chen Y, Gu HH, Perl Y, Halper M, Xu J. Expanding the extent of a UMLS semantic type via group neighborhood auditing. *J Am Med Inform Assoc.* Sep-Oct;2009 16:746–57. [PubMed: 19567802]
29. Morrey CP, Perl Y, Halper M, Chen L, Gu HH. A chemical specialty semantic network for the unified medical language system. *J Cheminform.* 2012; 4:9. [PubMed: 22577759]
30. Chen Y, Gu H, Perl Y, Halper M, Xu J. Expanding the extent of a UMLS semantic type via group neighborhood auditing. *J Am Med Inform Assoc.* Sep-Oct;2009 16:746–57. [PubMed: 19567802]
31. Chen Y, Gu H, et al. Overcoming an obstacle in expanding a UMLS semantic type extent. *J Biomed Inform.* Feb.2012 45:61–70. [PubMed: 21925287]
32. Morrey CP, Chen L, Halper M, Perl Y. Resolution of redundant semantic type assignments for organic chemicals in the UMLS. *Artif Intell Med.* Jul.2011 52:141–51. [PubMed: 21646001]
33. Mendonca EA, Cimino JJ, Campbell KE, Spackman KA. Reproducibility of interpreting “and” and “or” in terminology systems. *Proc AMIA Symp.* 1998:790–4. [PubMed: 9929327]
34. Geller J, Morrey CP, Xu J, Halper M, Elhanan G, Perl Y, et al. Comparing inconsistent relationship configurations indicating UMLS errors. *AMIA Annu Symp Proc.* 2009; 2009:193–7. [PubMed: 20351848]
35. Morrey CP, Geller J, Halper M, Perl Y. The Neighborhood Auditing Tool: a hybrid interface for auditing the UMLS. *J Biomed Inform.* Jun.2009 42:468–89. [PubMed: 19475725]
36. Peng Y, Halper MH, Perl Y, Geller J. Auditing the UMLS for redundant classifications. *Proc AMIA Symp.* 2002:612–6. [PubMed: 12463896]
37. Geller J, He Z, Perl Y, Morrey CP, Xu J. Rule-based support system for multiple UMLS semantic type assignments. *J Biomed Inform.* Feb.2013 46:97–110. [PubMed: 23041716]
38. (Dec 5). *The UMLS Semantic Network.* Available: <https://semanticnetwork.nlm.nih.gov/>



**Fig. 1.** The four highest levels of the Semantic Network, with the ten STs analyzed in our study highlighted in yellow. Six highlighted STs are in the first two levels and four more are in the third level. The latter are annotated with “few concepts are assigned this broad type” in the UMLS documentation.



**Fig. 2.** Example of Semantic Inversion: The more specific concept *Triangular* should not be assigned a semantic type that is more general than the semantic type of its parent *Shapes*.

Table 1

Erroneous ST Assignments for Top-Level STs

ST	Extent	# of concepts	SPC	LCA	SI	MC	# of errors	% of errors (95% CI)
Conceptual Entity	676	50	22	0	7	8	37	74 (61.8 – 86.2)
Chemical	27	27	12	0	2	5	19	70.4 (53.2 – 87.6)
Group	53	53	11	0	0	24	35	66 (53.2 – 78.8)
Anatomical Structure	127	127	13	0	49	7	69	54.3 (45.6 – 63.0)
Biologic Function	1398	50	1	0	17	9	27	54 (40.2 – 67.8)
Event	151	151	72	1	0	2	75	49.7 (41.7 – 57.7)
Activity	366	50	6	0	1	16	23	46 (32.2 – 59.8)
Entity	24	24	5	3	2	0	10	41.7 (22.0 – 61.4)
Phenomenon or Process	1596	50	3	0	16	1	20	40 (26.4 – 53.6)
Physical Object	47	47	4	0	1	8	13	27.7 (14.3 – 39.9)
Total	4465	629	149	4	95	80	328	52.1 (48.1 – 55.9)
% of all errors	–	–	45.4%	1.2%	29%	24.4%	–	–

ST: Semantic type

SPC: more specific ST classification needed

LCA: Lowest common ancestor

SI: semantic inversion

MC: Miscategorization

CI: Confidence interval

Table II

Samples of Erroneous Semantic Type Assignments of Different Kinds

Kind	Concept	ST	Suggested ST	Similar Concept	Comments
SPC	<i>Patient-related Identification code</i>	Entity	Intellectual Product	<i>Code (Intellectual Product)</i>	<b>Intellectual Product</b> is more specific than <b>Entity</b>
SPC	<i>NCI Director's Consumer Liaison Group</i>	Group	Professional Organization or Group	<i>Professional Organization or Group (Professional Organization or Group)</i>	<b>Professional Organization or Group</b> is more specific than <b>Group</b>
SPC	<i>Anterior Visual Pathway</i>	Anatomical Structure	Body Part, Organ or Organ Component	<i>Entire visual pathway (Body Part, Organ or Organ Component)</i>	<b>Body Part, Organ or Organ Component</b> is more specific than <b>Anatomic Structure</b>
SPC	<i>Tropospheric Ozone</i>	Chemical	Inorganic Chemical	<i>Oxygen Compounds, Unspecified (Inorganic Chemical)</i>	Ozone is allotrope of oxygen that is much less stable than the diatomic allotrope O <sub>2</sub> . <b>Inorganic Chemical</b> is more specific than <b>Chemical</b>
SPC	<i>Responsible observer</i>	Group	Professional or Occupational Group	<i>Observer (Professional or Occupational Group)</i>	<b>Professional or Occupational Group</b> is more specific than <b>Group</b>
SPC	<i>NCI Director's Consumer Liaison Group</i>	Group	Professional or Occupational Group	<i>NCI Advisory Board or Group (Professional or Occupational Group)</i>	<b>Professional or Occupational Group</b> is more specific than <b>Group</b> . <i>NCI Advisory Board or Group</i> is a grandparent of <i>NCI Director's Consumer Liaison Group</i> , therefore it is also an error of ST.
SPC	<i>negative regulation of hair-follicle maturation</i>	Biologic Function	Organism Function	<i>negative regulation of hair cycle (Organism Function)</i>	<b>Organism Function</b> is more specific than <b>Biologic Function</b>
LCA	<i>Products or substances for personal consumption</i>	Entity	Physical Object	<i>Substance (Substance) nonfood product (Manufactured Object)</i>	<b>Physical Object</b> is the LCA of <b>Substance</b> and <b>Manufactured Object</b>
LCA	<i>Radiofrequency interference</i>	Event	Phenomenon or Process	<i>Viral Interference (Natural Phenomenon or Process)</i>	<b>Phenomenon or Process</b> is the LCA of <b>Human-caused Phenomenon or Process</b> and <b>Natural Phenomenon or Process</b>
SI	<i>Reporter Position</i>	Conceptual Entity	Spatial Concept	<i>Location (Spatial Concept)</i>	Parent <i>Location</i> is assigned <b>Spatial Concept</b> , which is child of <b>Conceptual Entity</b>
SI	<i>positive regulation of response to external stimulus</i>	Biologic Function	Physiologic Function	<i>Regulation of response to external stimulus (Physiologic Function)</i>	Parent <i>Regulation of response to external stimulus</i> is assigned <b>Physiologic Function</b> , which is child of <b>Biologic Function</b>
SI	<i>Upper Gastrointestinal Tract</i>	Anatomical Structure	Body Part, Organ or Organ Component	<i>Gastrointestinal System Part (Body Part, Organ or Organ Component)</i>	Parent <i>Gastrointestinal System Part</i> is assigned <b>Body Part, Organ or Organ Component</b> , which is child of <b>Anatomical Structure</b>
SI	<i>positive regulation of conidiophore stalk development</i>	Biologic Function	Cell Function	<i>regulation of conidiophore stalk development (Cell Function)</i>	Parent <i>regulation of conidiophore stalk development</i> is assigned <b>Cell Function</b> , which is child of <b>Biologic Function</b>

Kind	Concept	ST	Suggested ST	Similar Concept	Comments
MC	<i>Nuclear Warfare</i>	Activity	Human-caused Phenomenon or Process	<i>Civil war (Human-caused Phenomenon or Process)</i>	Human-caused Phenomenon or Process is in Phenomenon or Process hierarchy
MC	<i>Civilization</i>	Activity	Idea or Concept	<i>Western world (Idea or Concept)</i>	Idea or Concept is in Entity hierarchy
MC	<i>Funding Applicant</i>	Group	Human	<i>Applicant (person) (Human)</i>	Human is in Physical Object hierarchy

**SPC:** more specific ST classification needed

**LCA:** lowest common ancestor

**SI:** semantic inversion

**MC:** miscategorization



Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

---

<sup>4</sup>From <http://en.wikipedia.org/wiki/Ozone>