

Sculpting the UMLS Refined Semantic Network

Zhe He¹, C. Paul Morrey², Yehoshua Perl³, Gai Elhanan⁴, Ling Chen⁵, Yan Chen⁵, James Geller³

¹Department of Biomedical Informatics, Columbia University, New York, NY, USA

²Department of Information Systems and Technology, Utah Valley University, Orem, UT, USA

³Department of Computer Science, New Jersey Institute of Technology, Newark, NJ, USA

⁴Halfpenny Technologies, Inc., Blue Bell, PA

⁵BMCC, City University of New York, New York, NY

Abstract

Background: The Refined Semantic Network (RSN) for the UMLS was previously introduced to complement the UMLS Semantic Network (SN). The RSN partitions the UMLS Metathesaurus (META) into disjoint groups of concepts. Each such group is semantically uniform. However, the RSN was initially an order of magnitude larger than the SN, which is undesirable since to be useful, a semantic network should be compact. Most semantic types in the RSN represent combinations of semantic types in the UMLS SN. Such a “combination semantic type” is called Intersection Semantic Type (IST). Many ISTs are assigned to very few concepts. Moreover, when reviewing those concepts, many semantic type assignment inconsistencies were found. After correcting those inconsistencies many ISTs, among them some that contradicted UMLS rules, disappeared, which made the RSN smaller.

Objective: The authors performed a longitudinal study with the goal of reducing the size of the RSN to become compact. This goal was achieved by correcting inconsistencies and errors in the IST assignments in the UMLS, which additionally helped identify and correct ambiguities, inconsistencies, and errors in source terminologies widely used in the realm of public health.

Methods: In this paper, we discuss the process and steps employed in this longitudinal study and the intermediate results for different stages. The sculpting process includes removing redundant semantic type assignments, expanding semantic type assignments, and removing illegitimate ISTs by auditing ISTs of small extents. However, the emphasis of this paper is not on the auditing methodologies employed during the process, since they were introduced in earlier publications, but on the strategy of employing them in order to transform the RSN into a compact network. For this paper we also performed a comprehensive audit of 168 “small ISTs” in the 2013AA version of the UMLS to finalize the longitudinal study.

Results: Over the years it was found that the editors of the UMLS introduced some new inconsistencies that resulted in the reintroduction of unwarranted ISTs that had already been eliminated as a result of their previous corrections. Because of that, the transformation of the RSN into a compact network covering all necessary categories for the UMLS was slowed down. The corrections suggested by an audit of the 2013AA version of the UMLS achieve a compact RSN of equal magnitude as the UMLS SN. The number of ISTs has been reduced to 336. We also demonstrate how auditing the semantic type assignments of UMLS concepts can expose other modeling errors in the UMLS source terminologies, e.g., SNOMED CT, LOINC, and RxNORM that are important for health informatics. Such errors would otherwise stay hidden.

Conclusions: It is hoped that the UMLS curators will implement all required corrections and use the RSN along with the SN when maintaining and extending the UMLS. When used correctly, the RSN will

support the prevention of the accidental introduction of inconsistent semantic type assignments into the UMLS. Furthermore, this way the RSN will support the exposure of other hidden errors and inconsistencies in health informatics terminologies, which are sources of the UMLS. Notably, the development of the RSN materializes the deeper, more refined Semantic Network for the UMLS that its designers envisioned originally but had not implemented.

Keywords: UMLS; Semantic Network; Refined Semantic Network; Abstraction Network; Refined Semantic Types; Intersection Semantic Types; Correction of Inconsistencies

Correspondence: zh2132@columbia.edu

DOI: 10.5210/ojphi.v6i2.5412

Copyright ©2014 the author(s)

This is an Open Access article. Authors own copyright of their articles appearing in the Online Journal of Public Health Informatics. Readers may copy articles without permission of the copyright owner(s), as long as the author and OJPHI are acknowledged in the copy and the copy is used for educational, not-for-profit purposes.

Introduction

The Unified Medical Language System (UMLS) [1,2], is derived from about 168 source terminologies. Its Metathesaurus (META) [3,4], contains over 2.9 million concepts. The UMLS Semantic Network (SN) [5-7] consists of 133 high-level, broad categories, called semantics types (STs) summarizing the content of META. One or more semantic types of the SN are assigned to each of the META concepts, describing the semantics of the concept by identifying its broad category or categories. For example, the semantics of *Dental Fistula*¹ is described by its assigned semantic type **Anatomical Abnormality**². The SN supports the ongoing integration of new and revised source terminologies into the UMLS [5].

We view the Semantic Network as an *abstraction network* of META. An abstraction network supports summarization of a repository of concepts by categorizing many concepts into a few broad categories. In [8,9] we introduced an alternative abstraction network for the UMLS, called the Refined Semantic Network (RSN). This introduction was motivated by two deficiencies of the SN, one implying the other. To explain these deficiencies, we first need to define the extent of a semantic type. The *extent* of a semantic type (ST) of the SN is defined as the set of META concepts that are assigned this semantic type.

For the SN, the extents of the semantic types are not necessarily disjoint. For example, there are 912 concepts that are assigned both **Disease or Symptom** and **Anatomical Abnormality**. Thus these 912 concepts are in two extents. However, an abstraction network is less effective in functioning as a summary if the extents of its semantic types are not disjoint, since it does not provide knowledge about the proportion of the overlaps of the extents of various semantic types.

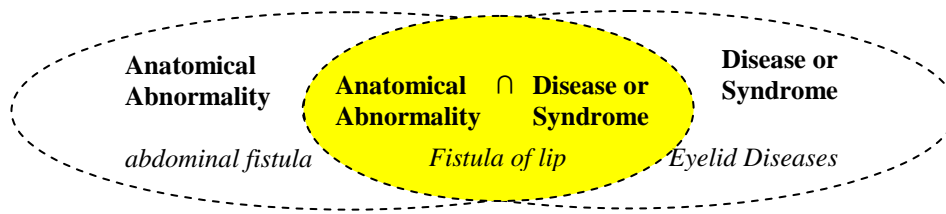


Figure 1. Example of a concept assigned two semantic types.

The second deficiency of the SN, implied by the first, is that the extents of many STs are semantically not uniform. For example, as shown in Figure 1, the concept *abdominal fistula* is assigned only **Anatomical Abnormality** while the concept *Fistula of lip* is assigned both **Anatomical Abnormality** and **Disease or Syndrome**. Hence, the extent of the ST **Anatomical Abnormality** is not semantically uniform, since some of its concepts are categorized only as **Anatomical Abnormality**, while others are categorized as **Anatomical Abnormality** and as **Disease or Syndrome**. In Figure 1, the overlapping part of the extents of **Anatomical Abnormality** and **Disease or Syndrome** is highlighted in yellow. An abstraction network is more effective in its summarization if each of its semantic types represents a semantically uniform set of concepts.

The Refined Semantic Network (RSN) [8,9] was introduced to overcome these two deficiencies of the SN. It has two kinds of *Refined Semantic Types* derived from the SN and META.

- (1) A *Pure Semantic Type* (PST) is assigned to concepts that were originally assigned only *one* semantic type. The name of a Pure Semantic Type is identical to the name of the original semantic type in the SN. The semantics of a Pure Semantic Type is the exclusive semantics of the corresponding original ST, whereby “exclusive semantics” means that the concepts assigned this semantic type are not assigned any other semantic type.
- (2) An *Intersection Semantic Type* (IST) represents a fixed combination of several STs that are all assigned to one or more concepts. An IST is not created for a combination of semantic types for which no concept appears in the UMLS. The compound semantics of an IST [8] is defined as the conjunction (AND) of the semantics of the combined STs. For example, an IST will be assigned to the concepts that are assigned both **Diseases or Syndrome** and **Anatomical Abnormality** from the SN. The name of this IST is **Disease or Syndrome** \cap **Anatomical Abnormality**. The symbol \cap is the mathematical intersection symbol and should be read as “intersected with.” For example, the concept *Fistula of lip* is assigned this IST. An IST is semantically uniform, since all concepts of its extent share the same compound semantics. The notion of Intersection Semantic Type (IST) is the most important theoretical construct in this research program. Figure 2 shows the IST **Disease or Syndrome** \cap **Anatomical Abnormality** (as a yellow box), with its parents **Anatomical Abnormality** and **Disease or Syndrome** from the original Semantic Network.

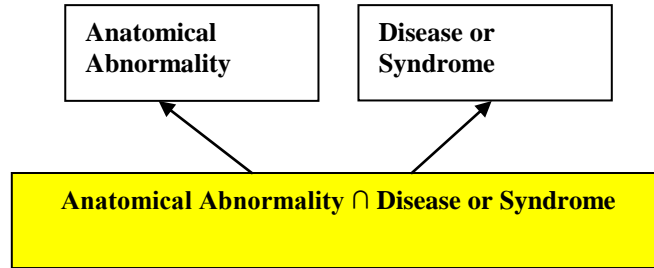


Figure 2. Example of an IST of two semantic types.

To summarize, the extent of any Refined Semantic Type is semantically uniform and the extents of all Refined STs of the RSN are disjoint. Thus, the RSN is an abstraction network that provides a better summarization of the content of META than the SN. For example, in the 2013AA release of the UMLS, the RSN shows that there are 2543 concepts that are anatomical abnormalities, 90,691 concepts that are diseases or syndromes and 989 concepts that are both anatomical abnormalities and diseases or syndromes. The SN does not make this kind of sharp distinction explicit, but the RSN does.

One more definition is required that pulls together ISTs and the sizes of extents. Whenever an IST has a small extent, i.e., this IST is assigned to at most six concepts, we will refer to the IST itself as a *small IST*. (The choice of “six” will be explained later.)

The utility of the RSN for auditing the UMLS was manifested in enabling several auditing methodologies. In [10-13] the utility of small ISTs to expose inconsistent or erroneous ST assignments was demonstrated. Group auditing techniques for large extents of Refined Semantic Types were described in [14,15]. Finally, improved categorization for conjugate and complex chemicals was explored in [16].

However, the first version of the RSN from 1998 had a major deficiency as an abstraction network. An abstraction network needs to be small to be effective, but for the 1998 release of the UMLS, the RSN had 1163 ISTs and thus was an order of magnitude bigger than the SN with its 132 STs for the 1998 release. This deficiency made the RSN a less attractive supplement for the SN as a UMLS abstraction network.

In [8] we conjectured that many of the small ISTs were erroneous and should not exist in the RSN. For example, a review of 100 out of 422 ISTs, assigned to only a single concept each, found 89 erroneous assignments. (Reminder: These 422 concepts are distinct!)

Furthermore 77 of the 1163 ISTs represented cases of *redundant ST assignments*. An assignment of an ST **A** to a concept *C* is defined as *redundant* if *C* is also assigned another ST **B**, when **B** IS-A **A**, i.e., **A** is a generalization or parent of **B**. Redundant assignments are forbidden in the UMLS [5] since they are implied. For example, in the 2011AA release of the UMLS, *Subungual swelling* is assigned both **Finding** and **Sign or Symptom**. The assignment of **Finding** is redundant since **Sign or Symptom** IS-A **Finding** in the SN. This assignment was removed in the next UMLS release.

Our plan at the time was that by an effort of removing of redundant ST assignments and other erroneous or inconsistent combinations of STs from the UMLS, only ISTs that stand for legitimate combinations of STs would remain, making the RSN considerably smaller.

Definition: An IST is considered *not legitimate* if its combination of STs satisfies any of the following:

- (1) The combination of semantic types is forbidden by the definitions or usage notes in the documentation of the semantic types of the Semantic Network. For example, the combination of **Anatomical Abnormality** and **Neoplastic Process** is forbidden.
- (2) The combination is a redundant semantic type assignment. For example, if a concept is assigned both **Finding** and **Sign or Symptom**, the assignment of **Finding** is redundant, since **Sign or Symptom** IS-A **Finding** in the Semantic Network, and thus the assignment of **Finding** to such concepts is redundant.
- (3) The semantic types of the IST are mutually exclusive in the real world, e.g. for sibling semantic types in the subhierarchy of **Organism**. For example, no real world concept is both a fish and a bird at the same time.
- (4) The semantic types of the IST do not refer to the same concept, but to two (or more) concepts with different real world semantics. An example for this case was presented in our 2003 paper [9] regarding the concept *Video Recording* and its child *Videotape recording*, which (in the 2008AB release of the UMLS) were (still) assigned both **Manufactured Object** and **Human-caused Phenomenon or Process**. This is a semantically impossible combination since an object cannot be a process. In our analysis [9] we realized that the **Manufactured Object** semantics referred to the *product of the recording* while the **Human-caused Phenomenon or Process** semantics referred to the *recording process* involved in producing this product. Indeed, in the current UMLS, both above concepts are assigned only **Manufactured Object**, similar to the 2008 assignments of *Video Recording*'s two other children *Videodisk recording* and *Videotape/Videodisc*.

Definition: An IST is considered *legitimate* if it is not illegitimate.

The legitimate ISTs deserve to be elevated to first class citizens in the RSN. Our assumption was that not too many legitimate ISTs will remain in the RSN after all the illegitimate ISTs have been removed. The legitimate ISTs occur mostly for chemical concepts where both a Structurally Viewed Chemical ST, and at least one Functionally Viewed Chemical ST, are expected, according to the definition of the **Chemical** ST [17].

After 1998 we embarked on a longitudinal study to achieve the goal of eliminating illegitimate ISTs from the UMLS in order to obtain a compact RSN. Naturally this was difficult, being outside of the National Library of Medicine (NLM), the curator organization of the UMLS, we have very limited influence on the development of the UMLS. This paper is dedicated to describing the process and steps used to “sculpt” a compact RSN out of its 1998 version and the results obtained. The term “sculpting” is used metaphorically, because a sculpture is created by

removing the excess material from a shapeless block of raw material. In the same way, the “correct” RSN with only legitimate ISTs should emerge from its initial version.

As will be reported, the goal of obtaining a compact RSN was achieved to a substantial degree, but it required a multiyear process. The process was slowed down by the phenomenon of ISTs that had been removed from the RSN being reintroduced by the NLM due to new, erroneous ST assignments in new UMLS releases. In [18], we introduced the AdviseEditor system, which can help the UMLS team with preventing the reintroduction of erroneous ISTs in the future, which would preserve the RSN as a compact Abstraction Network.

We stress again that the purpose of this paper is not to introduce new methods for auditing the UMLS, but to describe various techniques previously employed to transform the RSN into a compact abstraction network. These techniques were at the time published for their own sake, but are reviewed here for their role in sculpting the RSN (and not as novel research.) Specifically for this paper, we performed a comprehensive audit of 168 small ISTs in the 2013AA version of the UMLS.

Various terminologies are used in health informatics to support various needs. For example, SNOMED CT [19] serves for coding Electronic Health Records (EHR). LOINC [20] serves for reporting laboratory test results, and RxNORM [21] serves for prescription drugs. The accuracy of such terminologies is important for their proper use, especially in the realm of public health. For example, as a supplement for traditional public health surveillance, syndromic surveillance is being used in numerous states and localities to detect a potential large-scale biologic attack [22]. Hripsak et al. demonstrated the feasibility of using electronic health record data for syndromic surveillance, in which terminologies are used to encode the narrative clinical notes by natural language processing techniques [23]. However, this task is challenging due to complex grammar in free text as well as ambiguous concepts in the terminologies. Therefore, identifying and correcting ambiguities, inconsistencies, and errors in the terminologies may accelerate the adoption of standard terminologies in the Syndromic Surveillance System, which would improve public health. However, due to their size and complex modeling, errors and inconsistencies are unavoidable and Quality Assurance (QA) for each terminology is required.

However, QA of terminologies is difficult, requires experts with multiple training and requires extensive budgets.

To facilitate effective QA of a terminology, it is preferred to apply computational techniques that automatically find concepts with high likelihood of errors. Such techniques will improve the yield of QA resources available, where the yield is expressed in terms of number of errors found and corrected per effort spent.

One approach for effective QA of terminologies used in health informatics utilizes the fact that many of them are source terminologies for the UMLS. That is, a specific concept of the UMLS may be mapped into several concepts of various source terminologies. As it happens, the modeling of the concept in the various sources may not be consistent with one another and some modeling may even be outright wrong. While such cases are very difficult to detect by just performing QA on a specific terminology, the inconsistency or error may be manifested by the assignment of multiple semantic types for this concept. For example, one semantic type may follow the meaning of the concept in one source terminology, while another semantic type may

reflect the meaning of the same concept in the other terminology, or one of the semantic types may have been assigned by mistake. Furthermore, concepts with erroneous or inconsistent semantic types assignment may indicate other errors or modeling problems for that concept. However, the erroneous modeling of such concept came from one or more of the source terminologies. In this way, a hidden error or problem in the modeling of a concept in one of the UMLS source terminologies, is discovered and can be corrected due to improper combination of the semantic types in the UMLS. Examples of such phenomenon are demonstrated in the Result Section.

Methods

This paper describes the techniques, process, and results that enabled us to reshape the RSN into a compact abstraction network, materializing the vision defined more than a decade ago. The methodology framework of “sculpting” the RSN is illustrated in Figure 3. We will describe the techniques employed in the sculpting process in detail, pointing out the longitudinal history of the techniques invented in this framework.

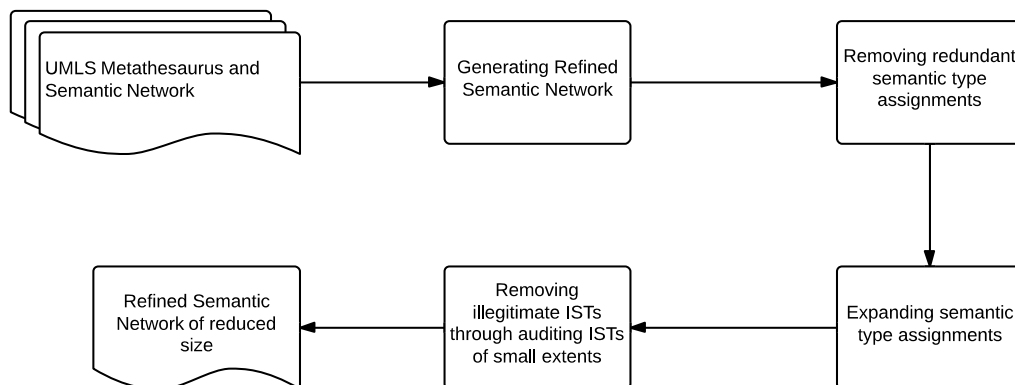


Figure 3. The methodology framework of “sculpting” the RSN for the UMLS.

Generating Database Tables for the Refined Semantic Network

To enable this longitudinal study of sculpting the RSN, we generated two database tables for the Refined Semantic Network for each release of the UMLS, starting with version 2006AC. We first used the MetamorphoSys tool developed by the NLM to generate the Oracle loading scripts of the UMLS Rich Release Format (RRF) tables. Then we loaded the RRF tables into our Oracle database system. In the original RRF schema, the “MRSTY” table stores a single semantic type assignment to a concept in one row. Therefore, if a concept is assigned multiple semantic types, there would be multiple rows in “MRSTY” for this concept. Based on “MRSTY,” we generated an RSN table called “COMBOS” to store the compound semantic type assignments to a concept. In this way, multiple rows for a concept assigned multiple semantic types (in multiple rows in the “MRSTY” table) are combined into a single row in the “COMBOS” table. In other words, the “COMBOS” table stores the assignment of a Refined Semantic Type (either a PST if a concept is

assigned a single semantic type or an IST if a concept is assigned multiple semantic types) to a concept in the RSN. We further generated another RSN table “CPTCNTBYST” (ConcePT CouNT BY Semantic Type) to store the number of concepts assigned each Refined Semantic Type with its description. Leveraging the database system’s capability of handling a wide range of queries, we can use SQL queries to retrieve ISTs of certain extents for human auditing in the sculpting process. The SQL scripts for generating the database tables “COMBOS” and “CPTCNTBYST” for the RSN are provided as supplementary files. Note that the scripts can be executed only after the UMLS RRF tables have been loaded.

Removing Redundant Semantic Type Assignments

In the framework of sculpting the RSN, we employed a fully automated algorithm to detect redundant ST assignments. As mentioned before, there were 77 ISTs in the 1998 UMLS release implementing forbidden redundant assignments [5]. In 2002, we designed an algorithm for detecting all META concepts with redundant ST assignments [24]. In 1998 there were 8622 such concepts, which we reported to National Library of Medicine. From that time on, we periodically monitored the UMLS for redundant ST assignments, reporting systematically to the NLM on our findings. Seemingly influenced by our publication [24] and repeated reports, the NLM implemented an automatic procedure that removes redundant ST assignments before each release of the UMLS [25].

Extending as new Form of Sculpting

To reiterate, we call this kind of action of eliminating erroneous ISTs from the RSN “sculpting.” The sculpting of the RSN was continued by *extending* some IST extents [14,15], which was done after detecting concepts missing appropriate semantic type assignments. That is, sculpting does not always involve removing erroneous ISTs, but always involves correcting ST assignments. In other words, sometimes, concepts are missing a necessary second ST assignment, and correcting this may increase the size of an IST extent that was not small to begin with. This phenomenon was demonstrated for the IST **Experimental Model of Disease** \cap **Neoplastic Process** which was enlarged from 33 to 948 concepts by Chen et al. [15], and was further expanded to 1397 concepts using another technique in work of Chen et al. [26]. Similarly, the IST **Governmental or Regulatory Activity** \cap **Intellectual Product** was expanded from 22 to 32 concepts [15]. The extent of the IST **Environmental Effect of Humans** \cap **Hazardous or Poisonous Substance** was enlarged from three to nine concepts, i.e., it was no longer a small IST [14].

Removing Illegitimate ISTs through Auditing ISTs of Small Extents

In our more than 15 years of research in QA of medical terminologies, we identified two recurring themes, regarding concentration of errors in medical terminologies [27]. Errors typically appear in *complex* concepts or in *unusual* concepts. The following rationale is offered. Modeling of complex concepts is more difficult than modeling of simple concepts, and thus they have a higher likelihood of (human) errors. For “unusual” concepts, the reason for the “uncommon” modeling may be the unique nature of these concepts, but there is also a high likelihood that the modeling is wrong, and this is why these concepts appear to be unusual.

The interpretation of “complex” or “unusual” varies from one terminology to another according to the different nature of various terminologies. Wang *et al.* have shown that complex concepts

in *overlapping partial areas* [28] have a high likelihood of errors in SNOMED CT [29,30]. (Space does not allow an explanation of partial areas.) If a partial area is small, i.e., it contains few concepts, we can label these concepts as being unusual. It has been shown that small partial areas contain relatively more errors in SNOMED CT and NCIt [27,31]. An IST consisting of multiple STs is more complex than a single ST, because of compound semantics. In [14,15,26], Chen has found many errors in the extents of ISTs, e.g., in **Experimental Model of Disease** \cap **Neoplastic Process**. A small IST is unusual, since out of 2.9 million concepts in the META, only a few concepts are assigned its ST combination. Thus, we hypothesized that ISTs assigned to only a few concepts are more likely to have concepts with inconsistent or erroneous ST assignments, since the concepts assigned such ISTs are both complex and unusual. Therefore, we conducted a study for auditing concepts of ISTs with small extents [10]. Our finding was that for ISTs with up to six concepts there is a higher likelihood of wrong ST assignments compared to concepts assigned an IST with a larger extent. If all the concepts assigned a specific small IST have an erroneous ST assignment, this IST disappears from the RSN, after the appropriate corrections have been made. This makes the RSN smaller, as desired.

Over the years, we have conducted several studies, e.g. [11-13], where a team of domain experts audited samples of small ISTs. We forwarded the consensus reached by our auditors to the UMLS editors for review. In some cases, the UMLS editors chose an alternative correction rather than the one suggested by our auditors, but the “erroneous” ISTs still disappeared from the RSN, whenever no concept was left with the combination of STs of this IST.

For this paper, we performed a new audit of all ISTs with small extents (1-6 concepts) left in the 2013AA UMLS release, removing inconsistent or erroneous semantic type assignments. The resulting RSN, with a smaller number of ISTs, is an outcome of this paper. The Java program used to generate a sample of concepts (concepts assigned small ISTs in this paper) is provided as supplementary file. Given a list of UMLS Concept Unique Identifiers (CUIs) as input, the program will generate a sample of corresponding concepts, including concept names, Refined Semantic Types, definitions, and contextual information of concepts, such as their parents, children, and siblings. The contextual information helps expose erroneous and inconsistent modeling of a concept. Note that the Java program can be executed only after the two RSN tables “COMBOS” and “CPTCNTBYST” have been generated.

Results

First, we will report on the progress of sculpting the RSN over multiple releases of the UMLS. Table 1 presents the information we monitored, including the number of concepts, number of STs and ISTs, number of concepts with redundant assignments and their ISTs, as well as the number of small ISTs with their extent sizes, the combined number of ISTs with extent sizes 1-6, and finally their numbers of concepts, for different UMLS releases. Some of this information is also illustrated in Figure 4.

Information was regularly collected starting with UMLS version 2006AC. During 2006-2007 our research group submitted reports of redundant and wrong ST assignments for small ISTs to the NLM. For example, for the 2006AC version, we submitted 42 erroneous, small extent IST assignments, 39 of which had one concept and three had two concepts each. The NLM implemented most of our corrections, causing many small ISTs to disappear. Note that we never

received feedback from the NLM regarding our error reports, but by reviewing the next releases of the UMLS we could track the changes, presumably caused by our reports.

Of these 42 small ISTs, 38 disappeared by the 2007AA version. One of these ISTs was **Mammal** \cap **Experimental Model of Disease** assigned to the concept *Knock-in Mouse*, with erroneous compound semantics; of course a mammal cannot be a disease. Another IST that disappeared, **Congenital Abnormality** \cap **Neoplastic Process**, which was assigned to *Port-Wine Stain*, was a forbidden combination of STs according to the UMLS usage note of the ST **Neoplastic Process** [17]. No change was made only for one IST **Gene or Genome** \cap **Enzyme**.

In three cases, the concept assignments were changed, but the IST remained in the RSN, because a new concept was simultaneously assigned the same IST by the UMLS editors.

(More about such occurrences will be discussed later.) In other words, in some cases new errors were introduced while old errors were being corrected.

We reiterate that the NLM did not always make the corrections that we suggested. However, the changes they made to the ST assignments still frequently resulted in the deletion of small ISTs. Nevertheless, the total number of ISTs between 2006AC and 2007AA was only reduced from 559 to 555. While some erroneous small ISTs disappeared, new ones were created due to the assignment of multiple STs to new concepts coming from new sources added to the UMLS or from new releases of existing UMLS sources.

A systematic decrease in the number of ISTs is evident in Table 1 from 2007AC till 2008AB including 2008AA. The number of ISTs went down from 532 in 2007AC to 397 in 2008AB, a reduction of 135 ISTs, 110 of which were small ISTs with a total of 235 concepts, including in particular 78 ISTs with one or two concepts each. The removal of such ISTs from the RSN is consistent with the finding of Gu *et al.* [10] that concepts assigned ISTs with extents of up to six concepts have a higher likelihood of erroneous ST assignments than concepts assigned larger extent ISTs.

Many erroneous assignments have been removed either due to our reports (e.g., [11]) or independently by the UMLS team. Furthermore, as mentioned in the previous section, the NLM implemented an automatic procedure for detecting all redundant assignments in the UMLS, which has been applied before any new UMLS release starting in 2008 [25]. As can be seen in Table 1, no redundant ST assignments were detected from the 2008AA to the 2013AA release, except for one case in 2011AA (reason unknown) that was subsequently corrected in 2011AB.

During 2009 – 2013 a plateau was reached, with about 400 ISTs, of which about 170 are small ISTs, containing a total of about 410-420 concepts. One may think that the RSN had reached a stable state during these years. However, the impression created by the numbers of ISTs and small ISTs is misleading.

Table 1. Progress of RSN over time

UMLS Release	#cpts	#STs	#ISTs	#cpts w/ redundant STs	#ISTs w/ redundant Assign	#ISTs w/ 1 cpt	#ISTs w/ 2 cpts	#ISTs w/ 3 cpts	#ISTs w/ 4 cpts	#ISTs w/ 5 cpts	#ISTs w/ 6 cpts	#ISTs w ≤ 6 cpts	#cpts in IST w ≤ 6 cpts
1998	476K	132	1163	8622	77	422	n/a	n/a	n/a	n/a	n/a	n/a	n/a
2001	800K	134	874	12161	40	322	113	64	35	28	25	587	1170
2006AC	1.4M	135	559	91	7	124	68	37	32	26	18	305	737
2007AA	1.4M	135	555	598	11	111	65	40	33	23	17	289	710
2007AC	1.5M	135	532	0	0	116	56	35	34	20	15	276	659
2008AA	1.6M	135	464	3	2	105	44	25	25	15	14	228	499
2008AB	1.9M	135	397	0	0	64	30	29	14	17	12	166	424
2009AA	2.1M	135	381	0	0	59	32	24	13	16	11	155	393
2009AB	2.2M	135	385	0	0	61	30	25	15	14	13	158	404
2010AA	2.2M	133	384	0	0	58	32	24	15	16	9	154	388
2010AB	2.4M	133	392	0	0	66	35	19	16	16	8	160	385
2011AA	2.4M	133	409	1	1	75	38	24	16	17	6	176	408
2011AB	2.6M	133	406	0	0	72	34	25	16	19	8	174	422
2012AA	2.6M	133	407	0	0	73	33	26	16	17	7	172	408
2012AB	2.8M	133	402	0	0	61	37	26	14	18	9	165	413
2013AA	2.9M	133	401	0	0	63	33	27	18	16	11	168	428
2013 Audit	2.9M	133	336	0	0	48	28	10	3	8	6	103	222

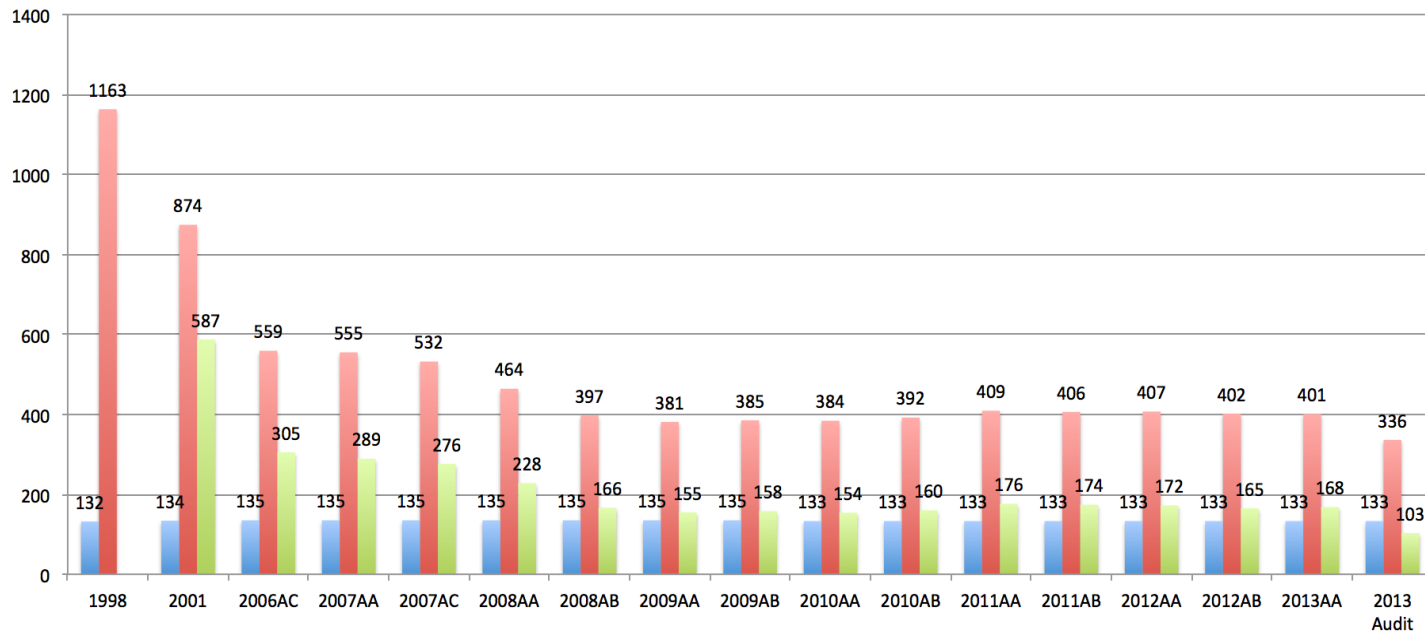


Figure 4. Progress of the Semantic Network, ISTs and ISTs with small extents. Blue bars show the number of semantic types in the UMLS Semantic Network. Red bars show the number of ISTs in the RSN. Green bars show the number of ISTs with small extents.

During the period from 2009 to 2013 two ongoing phenomena have been observed that have counteracting effects on the numbers of ISTs. From one side, erroneous ST assignments were detected by the UMLS team and as a result 69 erroneous ISTs of typically small extents disappeared (see Table 2). From the other side, new UMLS concepts were assigned semantic types and for 78 of them, new combinations of STs were created (see Table 2), leading to the addition of new ISTs of typically small extents. Many times those newly created ISTs are the same ones that had been removed from the RSN in earlier releases, because erroneous assignments of such ISTs were corrected.

According to Table 2, there are 35 such ISTs over the five releases 2011AA – 2013AA. Furthermore, 11 of these ISTs were added and deleted more than once during this period. These “oscillations” could have been avoided if the NLM would have adopted the RSN as an additional abstraction network for monitoring the UMLS. Besides our publications about the RSN and its use, the RSN was also presented at the NLM-sponsored workshop on “Future Directions of the Semantic Network” [32]. A recommendation how to avoid “oscillations” appears in the Discussion Section.

Table 2. Progress of IST removal in the past five releases

	2011AA	2011AB	2012AA	2012AB	2013AA	Total
ISTs	409	406	407	402	401	
Small ISTs	176	174	172	165	168	
New ISTs	23	17	13	14	11	78
Appeared Before	12	6	4	6	7	35
Repeated Previously	3	1	3	1	3	11
Number of Deleted ISTs	6	20	12	19	12	69

When we reviewed the new ISTs in the 2013AA and 2012AA releases of the UMLS, we found that most of them are illegitimate. For example, in Table 3 for the 11 new ISTs in the 2013AA release, the IST **Mental or Behavioral Dysfunction** \cap **Steroid** \cap **Pharmacologic Substance** is illegitimate, because a dysfunction cannot be a chemical. **Amino Acid, Peptide, or Protein** \cap **Pharmacologic Substance** \cap **Indicator, Reagent, or Diagnostic Aid** \cap **Element, Ion, or Isotope** is assigned to only one concept *Fluciclatide F18*, which is used as radioactive probe in PET imaging according to the definition of this concept. However, the UMLS usage note of ‘**Indicator, Reagent, or Diagnostic Aid**’ [33] states: “Radioactive imaging agents should be assigned to this type and not to the type ‘**Pharmacologic Substance**’ unless they are also being used therapeutically.” Thus, the assignment of ‘**Pharmacologic Substance**’ is deemed wrong.

In 2012AA, the IST **Carbohydrate** \cap **Chemical Viewed Functionally** was assigned to the concept *viridaphin A(1) glucoside* (see Table 4). It is surprising that a general semantic type such as **Chemical Viewed Functionally** is assigned to this concept. According to the rules of the UMLS [5], each concept should be assigned the most specific applicable ST. Our team member performing this audit proposed to change this semantic type assignment to a grandchild of **Chemical Viewed Functionally**, namely **Antibiotic**.

Finally, we report the results of an audit of the 428 concepts of the small ISTs of the 2013AA version. They were divided into two sets, 98 non-chemical concepts and 330 chemical concepts. The first set was reviewed by two domain experts, an MD, trained in medical terminologies

(G.E.) and a PhD who specialized in techniques for auditing medical terminologies after receiving training in Sports Medicine (Y.C.). The second set was audited by a Chemistry Professor (L.C.), experienced in auditing chemical concepts. All three auditors are co-authors. They used the Neighborhood Auditing Tool (NAT) [12] designed at NJIT and have previously audited UMLS ST assignments to concepts.

Table 3. New ISTs in UMLS release 2013AA.

New ISTs in 2013AA	Extent	Appeared also in Years							
Bacterium \cap Pharmacologic Substance	1	2012AA	2011AB	2011AA	2010AB	2010AA	2009AB	2008AA	2007AC
Congenital Abnormality \cap Finding	1	2011AA	2007AC	2007AB					
Laboratory or Test Result \cap Laboratory Procedure	1	2008AA	2007AC	2007AB	2007AA				
Pathologic Function \cap Anatomical Abnormality	1	2007AC	2007AB	2007AA					
Mental or Behavioral Dysfunction \cap Steroid \cap Pharmacologic Substance	1								
Medical Device \cap Indicator, Reagent, or Diagnostic Aid	4	2012AA	2008AA	2007AC	2007AB	2007AA			
Amino Acid, Peptide, or Protein \cap Pharmacologic Substance \cap Indicator, Reagent, or Diagnostic Aid \cap Element, Ion, or Isotope	1								
Carbohydrate \cap Pharmacologic Substance \cap Food	2								
Lipid \cap Pharmacologic Substance \cap Food	5								
Biomedical or Dental Material \cap Food	2	2008AA							
Biomedical or Dental Material \cap Element, Ion, or Isotope	1	2007AA							
Legend									
	IST removed once								
	IST removed twice								
	IST appeared the first time								
	IST appeared the second time								

Table 4. New ISTs in 2012AA

New ISTs in 2012AA	Extent	Appeared also in Years		
Bacterium \cap Eukaryote	1			
Therapeutic or Preventive Procedure \cap Biomedical or Dental Material	4			
Natural Phenomenon or Process \cap Indicator, Reagent, or Diagnostic Aid	1			
Medical Device \cap Indicator, Reagent, or Diagnostic Aid	1	2008AA	2007AB	2007AA
Medical Device \cap Clinical Drug	1	2010AB		
Qualitative Concept \cap Clinical Attribute	1			
Amino Acid, Peptide, or Protein \cap Biomedical or Dental Material \cap Inorganic Chemical	1			
Carbohydrate \cap Chemical Viewed Functionally	1			
Chemical Viewed Functionally \cap Inorganic Chemical	1			
Pharmacologic Substance \cap Vitamin \cap Indicator, Reagent, or Diagnostic Aid	2			
Pharmacologic Substance \cap Vitamin \cap Inorganic Chemical	2	2008AA	2007AB	2007AA
Pharmacologic Substance \cap Food	1	2008AA	2007AB	2007AA
Vitamin \cap Element, Ion, or Isotope	1			
Legend				
		IST removed once		
		IST removed twice		
		IST appeared the first time		
		IST appeared the second time		

Table 5 summarizes the results of auditing 29 small non-chemical ISTs from the 2013AA release. If all audit results were implemented in the 2013AA release, 16 out of 29 small non-chemical ISTs would disappear and 2 new non-chemical ISTs would be added, resulting in 15 such ISTs.

Table 5. Auditing impact on 2013AA non-Chemical ISTs of the sculpted RSN

Extent size of IST	Starting # of Non-Chemical ISTs 2011AA	# of Non-Chemical ISTs deleted by audit	Percentage of such ISTs deleted	# of Non-Chemical ISTs added by audit	Percentage of Non ISTs added	# of Non-Chemical ISTs after audit	Net reduction
1	7	5	71.4%	1	14.3%	3	57.1%
2	3	2	66.7%	0	0%	1	66.7%
3	5	3	60%	1	33.3%	3	60%
4	6	4	66.7%	0	0%	2	33.3%
5	2	1	50%	0	0%	1	50%
6	6	1	16.7%	0	0%	5	16.7%
Total	29	16	55.2%	2	6.9%	15	48.3%

For example, the IST **Congenital Abnormality** \cap **Finding** is only assigned to *Congenital abnormality of systemic artery*. However, the UMLS usage note of **Finding** [33] states that “Only in rare circumstances will findings be double-typed with either ‘Pathologic Function’ or ‘Anatomical Abnormality’.” **Congenital Abnormality** has an IS-A relationship to **Anatomical Abnormality**. Thus, the assignment of **Finding** should be removed. Consequently, this IST should disappear from the RSN.

Table 6 summarizes the results of auditing 139 small chemical ISTs from the 2013AA version. We see that 30 (= 139 - 109) small chemical ISTs were found correct and remained in the RSN. Also 58 new chemical ISTs were created in the auditing process, leaving a balance of 88 small chemical ISTs.

Table 6. Auditing impact on 2013AA Chemical ISTs of the sculpted RSN

Extent size of IST	Starting # of Chemical ISTs 2011AA	# of Chemical ISTs deleted by audit	Percentage of ISTs deleted	# of Chemical ISTs added by audit	Percentage of ISTs added	# of Chemical ISTs after audit	Net reduction
1	56	44	78.5%	33	58.9%	45	19.6%
2	30	19	63.3%	16	53.3%	27	10%
3	22	21	95.5%	6	27.3%	7	68.2%
4	12	11	91.7%	0	0%	1	91.7%
5	14	10	71.4%	3	21.4%	7	50%
6	5	4	80%	0	0%	1	80%
Total	139	109	78.4%	58	41.7%	88	36.7%

In some cases, an audit resulted in an ST combination which added a concept to the extent of an existing IST, which may have been large or small. For example, the concept *TrioMatrix* is the only concept assigned **Amino Acid, Peptide or Protein** \cap **Biomedical or Dental Material** \cap **Inorganic Chemical**. This is an implantable orthopedic device, namely, a surgical bone implant, composed of living or natural materials. Because **Amino Acid, Peptide, or Protein** is an **Organic Chemical**, it should not be assigned together with **Inorganic Chemical**. With the assignment of **Inorganic Chemical** removed, this concept is reassigned the very large IST **Amino Acid, Peptide or Protein** \cap **Biomedical or Dental Material**, while the previous IST disappears.

The results of the audit of version 2013AA appear in Table 1. The last row in Table 1 shows the impact of this audit on the size of the RSN. Only 15 small non-chemical ISTs and 88 small chemical ISTs are left in the RSN. The total number of ISTs (small and large) decreases to 336 (fourth column, Table 1).

The audit reports of both samples were submitted to the NLM for review. Based on past experience, we expect the recommendation to be at least partially incorporated into the UMLS, which will reduce the size of the RSN.

Figure 5 shows an excerpt of the RSN after the sculpting effort. All the ISTs are displayed as yellow boxes. Chemical semantic types are shown as red text. The part above the dashed blue line consists of the original semantic types from the Semantic Network. The part of Figure 5 below the dashed blue line shows the ISTs with at least one non-chemical intersecting ST and their parent ISTs even if all the STs of the parent ISTs are chemical, e.g., **Carbohydrate** \cap

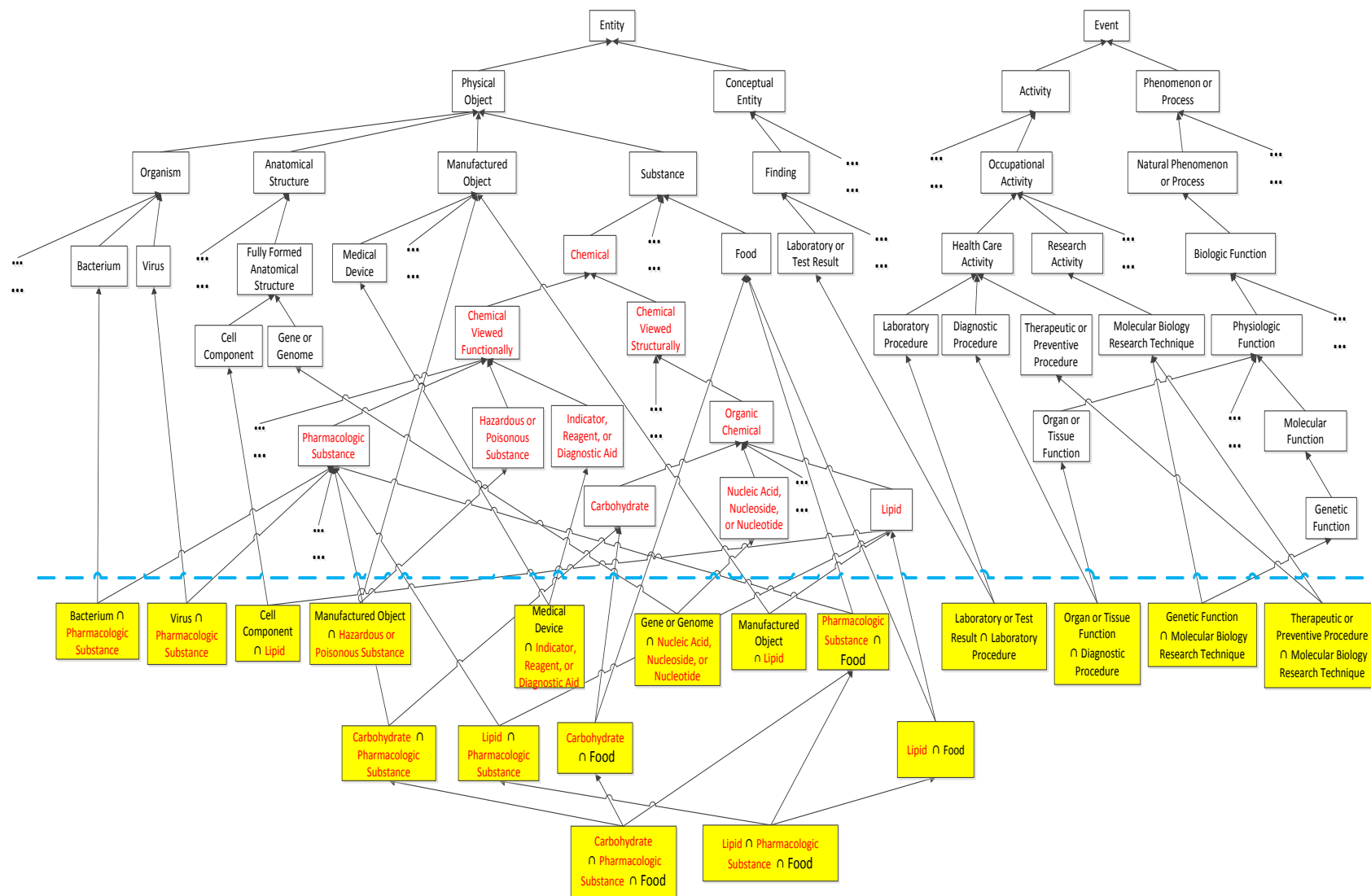


Figure 5. An excerpt of the RSN after sculpting. This figure shows all the ISTs with at least one non-chemical ST and their ancestors. All the Chemical STs are marked in red. All the ISTs are shown as yellow boxes.

Pharmacologic Substance. As can be seen in the figure, the parents of the ISTs combining two STs are their corresponding semantic types in the original Semantic Network. All those ISTs are in two rows immediately below the dashed blue line. The ISTs that combine three STs are located in the third row below them. The parents of those ISTs contribute the three constituent STs. Omitted parts of the SN are hinted at by dots.

In this paper, we advance in two ways beyond the auditing of small ISTs reported on in our previous publications [10,11]. One new feature is “group auditing” of small ISTs, that is, auditing a small group of semantically similar concepts as one unit, as opposed to auditing concepts one by one. Group auditing of small ISTs is expected to be more accurate and easier than auditing a list of concepts in random order. This is distinct from group auditing of large ISTs [14].

For example, the small IST **Human-caused Phenomenon or Process** \cap **Natural Phenomenon or Process** was assigned to four similar concepts *Chemical Hazard Release*; *Biohazard Release*; *Incidents, Biological* and *Accidents, Biological* from MSH. The first two are children of the concept *Accidents*, assigned **Phenomenon or Process** (in 2010AA), and assigned **Injury or Poisoning** (starting in 2010AB). The other two are concepts without parents or children. The definitions of the first two are almost perfectly parallel, “Uncontrolled release of a chemical (Biological material) from its containment that either threatens to, or does, cause exposure to a chemical (biological) hazard, such an incident may occur accidentally or deliberately.” Following the definitions and children listed (e.g. *Bhopal Accidental Release* assigned **Human-caused Phenomenon or Process**) these four concepts should be assigned only **Human-caused Phenomenon or Process**.

The second advanced feature is an important side effect of the group auditing of concepts of small ISTs, the discovery of other inconsistencies in such concepts or their neighbors. Typically, an erroneous ST assignment indicates a misconception or ambiguity of the concept, which may be manifested in other inconsistencies. A concept belonging to a small IST is algorithmically detectable, initiating a manual review of such a concept. However, there may be no known automatic method to detect the other inconsistencies found during this review. Their discovery is a byproduct of the review of small ISTs.

We illustrate several such inconsistencies found during the manual review of the ISTs in the previous example. Like the previous two concepts, *Accidents, Biological*, should have a parent *Accidents*, which in turn has a wrong parent *Injury*. The other isolated concept in the group, *Incidents, Biological* should have the concept *Incident* (from HL7V3.0 [34]) as a parent. Such a hierarchical relationship between concepts from two sources can be added by the NLM into the MTH source. *Incident*, by its definition, should be assigned **Phenomenon or Process** rather than **Idea or Concept**. The audit of ST assignments of these four concepts as a group suggested the exploration of other neighboring concepts, finding these other inconsistencies. At the same time, those errors suggest the correction of the modeling of concepts in individual health informatics terminologies, by e.g., adding IS_A relationships or a missing concept. These corrections were discovered only due to inconsistent multiple ST assignments in the UMLS.

Another example of group auditing appears with the IST **Manufactured Object (MO)** \cap **Self-help or Relief Organization (SHO)**. This IST is assigned only three concepts: *night shelter*,

social service facility, and *community resource center*. The assignment of **MO** to these three concepts is puzzling. All three concepts are from the Alcohol and Other Drug Thesaurus (AOD) [35].

Upon reviewing the context of this set, we see that *night shelter* has three siblings: *day shelter*, *dry shelter*, and *web shelter*, all children of *shelter homeless*. All of them are from AOD and assigned only **SHO**. *Shelter homeless* in turn has a sibling *community resource center* and a parent *social service facility* both assigned **MO** \cap **SHO**. Finally *social welfare* assigned only **SHO** is the parent of *social service facility*. Reviewing the context, the auditor suggested removing the **MO** ST from the assignment of these three first concepts for consistency.

However, this case of inconsistent ST assignment can be a trigger to review the AOD modeling. It seems that the assignment of **MO** ST was due to the use of the words “facility” and “center”, in two of these concepts, interpreting them to refer to the building hosting the self help organization. This interpretation exposes an ambiguity in the AOD modeling between the organization and the building hosting it. Our suggestion with regard to AOD modeling is to disambiguate by creating two concepts *social service center* with the **SHO** semantics that will be the parent of *shelter homeless* and *community resource center* and the child of *social welfare* and a concept *social service facility* with **MO** semantics referring to the building hosting it. This way an inconsistency in UMLS semantic type assignment exposes a modeling problem in the AOD source terminology which otherwise would be hidden.

Discussion

In the paper of McCray and Hole [7], which introduces the UMLS Semantic Network, the authors stated “The current scope of the network is quite broad, yet the depth is fairly shallow. We expect to make future refinements and enhancements to the network based on actual use and experimentation.”

This plan for further development of the SN was never executed, in spite of obvious needs. For example, describing the integration of the Gene Ontology (GO) [36] into the UMLS, Lomax and McCray [37] point to deficiencies of the SN in covering the Genomics field. While the UMLS META grew to be about 96-fold larger than in its first release [38], the SN changed very little, with a few semantic types being added or deleted over the years (See, for example, the third column in Table 1). Proposed extensions of Genomics coverage in the SN [39,40] were not implemented.

One may consider the RSN as a step towards fulfilling the above original vision of the designers of the UMLS Semantic Network, since it adds to the network depth by adding ISTs in a way that extends the SN downwards. Another important observation is that the RSN is derived from the SN and the ST assignments to META concepts in an intrinsic way without using any knowledge sources that are external to the UMLS. The extension provided by the RSN follows the same approach and is thus in line with the vision for the UMLS expressed at its founding.

The RSN helps identifying ISTs with proper compound semantics and treating them as legitimate first class citizens, while removing all the semantically invalid ST combinations. For example, in the 2013AA release of the UMLS, 85 ISTs are assigned to at least 100 concepts, 36 ISTs are

assigned to at least 500 concepts and 21 of these ISTs are assigned to at least 1000 concepts, demonstrating their validity as legitimate broad categories for META concepts.

Only 29 small *non-chemical* ISTs exist in the 2013AA release. According to our hypothesis [10], concepts assigned such small ISTs have a high likelihood of wrong or inconsistent ST assignments. Indeed, many such ISTs have already disappeared in past releases. We applaud the efforts of the NLM editorial and QA teams achieving the current situation, by preventing redundant ST assignments and eliminating many erroneous small ISTs. Furthermore, even for the current (2013AA) small, non-chemical ISTs, the hypothesis of Gu *et al.* [10] was found true in our recent audit report (see Table 5), according to which only 15 (about half) of the small non-chemical ISTs are legitimate, i.e., are meaningful in the real world.

The situation is different for small *chemical* ISTs. As mentioned earlier, ISTs are expected to exist for chemical concepts, due to their multiple structural and functional views. As a result there are 28 ISTs which represent combinations of four chemical STs. For example, 118 concepts are assigned **Amino Acid, Peptide, or Protein** \cap **Pharmacologic Substance** \cap **Immunologic Factor** \cap **Indicator, Reagent, or Diagnostic Aid**. While many of the small chemical ISTs are legitimate, Table 6 indicates that a large portion of them, $(109/139) = 78\%$ are erroneous. However, many (58) small chemical ISTs were added during the audit, when the concepts of the deleted ISTs were assigned correct semantic types. As a result, 88 small chemical ISTs were left in the RSN after our audit (see Table 6). The concepts of the other 51 $(109-58)$ small chemical ISTs were typically reassigned existing ISTs with larger extents, as shown in the example above. The contrast between the 88 small chemical and the 15 small non-chemical ISTs reflects the high frequency of categorizing chemical concepts by both structural and functional Chemical STs, as documented in the usage note for the **Chemical** ST of the UMLS [33].

In this paper, we stressed the success of group auditing of small ISTs in exposing other errors (besides semantic type assignments) as well. Such errors may not otherwise be detectable algorithmically. We recommend the auditing of concepts that were assigned small non-chemical ISTs in past UMLS releases, and of their neighboring concepts, for exposing other errors which may be hard to discover by a program. The storage of previous releases of the UMLS, can enable exposing such errors. Furthermore, these errors may expose errors in individual UMLS source terminologies, which otherwise, would be hard to expose.

Interestingly, once all erroneous ISTs will have been eliminated from the RSN, the hypothesis of [10], i.e., ISTs with small extents contain concepts with a relatively high likelihood of erroneous ST assignments, will not be true anymore. This is based on the expectation that the current NLM practice of re-assigning erroneous ISTs to new UMLS concepts will cease. This practice has turned the effort of sculpting the RSN into a Sisyphean task, since once an erroneous IST has been eliminated by correcting the erroneous ST assignments of its concepts, this IST often reappears in a future release, due to new erroneous semantic type assignments.

We recommend that the RSN should be used as a support tool for preventing re-assignment of illegitimate ISTs without hurting the efficiency of the UMLS team. This issue was the subject of another line of research of some of the authors [18]. In that work we analyzed the various reasons why some ST combinations should not be assigned to new UMLS concepts. These reasons include redundant ST assignments, detectable algorithmically [24] and conditions listed in the UMLS usage notes [33], as illustrated earlier. Among the reasons is also the mutual exclusion

between sibling STs in certain subtrees of the SN, e.g., in the subtree of **Organism** describing the animal kingdom.

Furthermore, an interactive, web-based system AdviseEditor was developed, which accepts as input a pair of STs, and determines whether this pair is legitimate or illegitimate (or whether more research is required for this pair). AdviseEditor can also process triples, quadruples and quintuples of semantic types in interactive mode and in batch mode [18].

We recommend that the UMLS team of the NLM will take advantage of the AdviseEditor tool to preserve the RSN as an additional compact abstraction network for the UMLS (in addition to the SN). Working this way will prevent many categorization errors in the future. Furthermore, preventing these errors will save the UMLS team the effort currently spent on meticulously correcting them.

Limitations

Some limitations are noteworthy in interpreting this study. First, the auditing of small ISTs was conducted by human experts. Thus, some suggestions might be subjective and arguable. Nevertheless, in this study, we tried to reduce the subjectivity by having multiple domain experts review the ISTs of small extents. Second, as we mentioned earlier, the NLM, as the curator organization of the UMLS, has the full control over its development. Therefore, we have limited influence on its development. According to the findings in this study and our past experience, even if the NLM did not adopt all of our suggestions to correct ambiguities, inconsistencies, and modeling errors in the UMLS, our auditing reports still played a positive role for its QA.

From a QA perspective, external auditing can be considered as a necessary task and an ethical advantage, because the NLM team cannot influence what external auditors want to investigate. Otherwise, there would be the appearance of a conflict of interest, which diminishes the credibility and integrity of the QA process. Third, we performed the auditing of source terminologies in the context of the UMLS, it might be difficult to make suggested changes in individual source terminologies in their own models, e.g. Description Logic. In the recent years, numerous domain ontologies are emerging for health informatics applications. Due the heterogeneous development models and domain knowledge of their curators, the quality issue has been recognized as one of a factor that has slowed down their adoption [41]. We suggest that a rigorous auditing methodology framework should be incorporated in the life cycle of domain ontologies.

As a final note, we would like to stress the importance of longitudinal studies in Medical Informatics. In Medicine, studies extending over 5 or more years are not uncommon. In Medical Informatics we have seen few such studies. The present paper shows that longitudinal studies are possible and fruitful in Medical Informatics.

Conclusions

We reported on a longitudinal study of the process of improving the UMLS as a result of auditing its semantic type assignments. The main instrument used in this sculpting is the auditing of small ISTs containing concepts with a high likelihood of erroneous or inconsistent ST assignments. Over the years, the external auditing of the UMLS has been shown to complement the internal auditing at the NLM. Numerous audit reports were submitted and reviewed by NLM team

members, who also performed their own auditing. The NLM also adopted automatic testing for redundant ST assignments before a new UMLS version is released. Furthermore, we conducted a dedicated, comprehensive audit of all 168 small ISTs in the 2013AA version for this paper that can support auditing of individual health informatics terminologies widely used for public health. As a result, after the audit is used to eliminate erroneous small ISTs, the RSN becomes a compact abstraction network with a size of the same order of magnitude as the SN, providing better comprehension support for the content of the META.

Acknowledgement

This work was partially supported by the NLM under grant R-01-LM008445-01A2.

References

1. Bodenreider O. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* 32(Database issue), D267-70. [PubMed http://dx.doi.org/10.1093/nar/gkh061](http://dx.doi.org/10.1093/nar/gkh061)
2. Humphreys BL, Lindberg DA, Schoolman HM, Barnett GO. 1998. The Unified Medical Language System: an informatics research collaboration. *J Am Med Inform Assoc.* 5(1), 1-11. [PubMed http://dx.doi.org/10.1136/jamia.1998.0050001](http://dx.doi.org/10.1136/jamia.1998.0050001)
3. Schuyler PL, Hole WT, Tuttle MS, Sherertz DD. 1993. The UMLS Metathesaurus: representing different views of biomedical concepts. *Bull Med Libr Assoc.* 81(2), 217-22. [PubMed http://dx.doi.org/10.1136/bmla.1993.081022](http://dx.doi.org/10.1136/bmla.1993.081022)
4. Tuttle MS, Sherertz DD, Olson NE, Erlbaum MS, Sperzel WD, et al. Using META-1, the first version of the UMLS Metathesaurus. *Proc 14th Annu Symp Comput Appl Med Care.* 1990. p. 131-5.
5. McCray AT, Nelson SJ. 1995. The representation of meaning in the UMLS. *Methods Inf Med.* 34(1-2), 193-201. [PubMed http://dx.doi.org/10.1136/infmed.1995.034012193](http://dx.doi.org/10.1136/infmed.1995.034012193)
6. McCray AT. 2003. An upper-level ontology for the biomedical domain. *Comp Funct Genomics.* 4(1), 80-84. [PubMed http://dx.doi.org/10.1002/cfg.255](http://dx.doi.org/10.1002/cfg.255)
7. McCray AT, Hole WT. The scope and structure of the first version of the UMLS Semantic Network. *Proc 14th Annu Symp Comput Appl Med Care*; Los Alamitos, CA. 1990. p. 126-30.
8. Gu H, Perl Y, Geller J, Halper M, Liu LM, et al. 2000. Representing the UMLS as an object-oriented database: modeling issues and advantages. *J Am Med Inform Assoc.* 7(1), 66-80. [PubMed http://dx.doi.org/10.1136/jamia.2000.0070066](http://dx.doi.org/10.1136/jamia.2000.0070066)
9. Geller J, Gu HH, Perl Y, Halper M. 2003. Semantic refinement and error correction in large terminological knowledge bases. *Data Knowl Eng.* 45(1), 1-32. [http://dx.doi.org/10.1016/S0169-023X\(02\)00153-2](http://dx.doi.org/10.1016/S0169-023X(02)00153-2)
10. Gu H, Perl Y, Elhanan G, Min H, Zhang L, et al. 2004. Auditing concept categorizations in the UMLS. *Artif Intell Med.* 31(1), 29-44. [PubMed http://dx.doi.org/10.1016/j.artmed.2004.02.002](http://dx.doi.org/10.1016/j.artmed.2004.02.002)

11. Gu HH, Hripcsak G, Chen Y, Morrey CP, Elhanan G, et al. 2007. Evaluation of a UMLS Auditing Process of Semantic Type Assignments. *AMIA Annu Symp Proc.* 2007, 294-98. [PubMed](#)
12. Morrey CP, Geller J, Halper M, Perl Y. 2009. The Neighborhood Auditing Tool: a hybrid interface for auditing the UMLS. *J Biomed Inform.* 42(3), 468-89. [PubMed](#) <http://dx.doi.org/10.1016/j.jbi.2009.01.006>
13. Gu HH, Elhanan G, Perl Y, Hripcsak G, Cimino JJ, et al. 2012. A study of terminology auditors' performance for UMLS semantic type assignments. *J Biomed Inform.* 45(6), 1042-48. [PubMed](#) <http://dx.doi.org/10.1016/j.jbi.2012.05.006>
14. Chen Y, Gu HH, Perl Y, Geller J, Halper M. 2009. Structural group auditing of a UMLS semantic type's extent. *J Biomed Inform.* 42(1), 41-52. [PubMed](#) <http://dx.doi.org/10.1016/j.jbi.2008.06.001>
15. Chen Y, Gu HH, Perl Y, Halper M, Xu J. 2009. Expanding the extent of a UMLS semantic type via group neighborhood auditing. *J Am Med Inform Assoc.* 16(5), 746-57. [PubMed](#) <http://dx.doi.org/10.1197/jamia.M2951>
16. Chen L, Morrey CP, Gu H, Halper M, Perl Y. 2009. Modeling multi-typed structurally viewed chemicals with the UMLS Refined Semantic Network. *J Am Med Inform Assoc.* 16(1), 116-31. [PubMed](#) <http://dx.doi.org/10.1197/jamia.M2604>
17. Wei WQ, Leibson CL, Ransom JE, Kho AN, Caraballo PJ, et al. 2012. Impact of data fragmentation across healthcare centers on the accuracy of a high-throughput clinical phenotyping algorithm for specifying subjects with type 2 diabetes mellitus. *J Am Med Inform Assoc.* 19(2), 219-24. [PubMed](#) <http://dx.doi.org/10.1136/amiajnl-2011-000597>
18. Geller J, He Z, Perl Y, Morrey CP, Xu J. 2013. Rule-based support system for multiple UMLS semantic type assignments. *J Biomed Inform.* 46(1), 97-110. [PubMed](#) <http://dx.doi.org/10.1016/j.jbi.2012.09.007>
19. IHTSDO. SNOMED CT Homepage [April 2014]. Available from: <http://www.ihtsdo.org>.
20. Forrey AW, McDonald CJ, DeMoor G, Huff SM, Leavelle D, et al. 1996. Logical observation identifier names and codes (LOINC) database: a public use set of codes and names for electronic reporting of clinical laboratory test results. *Clin Chem.* 42(1), 81-90. [PubMed](#)
21. Liu S, Ma W, Moore R, Ganesan V, Nelson S. 2005. RxNorm: Prescription for Electronic Drug Information Exchange. *IT Prof.* 7(5), 17-23. <http://dx.doi.org/10.1109/MITP.2005.122>
22. Henning K. Overview of Syndromic Surveillance: What is Syndromic Surveillance? 2004 [April 2014]. Available from: <http://www.cdc.gov/mmwr/preview/mmwrhtml/su5301a3.htm>.
23. Hripcsak G, Soulakis ND, Li L, Morrison FP, Lai AM, et al. 2009. Syndromic surveillance using ambulatory electronic health records. *J Am Med Inform Assoc.* 16(3), 354-61. [PubMed](#) <http://dx.doi.org/10.1197/jamia.M2922>

24. Peng Y, Halper MH, Perl Y, Geller J. 2002. Auditing the UMLS for redundant classifications. *Proc AMIA Symp.* 2002, 612-16. [PubMed](#)
25. Srinivasan S. Personal Communication. 2009.
26. Chen Y, Gu H, Perl Y, Geller J. 2012. Overcoming an obstacle in expanding a UMLS semantic type extent. *J Biomed Inform.* 45(1), 61-70. [PubMed](#) <http://dx.doi.org/10.1016/j.jbi.2011.08.021>
27. Halper M, Wang Y, Min H, Chen Y, Hripcsak G, et al. 2007. Analysis of error concentrations in SNOMED. *AMIA Annu Symp Proc.* 2007, 314-18. [PubMed](#)
28. Wang Y, Halper M, Wei D, Perl Y, Geller J. 2012. Abstraction of complex concepts with a refined partial-area taxonomy of SNOMED. *J Biomed Inform.* 45(1), 15-29. [PubMed](#) <http://dx.doi.org/10.1016/j.jbi.2011.08.013>
29. Wang Y, Halper M, Wei D, Gu H, Perl Y, et al. 2012. Auditing complex concepts of SNOMED using a refined hierarchical abstraction network. *J Biomed Inform.* 45(1), 1-14. [PubMed](#) <http://dx.doi.org/10.1016/j.jbi.2011.08.016>
30. Wang Y, Wei D, Xu J, Elhanan G, Perl Y, et al. 2008. Auditing complex concepts in overlapping subsets of SNOMED. *AMIA Annu Symp Proc.* 2008, 273-77. [PubMed](#)
31. Min H, Perl Y, Chen Y, Halper M, Geller J, et al. 2006. Auditing as part of the terminology design life cycle. *J Am Med Inform Assoc.* 13(6), 676-90. [PubMed](#) <http://dx.doi.org/10.1197/jamia.M2036>
32. NLM. The Future of the UMLS Semantic Network 2005 [Oct 8, 2013]. Available from: <http://mor.nlm.nih.gov/snw/>.
33. Definition of UMLS Semantic Types [cited 2012 Dec 5]. Available from: <http://semanticnetwork.nlm.nih.gov/Download/RelationalFiles/SRDEF>.
34. Simborg D, Sparks S, Buxton R, Klein J, Van Valkenburgh T, Quinn J, Campbell S, Carney D, Pine S, Glickman M, et al. HL7: the promise of tomorrow. Interview by Bill W. Childs. *US Healthc.* 1989;6(8):26, 8, 32.
35. Alcohol and Other Drug Thesaurus Homepage [April 2014]. Available from: <http://etoh.niaaa.nih.gov/AODV01/aodthome.htm>.
36. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet.* 25(1), 25-29. [PubMed](#) <http://dx.doi.org/10.1038/75556>
37. Lomax J, McCray AT. 2004. Mapping the Gene Ontology into the Unified Medical Language System. *Comp Funct Genomics.* 5(4), 354-61. [PubMed](#) <http://dx.doi.org/10.1002/cfg.407>
38. UMLS Reference Manual [cited 2013 March 4]. Available from: <http://www.ncbi.nlm.nih.gov/books/NBK9676/>.
39. Yu H, Friedman C, Rhzetsky A, Kra P. 1999. Representing genomic knowledge in the UMLS semantic network. *Proc AMIA Symp.* 1999, 181-85. [PubMed](#)

40. Cohen B, Chen Y, Perl Y. 2007. Updating the genomic component of the UMLS Semantic Network. *AMIA Annu Symp Proc.* 2007, 150-54. [PubMed](#)
41. He Z, Ochs C, Agrawal A, Perl Y, Zeginis D, Tarabanis K, Elhanan G, Halper M, Noy N, Geller J. A Family-Based Framework for Supporting Quality Assurance of Biomedical Ontologies in BioPortal. *AMIA Annu Symp Proc.* 2013, 581-90. [PubMed](#)

¹ Concepts are denoted by italics.

² Semantic types are denoted by a bold font.