Published in final edited form as:

J Am Soc Inf Sci Technol. 2013 October ; 64(10): 1963–1974. doi:10.1002/asi.22899.

Extending SemRep to the Public Health Domain

Graciela Rosemblat, Ph.D, Melissa P. Resnick, M.S., M.L.S, Ione Auston, M.L.S, Dongwook Shin, Ph.D, Charles Sneiderman, M.D., Ph.D, Marcelo Fiszman, M.D., Ph.D, and Thomas C. Rindflesch, Ph.D.

National Library of Medicine, NIH, Bethesda, MD

Abstract

We describe the use of a domain-independent methodology to extend a natural language processing (NLP) application, SemRep (Rindflesch, Fiszman, & Libbus, 2005), based on the knowledge sources afforded by the Unified Medical Language System (UMLS®) (Humphreys, Lindberg, Schoolman, & Barnett, 1998) to support the area of health promotion within the public health domain. Public health professionals require good information about successful health promotion policies and programs that might be considered for application within their own communities. Our effort seeks to improve access to relevant information for the public health profession, to help those in the field remain an information-savvy workforce. NLP and semantic techniques hold promise to help public health professionals navigate the growing ocean of information by organizing and structuring this knowledge into a focused public health framework paired with a user-friendly visualization application as a way to summarize results of PubMed searches in this field of knowledge.

Keywords

Health Promotion; Natural Language Processing; Semantic Predications; Public Health; Semi-Automated Ontology Development

INTRODUCTION

The need for improved access to high quality public health information has been stated by public health professionals for well over a decade. Further, recent infectious disease outbreaks such as severe acute respiratory syndrome (SARS) and Asian bird flu, along with growing concern about acts of bioterrorism, such as the 2001 anthrax episode in the US Postal Service, have brought unprecedented urgency to the information needs of the public health workforce (LaPelle, Luckmann, Hatheway Simpson, & Martin, 2006). As stated in that source, this information falls into various distinct categories, from newly identified health risks to reports on preventive health behaviors and new interventions known to be effective. Indeed, it has been estimated that as much as 50% of the health status of the American public is attributable to health behaviors and related factors (Mokdad, Marks, Stroup, & Gerberding, 2004). Public health practitioners share with their counterparts in

clinical medicine the task of helping the public be more aware of achieving and maintaining healthy behaviors—the former in more population-based approaches while the latter at the individual patient level. The substance of public health practice involves “organized community effort to address the public interest in health...to prevent disease and promote health” (at http://www.nap.edu/openbook.php?record_id=1091&page=7, IOM, 1998). At its core is program development which entails broad community involvement, promotion of a scientific basis for decision making, strategic alliances, and development of comprehensive public health policies to promote and encourage healthy behaviors. Health promotion at the community level takes many forms—from social marketing and targeted communication through various media (radio, television, internet), to joint health education efforts with schools, churches, work sites, and the like, to collaboration with health care providers on health messages and programs, to advocacy for and implementation of policies and legislation that create healthy environments. Public health professionals require relevant information about successful health promotion policies and programs but often lack the time and skills required to effectively search for well designed, evaluated studies and practices that might be considered for application within their own communities. Natural language processing and semantic techniques can facilitate these tasks by organizing and structuring this “knowledge” into a coherent conceptual public health framework. The current project aims to expand existing National Library of Medicine [NLM] semantic resources to facilitate advanced access to online documents on Health Promotion.

This paper describes the methodology implemented to extend a Unified Medical Language System (UMLS®)-based natural language processing application, SemRep (Rindflesch, Fiszman, & Libbus, 2005), to the previously unsupported domain of health promotion, within the broader context of public health. This approach leverages existing UMLS knowledge and integrates a newly defined semantic space with the semantic structure afforded by the UMLS (Humphreys, Lindberg, Schoolman, & Barnett, 1998).

Background

SemRep—SemRep (Rindflesch et al., 2005) is an NLP advanced information management application. It enhances document retrieval with the support of the UMLS knowledge sources (biomedical field) as the domain ontology. SemRep extracts relationships from biomedical text by mapping textual content to an ontology which represents the meaning of the text. SemRep extractions, called semantic predications, consist of assertions derived from sentences in PubMed titles and abstracts. A semantic predication contains three parts: a subject, an object, and the relation that links them. For example, from a sentence such as (1) in which the underlined expressions stand for the subject and object, respectively, SemRep derives the predication in (2):

- (1) Sertraline in the *treatment* of PTSD (Subject and object underlined)
- (2) Sertraline TREATS Stress Disorders, Post-Traumatic (Predication)

To form (2), SemRep extracts all three sentential components in (1) (the subject and object, and their binding relation), and uses MetaMap (Aronson, 2001) to map the subject and object arguments to concepts in the UMLS Metathesaurus. To establish their binding relation, SemRep relies on internal rules (called “indicator rules”) which map syntactic

elements--such as verbs, prepositions, and nominalizations--to predicates in the Semantic Network (e.g. TREATS, PREVENTS, CAUSES). For the example in (1), the indicator rule needed is given in (3):

(3) *treatment* (noun) → TREATS (Indicator Rule)

The UMLS Metathesaurus is composed of many biomedical vocabularies—such as Medical Subject Headings (MeSH®), Logical Observation Identifiers Names Codes (LOINC®), Systemized Nomenclature of Medicine Clinical Terms (SNOMED CT®)—and the terms it contains are grouped by meaning. Those terms that share the same meaning are classed under the same *concept*. For example, expressions such as “Post-traumatic Stress Disorder” and “PTSD” are classed under the UMLS concept ‘Stress Disorders, Post-Traumatic.’ Each Metathesaurus concept is assigned a semantic type such as “Disease or Syndrome” or “Gene or Genome,” which allows semantically similar concepts to be clustered under the same semantic type. For instance, the concept “Sertraline” in sentence (1) and its synonym “Zoloft” have the semantic type “Pharmacologic Substance,” while “PTSD” has the semantic type “Disease or Syndrome.” The predication in (2) presented earlier derives from the more abstract Semantic Network relation in (4), in which the semantic types take the place of the argument concepts:

(4) Pharmacologic Substance TREATS Disease or Syndrome (Semantic Types underlined)

Since the semantic types in (4) match the semantic types of the subject (*Sertraline*) and object (*Stress Disorders, Post Traumatic*) in (1) respectively, the semantic predication (2) is created, in which the corresponding Metathesaurus concepts are substituted for their respective semantic types ‘Pharmacologic Substance’ and ‘Disease or Syndrome.’ These semantic types are allowed as arguments of the TREATS relation, as specified in the UMLS Semantic Network, which stipulates co-occurrence restrictions for arguments of different predicates based on the semantic types of the arguments. For an indicator and its syntactic arguments to be interpreted as a semantic predication, the semantic types of the Metathesaurus concepts for the noun phrase arguments must match the semantic types allowed to serve as arguments of the given predicate in the Semantic Network. For example, for TREATS, the semantic type arguments shown in (4) are allowed.

Unlike the UMLS concept mapping process, information about how the subject-object binding relations are expressed in the text (e.g., *treatment* in (1)) and the semantic network relations to which the expressions in the text map (e.g., TREATS in (2)) is not in the UMLS, and discovering these mappings is done manually. Thus, bridging the gap between the textual expressions and the UMLS Semantic Network relations involves a manual process. For instance, in addition to ‘treatment’ as in (1), terms such as ‘management,’ ‘ameliorates,’ ‘heals,’ and ‘relief’ can also mean TREATS when used with the appropriate arguments. Extending SemRep to cover health promotion concepts and predicates not in the UMLS, or used with a different meaning than the ones intended for the medical domain, also entails discovering the links or relationships between domain concepts and the syntactic expressions that point to these relationships.

Semantic MEDLINE—Semantic MEDLINE (Rindflesch, Kilicoglu, Fiszman, Rosemblat, & Shin, 2011; Kilicoglu, Fiszman, Rodriguez, Shin, Ripple, & Rindflesch, 2008) is a National Library of Medicine (NLM) advanced information management application that allows the visualization of SemRep results. This tool helps manage results from search engines such as PubMed® by manipulating information in addition to documents. The information retrieved is summarized and presented in a graph as a network of concepts and relations extracted by SemRep. Figure 1 displays the network of concepts and relations derived from a PubMed search on ‘Post-traumatic Stress Disorder.’ The concept searched for is at the center in the lower part of the graph. The colors of the boxes represent the semantic groups to which the concepts belong. In Figure 1, pink boxes represent Disorders, grey boxes correspond to Therapeutic Procedures, yellow boxes correspond to Substances, and blue boxes, to Populations. While boxes represent concepts, color-coded directional arrows (color key on the right) represent the labeled relation that bind two concepts. By looking at the colored arrows (the relation) and the concepts they link, we can interpret entities such as “Cognitive therapy TREATS Stress Disorders, Post-Traumatic,” “Inflammation Process AFFECTS Stress Disorders, Post-Traumatic.”

Extending the Semantic Representation—SemRep was originally developed to extract semantic predications from the clinical literature and it has already been extended to the fields of Molecular Genetics (Rindflesch, Libbus, Hristovski, Aronson, & Kilicoglu, 2003) research, Pharmacogenomics (Ahlers, Fiszman, Demner-Fushman, Lang, & Rindflesch, 2007), and Disaster Information Management (Keselman, Rosemblat, Kilicoglu, Fiszman, Jin, Shin, & Rindflesch, 2010). As the UMLS—SemRep’s knowledge source—focuses on the medical domain, coverage for other disciplines is often incomplete or inadequate. Thus for domains other than biomedicine, the UMLS needs to be augmented with domain-specific concepts and semantic types, which requires ontological and terminological extensions to the system for generation of domain-relevant semantic predications.

There is a principled distinction in SemRep between general aspects and those specific to a particular domain. The underlying syntactic processing and argument identification are general, while indicator rules are domain specific, to a large extent. Many concepts and relations important in other fields may already exist in the UMLS (clinical domain) and are thus already incorporated into its ontological sources (MeSH). SemRep takes advantage of the already existing knowledge, and supplements it with domain-specific concepts and relations needed for interpretation, and which are either missing from the UMLS or inappropriately represented for the new domain. Due to this SemRep characteristic, only domain-specific ontological aspects need to be enhanced when extending coverage to new disciplines. New domain-specific entities are formalized and written in the same UMLS format so they can be seamlessly incorporated into the UMLS existing machinery. New concepts are allowed to co-occur both with other new domain concepts and others currently existing in the UMLS, to be used with existing and new relations, as illustrated in examples (6) to (12) in the Methods section.

One way to extend SemRep coverage beyond biomedicine to support predications in other domains, such as Public Health, is by developing a semantic representation underpinned by

a domain ontology. NLP-based systems often depend on ontologies to support a semantic representation. For example, in a newly developed semantic representation for automatic information extraction from dental clinical records (Irwin, Harkema, Christensen, Schleyer, Haug, & Chapman, 2009), the contents of semantic representations were found to be excessively document-bound and application-specific to be useful in modeling domain-relevant concepts and relations. While results in domain ontologies rarely fit well with the information in actual text, building a new ontology for each specific application undermines two important characteristics of ontologies, namely, that they facilitate both knowledge sharing and data reuse (Bodenreider, 2008), thus laying the groundwork for future semantic integration (Noy, 2009). These characteristics of ontologies play a fundamental role in supporting the extension of SemRep coverage to a new domain. Rather than creating a new ontology for a new domain, we profit from UMLS knowledge that is common to both domains, focusing only on the development of the entities needed for domain-specific knowledge representation and which are not fully or adequately represented in the UMLS. Thus we take advantage of the distinction between the general and domain-specific aspects in SemRep. This affords us considerable savings in time and labor by enabling the reuse of existing knowledge, and facilitates the semantic integration (Noy, 2009) of the new knowledge with the UMLS content, as described in the Methods section.

Literature Review—Initiatives such as the Healthy People reports from the US Department of Health and Human Services provide a framework for identifying the most significant preventable threats to health and establish national goals to reduce them. These initiatives have prompted a strong move to empower people to take charge of their own health through disease prevention and healthier lifestyles, which represent the essence of Public Health. Undoubtedly, medical findings have important implications to public health and so do other factors that influence behavior choices. A major effort for improving access to relevant information for the Public Health workforce has been the Partners in Information Access for the Public Health Workforce initiative (<http://phpartners.org/hp2020/index.html#whatispartners>, retrieved December 2011), a multi-agency collaboration to provide public health professionals with timely, convenient access to information resources to help them improve the health of the American public. The most heavily used section of the Partners website has been one specific component of this initiative, the Healthy People 2020 Structured Evidence Queries (<http://phpartners.org/hp2020/index.html#whatisiap>, retrieved December 2011). It consists of pre-formulated PubMed search strategies developed and designed by librarians working with subject matter experts to assist the public health workforce in achieving Healthy People 2020 goals (<http://healthypeople.gov/2020/Consortium/HP2020Framework.pdf>) and objectives (<http://healthypeople.gov/2020/topicsobjectives2020/pdfs/HP2020objectives.pdf>) by helping them quickly and easily find relevant information from NLM's MEDLINE database.

Access to manageable relevant information of the type just described could be improved by an architecture of information extraction that allows the organization of the information retrieved to be mapped against an ontological knowledge-based system. There is wide consensus on the importance of building domain models or ontologies in developing knowledge based systems, as ontologies provide a structural framework for natural language

processing (NLP) applications. However, to be useful for purposes or domains different from the ones they were originally designed for, the ontology coverage often needs to be extended. Extension of existing ontologies is highly recommended, as explained in Bodenreider (2008) and Noy (2009) over creating separate ontologies as in (Irwin et al., 2009). Various other projects have extended ontologies and terminologies to meet different needs or for other applications. In the biomedical domain, for example, an expansion of terms relevant to muscle biology was incorporated into the Gene Ontology (GO Ontology) (Feltrin, Campanaro, Diehl, Ehler, Faulkner, Fordham,...Deegan, 2009); Chen, Morrey, Gu, Halper, and Perl (2009) developed a refined UMLS Semantic Network to distinguish between chemical complex types and conjugates in the UMLS.

Our literature search on public health ontologies retrieved a few thesauri, such as the European Multilingual Thesaurus on Health Promotion (<http://ring.ciard.net/kos/european-multilingual-thesaurus-health-promotion>, retrieved March 2012) and the World Health Organization Health Promotion Glossary (<http://www.who.int/healthpromotion/about/HPG/en/>). Several other works retrieved dealt with various specific aspects of an ontology, unrelated to health promotion (Derose, Schuster, Fielding, and Asch, 2002; Kostova, Kumar, Roy, Weinberg, Madle, and Carson, 2005; OntoReason [OTR] Public Health Ontology, 2007, in http://protege.stanford.edu/conference/2007/presentations/02.01_Lynch.pdf, retrieved Nov 2011). Some of these works often mapped to formal existing terminologies, with the main focus on biosurveillance and alert systems (Mirhaji, 2009; Srinivasan, Kunapareddy, Mirhaji, & Casscells, 2006; Lynch, Cunningham, Schripsema, Morris, & Rhodes, 2007). For example, Mirhaji (2009) uses a few UMLS Metathesaurus concepts and semantic types to model LOINC axes as a formal ontology to represent laboratory tests for automated classification and decision support. Biosense (Lynch et al., 2007) uses the OTR Ontology for automatic case classification of incoming HL7 messages, analyzed by content “to determine syndromic classification and routing.” Its ontological knowledge representation is based on HL7 complex data types, while concept representation is based on SNOMED (Stearns, Price, Spackman, and Wang, 2001), to which syndromes in the International Classification of Diseases –9th Revision– Clinical Modification (ICD-9-CM) were mapped. The ontological concepts support pattern matching and differential diagnosis with an expert system implementation platform.

The preceding systems describe concept mapping for knowledge representation but do not aim to establish how these concepts are related beyond a taxonomic relationship (Srinivasan et al., 2006). The focused ontologies that support the cited models do not include health promotion. Further, a model with the coverage and internal consistency of SNOMED CT in medicine (Stearns et al., 2001), for example, is not available for Public Health.

Our work differs from previous endeavors in this field in that it implements the mapping of the new knowledge (represented by both concepts and relations) in a previously unsupported domain to the UMLS knowledge sources. This can be done because we implement the methods in a flexible way within the framework of the existing UMLS architecture. Concepts in higher-level ontological knowledge sources rarely achieve the desired level of specificity required for a (sub)domain, and when they do, mappings often require adaptations that render this task too labor-intensive (Irwin et al., 2009; Kuziemsky & Lau,

2010, among others). This work is unique in that it achieves these mappings with seamless integration with the existing machinery, facilitating efficiency and reuse of ontological entities and allowing us to benefit from the synonymy information in the UMLS (Humphreys et al., 1998).

Applying SemRep to health promotion requires modifications to the UMLS on which the application relies. There are various methods for building or extending domain ontologies, as reviewed by many works (e.g., Buitelaar, Cimiano, & Magnini 2005). While we use a similar bottom-up approach as in Buitelaar et al. (2005), we differ in that we integrate the new knowledge into a higher-level knowledge source (UMLS), either mapping to it or framing the new entities in accordance with its constructs (e.g., semantic types). Our focus is on ontological and terminological supplementation to the Metathesaurus and the Semantic Network for domain-relevant coverage, enriching our application to recognize domain-specific classes of concepts while relying on the semantic structure of the UMLS. Hierarchical relations as defined in the UMLS semantic types and predicates only apply to UMLS semantic types and thus do not extend to new domain-specific semantic types and relations. However, many UMLS relations that SemRep supports are nonhierarchical (e.g., COMPLICATES, DIAGNOSES, among others). The Metathesaurus requires augmentation with new concepts and semantic types, and the Semantic Network requires augmentation with new domain-specific relations. This project benefited from an automated program developed in-house designed to facilitate the identification of core domain concepts not in the UMLS. In the sections that follow, we describe how SemRep has been enriched to recognize the new classes of concepts that play a role in this domain as well as new domain-relevant semantic relations, such as ADDRESSES, HINDERS, and PROVIDES, to name a few. The resulting network of concepts and relations is supported by a semantic representation which requires the definition of the ontological space of Health Promotion. The core domain aspects will be articulated below.

Ontological considerations—In defining the ontological space of health promotion, the key players and events must be described. From a public health perspective, disease prevention and health promotion deal with desirable changes of a certain magnitude (e.g., healthier lifestyles) in a community, and the most effective way to carry them out. For example, demographics and physical and social environments can be used as channels or venues to foster healthier behavioral changes; organizational settings such as community clinics or hospitals serve populations by targeting certain problems or conditions, and the programs or policies that address them can be specifically designed for the target population. By monitoring undesirable or risky health behaviors and incidence of disease and mortality rates which define public health issues in different populations, policies and programs to address them can be developed and implemented to help mitigate these health issues before they become a public health problem. Recognizing that differing demographic characteristics can trigger disparities in access to--or delivery of--health care, the policies and programs aim at eliminating barriers to accessibility to healthcare as part of a sustained effort towards the elimination of health disparities, defined as the difference in “*health status between a defined portion of the population and the majority*” in Carter-Pokras & Baoljet (2002). Evaluation instruments measure the results and outcomes of the programs

and policies implemented in order to determine their effectiveness. This process leads to newer need-driven goals, thus starting the cycle anew.

Semantic Modeling—From the description in the preceding section, several components come to the forefront: the target population categorized into various groups, the behavior or condition to be addressed, the development of policies or methods to address it, assuring their proper implementation and effectiveness (measuring outcomes), in compliance with health laws and ethical principles (quality control), and the elimination of health disparities. Programs may be described in terms of type, venue, implementer, attributes, goal(s), and target population. Timely implementation of policy is essential to prevent any health issues from becoming public health issues. Policy development involves governmental agencies (local, state, federal) providing information, guidance, and funding. Relevant aspects of the targeted population and the targeted behavior/conditions (and degree of threat they pose) will guide the design and methods to be implemented. Elimination of health disparities focuses on raising awareness about cultural sensitivities, system of values and beliefs in a given population, and linguistic/minority appropriateness of methods and implementers. Measuring outcomes and program policy effectiveness may in turn give rise to new policies or programs.

Developing a Semantic Representation—Since health promotion is not well represented in the UMLS, developing a semantic representation starts with the development of a UMLS-compatible domain ontology to enhance the UMLS semantic space and bridge this gap. We are guided by compatibility with our own SemRep application and importantly, semantic analysis of domain texts (mentioned in Resnick, Ripple, Auston, & Rindflesch, 2010). The full approach will be described in the Methods section.

The process starts with a thorough analysis of texts representative of health promotion, provided by a domain specialist. This allows the identification of the core concepts and major topics, and facilitates the identification of concepts and relations in the sentences. Additionally, texts are also processed through SemRep, which partitions the abstracts into sentences and generates the baseline predications. The corpus collected is also processed through a program we developed that attempts to map noun phrases to already existing Metathesaurus concepts. This program separates the UMLS concept-mapping noun phrases from the others. The output files are used for preliminary assessment of UMLS coverage in the new domain. Domain-relevant concepts and relations not included in the UMLS are added to supplementary files, so new, more domain-appropriate semantic types of concepts can be defined. Those already in the UMLS are redefined if needed. Finally, the semantic relations that hold between the concepts are semantically defined and labeled.

METHODS

Approach

Our approach draws upon the generic ontology engineering process described in Kuziemy and Lau (2010), who used an ontology to support queries as part of a clinical decision support system. Our bottom-up approach shares elements of the ontology-learning layer cake described in Buitelaar et al., (2005) and the process in Kuziemy and Lau (2010),

working up from terms, to synonyms, concepts, and relations, to formalize rules at the top. However, as our system maps to a higher-level ontology, we adapted the process to accommodate the use of the UMLS.

Document Collection—The third author of this manuscript, (IA), a domain specialist, processed three of the Healthy People 2020 Structured Evidence Queries with a focus on three health promotion topics: smoking cessation, asthma, and osteoporosis. Each of those topics was included in the PubMed search that resulted in a training set of almost 800 PubMed citations with the targeted representation and scope.

Training Set—From the citations retrieved, 775 sentences were extracted for later categorization. An illustration of the complex content-rich sentences we analyzed is given in Table I:

Main Themes—To identify the main domain themes for this work, we manually classified the set of 775 sentences into subcategories, guided by topic similarity. The main categories were not pre-defined but stemmed from the results of our manual clustering. Most sentences contained elements that belonged to more than one theme, fitting more than one category. We identified three major themes or categories and, within the larger context of essential public health services, two publications (Public Health Functions Steering Committee, “Public Health in America” (1994) and the Institute of Medicine 1988 report) provided post-analysis external validation about their correctness. The three categories are:

Assessment has to do with identifying and monitoring populations at-risk to identify health problems.

Policy Development deals with the design and implementation of programs and policies to address the health problems identified.

Assurance deals with ensuring that the programs and policies are effective and appropriately implemented.

Defining the Semantic Space—The Assessment and Policy Development sentences in Table II are representative of their respective themes, while the Assurance sentence suitably refers to the results of the program implemented—program benefits and efficacy—but also displays elements from the other two themes. This is distinctive of sentences in this category, as the expression of effectiveness, results, or impact often includes the entities to which they apply, to complete the meaning of the sentence. Differentiating entities as per the categories is a clear-cut task, but one sentence may contain parts of different categories, with different entities in the sentence belonging to different themes. The sentence is the unit that holds all those entities, which renders the delineation in terms of categories not so clear-cut at the sentence level, where the various entities interact. When classifying sentences, the same sentential unit may be duplicated in different areas. This is especially the case with the Assurance category sentences, as in the process of conceptualizing where the various actors fit and how they interact, it became apparent that the Assurance category ties together all other categories and subcategories in this domain, but with clear elements that set it apart from the other two. For example, differences in demographic factors or population groups

(Assessment category) often lead to health disparities, and to fulfill the health promotion goal of eliminating disparities, programs and interventions are planned, designed, and implemented (Policy Development). Measures are then taken to assure that all the conditions are in place for the proper implementation of the programs as designed with the expected outcomes (Assurance category). These interrelationships also account for the many more sentences in this category (N=474) with respect to the others. Building a model which can point out potential correlations between “Assurance” statements and “Assessment” (such as in demographic factors or health issues) would provide valuable insights, but determining the extent to which greater refinement of the relationship between “Assurance” and the other categories might be helpful is beyond the scope of this work.

Defining the Conceptual Space—The process of sorting the sentences into subcategories was facilitated by the analysis of the conceptual space (conceptualization) as detailed earlier. The different subcategories linked to the core categories represent the ontology scope.

Identification of concepts and relations: Semantic analysis forms the foundation of the semantic representation of the domain and leads to the discovery and labeling of the new concepts and their interactions. Sentence analysis revealed terms that describe the different subcategories and helped identify core concepts and the actions associated with them. Important concepts in the ASSURANCE subcategory *Eliminating Disparities* (Table III), for example, have to do with barriers such as cultural beliefs, linguistic factors, access to healthcare, among others. Many core concepts and relations refer to ways to reduce or eliminate lack of competence, such as raising awareness or imparting better communication skills to foster more competent care. Combined with the output of our automated in-house program mentioned earlier, this analysis allowed us to earmark those core concepts and relations not in the UMLS, which led to the addition of new concepts and relations, and the contextualization or redefinition of existing ones, for appropriateness in this new domain.

Concepts added: A total of 194 core domain concepts not in the UMLS Metathesaurus were identified and added to supplementary files, along with their synonyms. Of those, 138 concepts were added with an already existing semantic type, as illustrated in Table IV, while 56 concepts were added with new domain-specific semantic types (Table V), assigned to these new concepts based on linguistic and contextual analysis.

Concepts redefined: UMLS semantic types that are too vague and lack specificity, such as Functional Concept, Intellectual Product, or Idea or Concept, are not used by SemRep’s semantic predications. Although assigning these semantic types works well for an upper level knowledge source, they group together a wide-range of concepts of very dissimilar nature. For UMLS concepts that appeared to be relevant in the domain texts and displayed those semantic types, the UMLS-assigned semantic type was changed to capture more adequately the meaning as used in this domain (Table VI). A total of 130 UMLS-mapping concepts were contextualized this way. Missing synonyms for existing concepts were also added if needed (Table VII). For example, within the context of smoking cessation programs, an important public health initiative, the expression ‘quitter’ was added as a

synonym of UMLS concept ‘former smoker’ because it frequently appeared in our training set with that meaning. While ‘quitter’ can have other negative connotations, within the focused context of public health/smoking cessation, the mapping to ‘former smoker’ is quite unequivocal.

Relations added: New domain-specific relations were created if needed, to capture more precisely the semantic nature of the interactions in this domain. This process led to semantically appropriate predications, formally extending SemRep coverage. In this project, fifteen new domain relations were identified, the most productive ones being ADDRESSES, ANALYZES, EVALUATES, HAS_ACCESS_TO, IMPEDIMENT_TO, MEASURES, PROMOTES, PROVIDES, SERVES, and TARGETS. To allow for a richer representation, we made a distinction between the relations TARGETS and ADDRESSES, so that in the case of the former, the object would be the audience or population group that a given program or intervention targets (*target population*), while in the case of ADDRESSES, the object would be the condition or health behavior that the program or intervention aims to alleviate.

Relations redefined: The process of supplementing current UMLS relations involved more than identifying the new relations to be added. We also needed to redefine allowable arguments (with existing semantic types) for existing UMLS relations, in ways not initially specified in the Semantic Network, as in (5), and integrate newly created semantic types with existing UMLS relations, as shown in (6):

(5) *Racial Discrimination* [Social Behavior] CAUSES *Disparities* [Finding]

(6) *Project* [Program Type] PREVENTS *Tobacco Use* [Finding]

Lastly, we also needed to integrate new and existing UMLS semantic types with the new relations, as in (7–8):

(7) *Intervention Program* [Program Type] PROMOTES *Healthy weight* [Healthy Behavior]

(8) *Educational Program* [Educational Activity] TARGETS *Elderly* [Age Group]

RESULTS

After the addition of the new and redefined concepts, semantic types, and relations, and given the sentence in (9), SemRep can now extract a predication such as (10):

(9) ...community-based projects should focus on healthier lifestyles

(10) Project TARGETS Healthy lifestyle

‘Project’ in (10) is a newly added concept, with the new semantic type ‘Program Type.’ TARGETS is a new relation to which the lexical expression ‘*focus on*’ points, and ‘Healthy lifestyle’ is a new concept with the new semantic type ‘Healthy Behavior.’ Again, the new concepts and relations follow the UMLS-based Metathesaurus and Semantic Network formats. Concepts participate in different relations, and determining which concepts

participate in which relations (both newly created and current) results in a semantic representation of the domain.

Figure 2 shows a partial model of the resulting semantic representation after implementing the changes described earlier. Color-coded arrows (key on the right) indicate the labeled relations. In this figure, red and pink arrows (TREATS, PREVENTS) represent current UMLS predicates (relations); arrows in various shades of green represent newly added relations (TARGETS, VENUE_FOR, PROMOTES, ...), as in the predication ‘Interventions PROMOTES Health promotion,’ in which both arguments have the semantic type Healthcare Activity. While arrows correspond to relations, boxes represent concepts. All concepts with the same semantic type are shaded with dark/light shades of the same color. Lighter color boxes contain current UMLS concepts with the originally assigned semantic type, as in ‘Hispanics’ (Population Group) and ‘Low Back Pain’ (Disease or Syndrome). Darker boxes with italics represent new concepts added, often with new semantic types (e.g., Church coalition, Clinical Setting). Lastly, darker boxes not in italics represent existing UMLS concepts for which the semantic type was redefined, such as ‘Smokers,’ ‘Smoking,’ ‘Healthful behavior,’ and ‘Physical Activity.’

Evaluation

The evaluation set consisted of 2,163 sentences from 218 abstracts distinct from those used in the training set. PubMed searches were carried out on 1/23/2012, restricted to citations with abstracts. Search terms were “Obesity” AND “Health Promotion,” “Smoking Cessation” AND “Health Promotion” as MeSH major topics. Judgments were assigned by a physician (CAS) with experience in natural language processing, who did not participate in any stage of the development of this enhanced version of SemRep. In the evaluation set, SemRep identified 773 predications. Of these, 658 were correct (true positive) and 115 were false positives, yielding precision of 0.85.

Error Analysis

Analysis of enhanced SemRep output indicated that 110 of the 115 false positive errors were due to shortcomings in SemRep or the knowledge sources it is based on, independent of the extension to the public health domain, with most mistakes (64.4%) due to infelicitous mappings to UMLS concepts or their assigned semantic types. The error analysis determined the following main categories of errors, discussed below:

UMLS-related errors—Infelicitous MetaMap mapping to UMLS concepts: one example is the mapping of a broader term to a more restrictive one. For example, all instances of ‘model’ in the dataset were erroneously mapped to the narrower UMLS concept ‘Biological models,’ as in (11). This error alone accounted for 5 false positives.

- (11) *In this study we use a recently developed semi-varying coefficient model to elucidate this relationship.*

Predication: Clinical Research USES Biological Models

Argument Identification: Current Metathesaurus content interacts infelicitously with the lack of a concept appropriate for this domain, such as the multi-term expression “*ill-effects*” in (12). Thus the expression ‘ill’ got mapped to the UMLS concept ‘malaise,’ triggering a false positive predication:

- (12) *Tobacco use is a health hazard and its use is attributed to a lack of knowledge regarding the ill-effects of tobacco.*

Predication: Tobacco AFFECTS Malaise

Ambiguity: This is a well-known problem in NLP. (13) and (14) below illustrate part-of-speech ambiguity and lexical ambiguity respectively. In the former, ‘individual’ is taken as a noun although the adjective is being used. In (14), ‘plant’ refers to ‘Engineering plant’ (a concept missing in the UMLS) as opposed to a botanical entity.

- (13) *Qualitative study based on individual in-depth interviews with GPs and their patients.*

Predication: Qualitative Research TARGETS Individual

- (14) *Obesity and work: proposal for a multidisciplinary intervention model for prevention and its application in an engineering plant*

Predication: Obesity PROCESS_OF Plants

SemRep processing—Over 20% of all false positives were due to SemRep processing errors, such as failure to recognize coordinated pre-modifiers, negated structures, prepositional phrase attachment, to name a few. In (15), the coordinated pre-modifier ‘*health*’ is taken as the argument of the predication instead of the correct ‘*health inequalities*’:

- (15) *This study highlights the importance of understanding the meaning of smoking in different social contexts so future tobacco control interventions can be developed to reduce health and social inequalities.*

Predication: Health Intervention PREVENTS Health

Missing or erroneous rules—Slightly over 12% of false positives were due to the co-occurrence of the semantic types of the arguments of the predication not included in the allowable rules. In these cases, the true argument is ignored and another argument with a semantic type allowed by an existent rule may be used instead, triggering a false positive. For example, in (16), the objects measured or evaluated were ‘*height*’ and ‘*weight*,’ with UMLS-assigned semantic types not considered for the object arguments of EVALUATES. Instead, *behavior* (‘*girls’ behaviors*’) was used:

- (16) *At baseline and follow-up girls’ behaviors were self-reported, height and weight were measured by study staff*

Predication: Staff EVALUATES Behavior

DISCUSSION

Analysis results of enhanced SemRep output are fairly typical of general SemRep, with most errors related to the issues of NLP and inadequate mappings. This is not surprising, as most of the current research in this area is geared towards having better mappings and more accurate NLP processing. Domain-inadequate mappings are an unwelcome consequence when using any higher-level knowledge source since it lacks the required level of granularity for specific domains. In working towards a solution for this issue, we have developed a mechanism to prevent those infelicitous mappings when dealing with domain concepts, either by specifically canceling some mappings, or replacing them with ones that are more contextually appropriate, grounded in the texts, as needed. Argument identification plays an important role in any NLP system, and SemRep is no exception. It goes beyond missing concepts. For example, a domain-appropriate UMLS term with the wrong semantic type for this domain will be missed by SemRep. This would require changing the semantic type so it adequately captures the meaning as used in this domain, as the examples in Table VI show, facilitating the generation of a true positive predication if adequately specified in the co-occurrence rules. Missing UMLS concepts and missing SemRep rules, such as extending the semantic types allowed to co-occur in specific predicates, are easily incorporated into SemRep's supplementary domain files, but only after relevant texts bring the missing entities to light. We are guided by the data to avoid overgeneration of relationships, which could result in false positive predications. There are no across-the-board solutions for any single issue.

Most SemRep issues are not domain-dependent, and over the years we have resolved many recurrent problems, such as Word Sense Disambiguation (Humphrey, Rogers, Kilicoglu, Demner-Fushman, & Rindflesch, 2006). The issues highlighted in this project that are due to SemRep processing are well known to us, as they occur in general SemRep. Improvements on each area constitute a research project of its own, and this process is ongoing. Program improvements and efforts to elicit more accurate mappings notwithstanding, it is impossible to attain 100% accuracy with an automatic text analysis methodology such as SemRep. Further, punctuation errors, misspellings, and related issues in text often get in the way as well. Despite these drawbacks, SemRep users find it quite useful having a Web application in which large-scale information retrieval is summarized automatically and visualized as a network of semantic predications. (Miller, et al. 2012, Jonnalagadda, S. *et al.*, 2012).

CONCLUSIONS

Exploiting Semantic Predications in the Public Health domain

Despite the availability of the information to meet the diverse needs of public health professionals, the barriers they experience to accessing public health information has been well documented (LaPelle et al., 2006; Revere, Turner, Madhavan, Rambo, Bugni, Kimball, & Fuller, 2007). Some of these barriers include time, resource reliability, trustworthiness/credibility of information, and information overload (Revere et al., 2007). Public health professionals need a way to search through the available literature to find the necessary information. SemRep and Semantic MEDLINE can provide a way to assist public health professionals find the much needed information in [PubMed].

The ability to extract semantic predications from health promotion documents underpins the possibility of exploiting information management applications in the public health domain. Semantic MEDLINE (Kilicoglu et al., 2008), an advanced information management UMLS-based application, is intended to assist users by helping them manage the results of PubMed searches. Since it relies on SemRep—which in turn relies on UMLS knowledge—to extract semantic predications from PubMed titles and abstracts, it is mainly limited to biomedicine, and uses NLP, automatic summarization, and visualization to represent key entities in a given domain. Extracted predications can be visualized as an informative graph with links to the original citations. Its application in the Health Promotion arena depends on the development of an ontology to support it (with increased precision) with managing, summarizing, and visualizing the results of searching health-related information documents. The ontology is formally represented as subject-verb-object propositional-like semantic predications as presented in the UMLS Semantic Network, compatible with SemRep and also with the Semantic Web (Berners-Lee, Handler, & Lassila, 2001). Implementation of new domain knowledge focuses on compatibility with SemRep (i.e., formatting) and the UMLS, knowledge integration with existing UMLS concepts and relations (including synonyms for new concepts as well as new synonyms for existing concepts), while consonant with the ontological constructs on which SemRep depends, namely, Metathesaurus concepts categorized into semantic types, the indicator rules, and the Semantic Network predications consisting of semantic predicates with arguments of specific semantic types.

Creating a new semantic representation entails updating and correcting the implemented ontology, as needed. Brewster, Alani, Wilks, and Dasmahapatra (2004) suggest evaluating the fit of an ontology to a certain corpus of texts. In our system, this evaluation can be done with the visualization architecture of information extracted, mapped against the ontology, and condensed into a network of core domain concepts and their binding relations. While this project is still in the laboratory phase, we are considering having it explored with potential users through well-known resources such as NLM's Public Health Partners (accessed at <http://www.nlm.nih.gov/nno/partners.html>) for ways in which it can be refined and more widely applied to public health practice. As demonstrated by the results of our preliminary evaluation, such a tool can go a long way towards improving access to manageable relevant information for public health practitioners and for the public health community at large, an important step in assisting the members of all communities in achieving the goal of becoming active participants in bringing about the desirable changes that are needed for a healthier society.

Acknowledgments

This research was supported in part by an appointment to the NLM Associate Fellowship Program sponsored by the National Library of Medicine and administered by the Oak Ridge Institute for Science and Education. This study was also supported by the Intramural Research Program of the National Institutes of Health, National Library of Medicine.

References

- Ahlers CB, Fiszman M, Demner-Fushman D, Lang FM, Rindflesch TC. Extracting semantic predications from Medline citations for pharmacogenomics. *Pacific Symp on Biocomp*. 2007;209–220.
- Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. *Proc AMIA Symp*. 2001;17–21. [PubMed: 11825149]
- Berners-Lee T, Handler J, Lassila O. The Semantic Web. *Scientific American*. 2001; 284(5):34–43. [PubMed: 11396337]
- Bodenreider O. Biomedical Ontologies in Action: Role in Knowledge Management, Data Integration and Decision Support. *Yearb Med Inform*. 2008; 47:67–79. [PubMed: 18660879]
- Brewster CA, Alani H, Wilks Y, Dasmahapatra S. Data driven ontology evaluation. *Proc LREC-2004*. 2004
- Buitelaar, P.; Cimiano, P.; Magnini, B., editors. *Ontology Learning from Text: Methods, Evaluation and Applications*. Amsterdam: IOS Press; 2005.
- Carter-Pokras O, Baoljet C. What is a “Health Disparity”? *Public Health Reports*. 2002; 117:426–34. [PubMed: 12500958]
- Chen L, Morrey CP, Gu H, Halper M, Perl Y. Modeling multi-typed structurally viewed chemicals with the UMLS Refined Semantic Network. *J Am Med Inform Assoc*. 2009 Jan-Feb;16(1):116–31. [PubMed: 18952946]
- Deroose, SF.; Schuster, MA.; Fielding, JE.; Asch, SM. *Public Health*. 2002. *Public Health Quality Measurement: Concepts and Challenges*.
- European Multilingual Thesaurus on Health Promotion. European Commission. 2001. Retrieved March 2012 from <http://ring.ciard.net/kos/european-multilingual-thesaurus-health-promotion>
- Feltrin E, Campanaro S, Diehl AD, Ehler E, Faulkner G, Fordham J, Gardin C, Harris M, Hill D, Knoell R, Laveder P, Mitterpergher L, Nori A, Reggiani C, Sorrentino V, Volpe P, Zara I, Valle G, Deegan née Clark J. Muscle Research and Gene Ontology: New standards for improved data integration. *BMC Med Genomics*. 2009; 2:6. [PubMed: 19178689]
- Health Promotion Glossary. World Health Organization; 2005. Posted to <http://www.who.int/healthpromotion/about/HPG/en/>
- Healthy People 2020 Structured Evidence Queries. Partners in Information Access for the Public Health Workforce. Posted to <http://phpartners.org/hp2020/index.html#whatisiap>
- Humphreys BL, Lindberg DA, Schoolman HM, Barnett GO. The Unified Medical language System: An informatics research collaboration. *J Am Med Inform Assoc*. 1998 Jan-Feb;5(1):1–11. [PubMed: 9452981]
- Humphrey SM, Rogers WJ, Kilicoglu H, Demner-Fushman D, Rindflesch TC. Word sense disambiguation by selecting the best semantic type based on Journal Descriptor Indexing. *Journal of the American Society for Information Science and Technology (JASIST)*. Jan; 2006 57(1):96–113. [PubMed: 19890434]
- Institute of Medicine. *The future of Public Health*. National Academy Press; Washington, D.C: 1998. Report posted to http://www.nap.edu/openbook.php?record_id=1091&page=7, Full report downloadable from <http://www.iom.edu/Reports/1988/The-Future-of-Public-Health.aspx>
- Irwin Y, Harkema H, Christensen L, Schleyer T, Haug P, Chapman W. Methodology to develop and evaluate a semantic representation for NLP. *Proc, AMIA Annual Symp Proc*. 2009:271–275.
- Jonnalagadda, S.; Del Fiol, G.; Medlin, R., Jr; Weir, C.; Fiszman, M.; Mostafa, J.; Liu, H. Automatically extracting sentences from Medline citations to support clinicians’ information needs. *Proc., 2nd IEEE Conference on Healthcare Informatics, Imaging, and Systems Biology*; 2012. in press
- Keselman A, Rosemblat G, Kilicoglu H, Fiszman M, Jin H, Shin D, Rindflesch TC. Adapting semantic natural language processing technology to address information overload in influenza epidemic management. *Journal of the American Society for Information Science and Technology (JASIST)*. Dec; 2010 61(12):2531–2543.

- Kilicoglu, H.; Fiszman, M.; Rodriguez, A.; Shin, D.; Ripple, AM.; Rindflesch, TC. Semantic MEDLINE: A Web application to manage the results of PubMed searches. Proceedings of the Third International Symposium for Semantic Mining in Biomedicine; 2008; 2008. p. 69-76.
- Kostova, P.; Kumar, A.; Roy, A.; Weinberg, J.; Madle, G.; Carson, E. Ontological Principles of Disease Management from Public Health Perspective: a Tuberculosis Case Study. City University; London, UK: 2005.
- Kuziemsky CE, Lau F. A four stage approach for ontology-based health information system design. Artificial Intelligence in Medicine. 2010; 50:133–48. [PubMed: 20510592]
- LaPelle NR, Luckmann R, Hatheway Simpson E, Martin ER. Identifying strategies to improve access to credible and relevant information for public health professionals: a qualitative study. BMC Public Health. 2006; 6:89.10.1186/1471-2458-6-89 [PubMed: 16597331]
- Lynch, CO.; Cunningham, C.; Schripsema, E.; Morris, T.; Rhodes, B. A biosurveillance platform for BioSense Message analysis using integrated reference ontologies and intelligent agents. Proc. Twenty-Second Conference on Artificial Intelligence (AAAI-07); 2007. p. 86-91.
- Miller CM, Rindflesch TC, Fiszman M, Hristovski D, Shin D, Rosemblat G, Zhang H, Strohl KP. A closed literature-based discovery technique finds a mechanistic link between hypogonadism and diminished sleep quality in aging men. Sleep. 2012; 35(2):279–285. [PubMed: 22294819]
- Mirhaji P. Public health surveillance meets translational informatics: A desiderata. Journal of Laboratory Automation. 2009; 14:157.
- Mokdad AH, Marks JS, Stroup DF, Gerberding JL. Actual causes of death in the United States, 2000. JAMA. 2004; 291:1238–45. [PubMed: 15010446]
- Noy, NF. Ontology Mapping. In: Staab, S.; Studer, R., editors. Handbook on Ontologies, International Handbooks on Information Systems Part. Vol. 4. 2009. p. 573-590.
- OntoReason Public Health Ontology. Cited in Lynch CO, Cunningham C, Schripsema E, Morris T, Rhodes B. 2007 A Protégé Ontology as the Core Component of a BioSense Message Analysis Framework. Nov. Accessed at http://protege.stanford.edu/conference/2007/presentations/02.01_Lynch.pdf 2011; Company information at <http://ontoreason.com/>
- Partners in Information Access for the Public Health Workforce. Reports posted to <http://phpartners.org/hp2020/index.html#whatispartners>
- Public Health Functions Steering Committee. Public Health in America. 1994. Available at www.health.gov/phfunctions/public.htm; <http://www.health.gov/phfunctions/project.htm>
- Resnick MP, Ripple AM, Auston I, Rindflesch T. An Ontology for Public Health to Support Enhanced Information Retrieval. AMIA Annual Symposium. 2010; 2010:50.
- Revere D, Turner AM, Madhavan A, Rambo N, Bugni PF, Kimball AM, Fuller SS. Understanding the information needs of public health practitioners: a literature review to inform design of an interactive digital knowledge management system. Journal of Biomedical Informatics. 2007; 40:410–421. [PubMed: 17324632]
- Rindflesch TC, Fiszman M, Rosemblat G, Shin D. Semantic MEDLINE: An advanced information management application for biomedicine. Information Systems and Use. 2011; 31(1,2):15–21.
- Rindflesch, TC.; Fiszman, M.; Libbus, B. Semantic interpretation for the biomedical research literature. In: Chen; Fuller; Hersh; Friedman, editors. Medical informatics: Knowledge management and data mining in biomedicine. Springer; 2005. p. 399-422.
- Rindflesch, TC.; Libbus, B.; Hristovski, D.; Aronson, AR.; Kilicoglu, H. Semantic relations asserting the etiology of genetic diseases. Proc AMIA Symposium; 2003; 2003. p. 554-8.
- Srinivasan, A.; Kunapareddy, N.; Mirhaji, P.; Casscells, SW. Semantic web representation of LOINC: an ontological perspective. AMIA Annual Symp Proc; 2006. p. 1107
- Stearns, MQ.; Price, C.; Spackman, KA.; Wang, AY. SNOMED clinical terms: overview of the development process and project status. Proc. AMIA Symposium; 2001; 2001. p. 139-43.
- U.S. Department of Health and Human Services. Healthy People reports. Initiative posted at www.healthypeople.gov
- U.S. National Library of Medicine. Project Partners. Partners in Information Access for the Public Health Workforce. Accessed at <http://www.nlm.nih.gov/nno/partners.html>

Figure 1.
Semantic MEDLINE graph focusing on Post-traumatic Stress Disorder.

Figure 2.

Health Promotion Semantic Representation. Color-coded arrows indicating relations between concepts correspond to the named predicates (key on the right): Arrows in various shades of green correspond to new domain predicates (ADDRESSES, PROMOTES, PROVIDES, TARGETS, VENUE_FOR); red and pink arrows correspond to current UMLS predicates (TREATS, PREVENTS). Light-colored boxes show existing UMLS concepts with UMLS-assigned semantic types; darker boxes with italics display new concepts with either existing or new semantic types; darker boxes without italics correspond to existing UMLS concepts with redefined semantic types.

Table I

Examples of content-rich sentences analyzed for this project.

PMID	Example sentence
18415827	<i>The goal of this study was to document similarities and differences in beliefs about the causes, symptoms, and treatments of asthma across two sites and two Latino ethnic sub-groups of children who remain the most at risk for asthma morbidity.</i>
15671457	<i>Experiences of racial-ethnic harassment may contribute to tobacco use behaviors among some African American young adults.</i>
16523248	<i>The aim of this study was to determine the utility of screening tools, namely heel ultrasound, the osteoporosis self- assessment tool (OST), weight-based criterion (WBC) and body mass index (BMI), in screening for low bone mass in African American men.</i>

Table II

Breakdown by theme of the total number of sentences in the training set.

THEME	Number of Sentences	EXAMPLE SENTENCE
ASSESSMENT	180	<i>African Americans had slightly higher prevalence of knee symptoms.</i>
POLICY DEVELOPMENT	121	<i>Health care interventions focused on the roles of nurses to deliver culturally tailored health information.</i>
ASSURANCE	474	<i>Spanish-speaking participants of an arthritis self-management program demonstrate short- and long-term benefits (improved health behaviors, health status, and self-efficacy).</i>

Table III

Main subcategories within the three domain themes presented in Table II.

Category I: ASSESSMENT	
Subcategories ^{1,2}	Demographic factors, including societal issues and health-related behaviors
	Health issues that afflict populations, including disorders, prevalence, monitoring, risk-assessment
Category II: POLICY DEVELOPMENT	
Subcategories	Programs, interventions, initiatives
	Program type, design and purpose, target audience, tools, venues, implementers
	Program implementation
Category III: ASSURANCE	
Subcategories ³	Assure accessibility for all populations, eliminating disparities
	Assure design, tools, and training of workforce are optimized for effective results and tailored to target audience
	Assure adequate tools/resources are available and used in the implementation
	Assure effectiveness and compliance with ethical principles; measurement of outcomes and evaluation

¹ Differences in these subcategories lead to differences in health status and identify health disparities.

² Steps taken to improve health status and eliminate health disparities and potential issues give rise to Category II.

³ Measures of effectiveness and quality control

Table IV

Examples of new concepts added with an existing semantic type.

New Concept	UMLS Semantic Type
<i>Seminar</i>	Educational Activity
<i>Practitioner</i>	Population Group
<i>Preventative Effort</i>	Health Care Activity

Table V

Examples of new concepts added with a newly created semantic type.

New Concept	New Semantic Type
<i>Project</i>	Program Type
<i>Nonsmoking behavior</i>	Healthy Behavior
<i>Forum</i>	Venue

Table VI

Examples of redefined UMLS concepts.

UMLS Concept	From UMLS-assigned Semantic Type	To Domain-Appropriate Semantic Type
<i>Course</i>	Temporal Concept	Educational Activity
<i>Surveys</i>	Intellectual Product	Research Activity
<i>Counselor</i>	Finding	Professional Group

Table VII

Examples of new synonyms added for UMLS concepts.

New Synonym	UMLS Concept	UMLS-assigned Semantic Type
<i>Elders</i>	Elderly Person	Age Group
<i>Subpopulation</i>	Population Group	Population Group
<i>Quitter</i>	former smoker	Population Group

Table VIII

Main categories of errors.

E.g.#	Problem type	Subtype	Subtotal	N=	Percentage	
	<i>UMLS-related errors</i>			74		64.4%
(11)		Infelicitous UMLS concept mapping or infelicitous semantic type in the UMLS concept	49		42.6%	
(12)		Argument identification due to missing concepts	20		17.4%	
(13)		Part-of-Speech Ambiguity	3		2.6%	
(14)		Other ambiguity	2		1.8%	
(15)	<i>SemRep processing</i>		26	26	22.6%	22.6%
(16)	<i>Missing or erroneous rules</i>			15		13.0%
		Domain-specific	10		8.7%	
		General SemRep	5		4.3%	
	TOTALS		115	115	100%	100%