Controlled Health Thesaurus for the CDC Web Redesign Project

Fabio Almeida, MD; Mamie Bell, MLn; Eliana Chavez, MD, MPhil

Abstract

A considerable number of robust vocabularies and thesauri have been developed for the healthcare and biomedical domain. No single vocabulary, however; provides complete coverage of the information needs from a public health perspective. The results of an investigation of vocabulary sources for the development of a comprehensive controlled vocabulary for the public health domain at the Centers for Disease Control and Prevention (CDC) is presented.

Introduction

To manage the CDC's web content, a controlled health thesaurus is being developed to ensure consistent metadata tagging and accurate information retrieval. Existing vocabularies may serve as content starting points to avoid duplication of work [1,2]. Our criteria for vocabulary selection included concept coverage, robustness of synonyms and lexical variants, vocabulary structure and free availability for public use.

Methods

Concepts and terminology important to the CDC's domain were identified using standard methods [3] and encompassed the following internal data: an inventory of databases and controlled vocabularies, search logs, indexes and glossaries, concepts extracted (noun-phrase) from key content and expert review.

1) Eight internal vocabularies (CDC inventory) were parsed and combined into a single database structure to be used for data analysis.
2) Search logs for 14 months were parsed to extract search strings and queried for terms with multiple occurrences (range of 2-227,000). A threshold of 500 occurrences was selected, which provided the top 2,500 unique search terms.
3) Terms from indexes and glossaries were extracted and decomposed into individual noun-phrases.
4) Key "content" pages were identified from WebTrends™ data, providing the top 1000 content pages. An additional 750 pages were added to the review from direct expert navigation on the CDC web site. Using concept extraction software, up to 100 concepts were extracted and ranked from each page. A total of 133,000 concepts were extracted representing 10,765 unique strings.

The data from all collection methods was combined into a single data structure. Automated and computer assisted matching/coding, vocabulary source and semantic grouping analysis using the Unified Medical Language System (UMLS) (Give provider information or reference) was performed.

Results

The results are shown in the Table below:

Source for Analysis	Unique Strings	Coded to UMLS	%
CDC Inventory	7174	4276	60%
Search Logs	2500	1691	68%
Indexes and Glossaries	2375	1153	48%
Concept Extraction	10765	3638	34%
All sources	19,475	7953	41%

The top ranking source vocabularies where MeSH (5608), REED (4206), SNOMED (3941), CSP (3248), AOD (3188), LCH (2361), MTH (2027), and MEDRA (1893). Incremental coverage was identified by combining several vocabularies with MeSH and MTH. The combination of MeSH/MTH/CSP/AOD provided good total coverage of coded matches (86%), which was comparable to MeSH/MTH and SNOMED. The addition of SNOMED to MeSH/MTH/CSP/AOD provided only a modest additional increase in coverage.

Discussion

Our vocabulary analysis demonstrated that no single or combination of existing standardized vocabularies provides complete concept coverage to fulfill the needs of a broad public controlled health thesaurus. Only about 50% coverage on average can be achieved as a starting point.

The Medical Subject Heading (MeSH) Thesaurus, used along with MTH/CSP/AOD, best meets our overall selection criteria and represents the optimal starting point. The vocabulary analysis revealed that approximately 11,000 strings were not matched to any of the UMLS vocabularies. While some of these "gap" terms can be filtered out as lexical variants or foreign language terms, the remaining terms represent specific public health areas such as occupational health and safety, environmental health and injury prevention. These terms will be the focus of vocabulary development as our project proceeds.

References

1. Chute CG, Cohn SP, Campbell KE, Oliver DE, Campbell JR. The content coverage of clinical classifications. For The Computer-Based Patient Record Institute's Work Group on Codes & Structures. J Am Med Inform Assoc 1996 May-Jun;3(3):224-33

2. Coletti MH, Bleich HL. Medical subject headings used to search the biomedical literature. J Am Med Inform Assoc 2001 Jul-Aug;8(4):317-23

3. National Information Standards Institute. *Guidelines for the Construction, Format, and Management of Monolingual Thesauri*. Bethesda, MD: NISO Press, 1994. (ANSI/NISO Z39.19-1993)