Title: Detecting Misaligned and Missing Concepts in SNOMED CT using Structural and Lexical Patterns

Presenter: Licong Cui\textsuperscript{a}, Wei Zhu\textsuperscript{a}, Shiqiang Tao\textsuperscript{a}, James T Case\textsuperscript{b}, Olivier Bodenreider\textsuperscript{b}, GQ Zhang\textsuperscript{a}
\textsuperscript{a}: University of Kentucky, Lexington, USA; \textsuperscript{b}: National Library of Medicine, USA

Audience
Developers and users of SNOMED CT interested quality assurance.

Objectives
1. To learn about quality assurance challenges in SNOMED CT;
2. To understand how structural and lexical approaches can help identify misaligned and missing concepts;
3. To describe the contribution of visual representation of subgraphs to quality assurance.

Abstract
Objective: Quality assurance of large ontological systems such as SNOMED CT is an indispensable part of the terminology management lifecycle. We introduce a hybrid structural-lexical method for scalable and systematic discovery of novel anomalies in SNOMED CT. The structural component is based on shared \textit{isa} relations to other concepts. The lexical component leverages shared words in descriptions between concepts.

Material and Methods: All non-lattice subgraphs (the structural part) in SNOMED CT are exhaustively extracted. Four types of lexical patterns (the lexical part) are identified among the concepts involved in non-lattice subgraphs. Non-lattice subgraphs exhibiting such lexical patterns are often indicative of misaligned and missing concepts.

Results: Applying our hybrid structural-lexical method to the September 2015 version of SNOMED CT (U.S. edition), we extracted 171,011 non-lattice subgraphs, among which 6,801 matched the lexical patterns. A subset of 2,046 small non-lattice subgraphs with sizes 4 to 6 with lexical patterns was obtained. A random sample of 100 of these subgraphs was selected, visualized and manually reviewed by two domain experts. Of these, 59 (59\%) revealed errors confirmed by the experts. The most frequent type of error was missing \textit{isa} relations due to incomplete or inconsistent modeling of the concepts.

Discussion: The combined non-lattice and lexical-based anomalies have not been uncovered by other existing ontology quality assurance approaches known to date. Non-lattice subgraphs of sizes 4, 5 and 6 can be easily visualized for manual inspection by experts. It also makes sense to investigate them first, because they are often included in larger subgraphs.

Conclusions: Our hybrid structural-lexical method is innovative and effective in detecting SNOMED CT anomalies that have escaped existing quality assurance processes.

References