Next generation ontology browser

Tomasz Adamusiak MD PhD[1*], Naoki Shimoyama[1], Marek Tutaj[1] and Mary Shimoyama PhD[1,2*]

[1] Human and Molecular Genetics Center, Medical College of Wisconsin, Milwaukee, WI, United States
[2] Department of Surgery, Medical College of Wisconsin, Milwaukee, WI, United States

ABSTRACT

The eleven million concept names integrated within the Unified Medical Language System (UMLS), developed and maintained by the U.S. National Library of Medicine, make it an unparalleled resource in biomedical sciences. A resource that due to its sheer complexity remains often underused.

We developed a state of the art UMLS browser with a specific clinical focus on the three clinical terminologies: SNOMED CT, LOINC, and RxNorm. These terminologies are quickly becoming the cornerstone of U. S. health information interchange fuelled by the Meaningful Use legislation.

While earlier works have shown the usefulness of each of these ontologies separately, this is the first work that demonstrates all three integrated into a single, data-driven workflow.

A demo of the application (user: icbodemo, password: icbodemo) is available from:

https://clinminer.hmgc.mcw.edu

1 INTRODUCTION

The title is a reference to a seminal paper on next generation phenotyping of electronic health records (Hripcsak and Albers 2012) advocating the need for new tools capable of consuming the oncoming explosion of electronic medical data. The Health Information Technology for Economic and Clinical Health (HITECH) Act, enacted as part of the American Recovery and Reinvestment Act of 2009 (ARRA), positioned the Meaningful Use of interoperable Electronic Health Records as a critical goal and encouraged nationwide EHR adoption. From the standpoint of semantic interoperability, the most interesting development is the recently released Meaningful Use Stage 2 Rules as they define the mandatory vocabularies for health data exchange. In particular, the Consolidated Health Informatics (CHI) initiative recommended the following three terminologies for EHRs: SNOMED CT, LOINC, and RxNorm.

SNOMED CT (Systematized Nomenclature of Medicine, Clinical Terms) is one of the most comprehensive, multilingual medical terminologies in the world (Stearns et al. 2001). **LOINC** (Logical Observation Identifiers Names and Codes) is a universal standard for identifying laboratory observations and is considered the *lingua franca* of clinical observation exchange (McDonald 2003). **RxNorm** is a standardized nomenclature for generic and branded drugs integrating a number of different drug resources (Parrish et al. 2006). All three terminologies are integrated within the Unified Medical Language System (UMLS) (Lindberg, Humphreys, and McCray 1993).

The Clinical Informatics team at the Medical College of Wisconsin has developed ClinMiner, a clinical research portal for clinical and diagnostic information on patients in genetics clinics and clinical sequencing programs, as well as other clinical research projects.

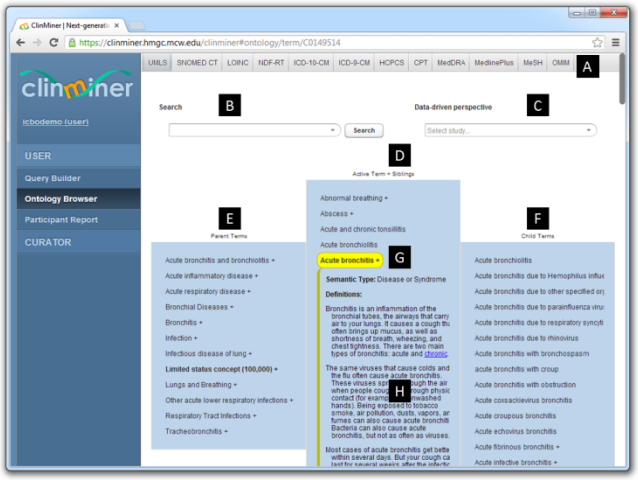

Fig. 1. Screenshot of the ClinMiner ontology browser. Tabs allow for switching between different sources and UMLS (A). **Searching.** Typing a query into the input field (B) brings up a list of suggested search terms. Results are displayed in the middle pane (D). **Browsing.** The currently active term is highlighted in yellow (G) in the middle pane (D) together with its siblings. Parent terms of the active term are displayed in the left pane (E) and child terms are displayed in the right pane (F). Meta data (H) is shown below the term. Selecting a study from the listbox box (C) switches the data-driven perspective that only shows a fragment of the ontology tree relevant to a particular study. A larger version of the figure is available at: http://dx.doi.org/10.6084/m9.figshare.155879

Data for the system consists of many clinical and referral documents the patients have accumulated throughout their clinic and diagnostic histories, and are standardized through the three Meaningful Use Ontologies: SNOMED CT, RxNorm and LOINC, and the UMLS. The challenge here is that these terminologies encompass hundreds of thousands

* To whom correspondence should be addressed: tadamusiak@mcw.edu; shimoyama@mcw.edu

of clinical concepts and have intricate and complex hierarchies.

2 METHODS

The terminology core service is provided by a UMLS oracle-based installation configured with a standard set of over 150 terminologies. For the sake of usability, only a preconfigured subset is displayed as navigable tabs at the top of the browser window (see Figure 1A). However, all the available UMLS sources are used in search and in synonym expansion.

Indexing was implemented using Oracle Text – a powerful feature of Oracle databases that provides unparalleled, Google-like indexing and text classification capabilities. Individual tokens are using the tf–idf (term frequency, inverse document frequency) algorithm reflecting how often a particular string occurs in the corpus (Salton 1991). A query relaxation approach was used instead of an advanced query interface that typically negatively impacts user experience. Every search sequence starts with an exact phrase match, progresses into matching all tokens in a close sequence (*NEAR* PL/SQL operator), all words matched (*AND*) in a phrase, most words matched (*ACCUM*), and finally fuzzy matching and wildcard expansion.

Visualizing UMLS Metathesaurus is difficult because of its size and lack of obvious starting points for exploration. In an approach similar to (Patel and Cimino 2007) all UMLS concepts were ranked according to their *branching factor* (number of children) and number of unique source mappings. One hundred top-ranked concepts were then selected from each of the sources separately to achieve equal representation in the final result set of 167 nodes and 230 edges (some concepts overlapped). LOINC codes and parts were considered independently due to their different nature(Adamusiak and Bodenreider 2012).

In order to minimize user effort involved in browsing large hierarchies the ontology graph was treated as minimum Steiner tree problem (Gilbert and Pollak 1968). This results in a more compact ontology tree, which includes only the directly annotated concepts and their close neighborhood, or in other words, identifies concepts and their relations that more effectively partition the underlying data. This also identified orphaned nodes that were otherwise disconnected from hierarchy, placing them at the root of the tree.

3 RESULTS

Interpreting a query string simultaneously using different operator combinations allows for a more concise query design. For example, if a user enters a query *rash on examination*, the application can interpret the query in parallel as a single phrase '*rash on examination*' and *rash OR on OR examination* to increase recall. Fuzzy and wildcard matching typically provide the most results at the expense of precision, but are also automatically ranked lower than exact matches if such exist.

For example, searching for a misspelled *myleoid leukemia*, returns *C0023470 Myeloid Leukemia, C0001815 Primary Myelofibrosis, C0023467 Leukemia, Myelocytic, Acute*, etc.

The ontology browser discussed in this paper is a component of a larger system that also includes data entry forms, patient reports, advanced querying and data export. A demo of ClinMiner (user: icbodemo, password: icbodemo) is available at:

https://clinminer.hmgc.mcw.edu

4 DISCUSSION

This is the world's only clinical application that integrates all three terminologies in a single workflow and actually goes beyond Meaningful Use, as any terminology integrated within the UMLS can be used to annotate, visualize and query data. For example, legacy information annotated with ICD-9 codes can be just as easily integrated in the system.

While physicians rarely have to deal with ontology hierarchies directly, these are indispensable in clinical research to facilitate query expansion, building transitive closures, and data validation and reconciliation. Large terminologies have their unique challenges and are often too complex to use directly. On the other hand, maintaining local application ontologies requires considerable resources and can suffer from uncontrolled expansion beyond the initial scope. Data management through ClinMiner is a solution to the complexity and difficulties making annotations with UMLS.

REFERENCES

Adamusiak, Tomasz, and Olivier Bodenreider. 2012. "Quality Assurance in LOINC Using Description Logic." *AMIA ... Annual Symposium Proceedings / AMIA Symposium. AMIA Symposium* 2012 (January): 1099–108.

Gilbert, E N, and H O Pollak. 1968. "Steiner Minimal Trees." *SIAM Journal on Applied Mathematics* 16 (1): 1–29. doi:10.1137/0116001.

Hripcsak, George, and David J Albers. 2012. "Next-generation Phenotyping of Electronic Health Records." *Journal of the American Medical Informatics Association : JAMIA* (September 30). doi:10.1136/amiajnl-2012-001145.

Lindberg, D A, B L Humphreys, and A T McCray. 1993. "The Unified Medical Language System." *Methods of Information in Medicine* 32 (4) (August): 281–91.

McDonald, C. J. 2003. "LOINC, a Universal Standard for Identifying Laboratory Observations: A 5-Year Update." *Clinical Chemistry* 49 (4) (April 1): 624–633. doi:10.1373/49.4.624.

Parrish, Fola, Nhan Do, Omar Bouhaddou, and Pradnya Warnekar. 2006. "Implementation of RxNorm as a Terminology Mediation Standard for Exchanging Pharmacy Medication Between Federal Agencies." *AMIA ... Annual Symposium Proceedings / AMIA Symposium. AMIA Symposium* (January): 1057.

Patel, Chintan O, and James J Cimino. 2007. "A Scale-free Network View of the UMLS to Learn Terminology Translations." *Studies in Health Technology and Informatics* 129 (Pt 1) (January): 689–93.

Salton, G. 1991. "Developments in Automatic Text Retrieval." *Science* 253 (5023): 974–980.

Stearns, M Q, C Price, K A Spackman, and A Y Wang. 2001. "SNOMED Clinical Terms: Overview of the Development Process and Project Status." *Proceedings / AMIA ... Annual Symposium. AMIA Symposium* (January): 662–6.