A Protégé Ontology as the Core Component of a BioSense Message Analysis Framework

Cecil O, Lynch[1], Craig Cunningham[1], Eric Schripsema[1], Tim Morris[2], Barry Rhodes[2]

[1] OntoReason, LLC, Salt Lake City, UT, [2] US Centers for Disease Control and Prevention, Atlanta, GA

Abstract

In this paper we describe a prototype application for real time BioSense message analysis and classification using a Protégé public health ontology built on an HL7 Version 3 Restricted Message Information Model (RMIM) and incorporated as the core component of a model driven, services based architecture.

Introduction

BioSense is a national biosurveillance program initiated by the US Centers for Disease control and Prevention (CDC) as part of the Public Health Information Network for the purpose of early event detection, quantification and spatio-temporal visualization of public health events and risks[1]. It currently receives anonymized data in the form of HL7 2.5 messages from more than 350 private and public acute care hospitals, 1 major commercial clinical laboratory systems, 866 Veterans Affairs healthcare facilities and 366 Department of Defense healthcare facilities in the US[2]. Additionally, all Poison Control Centers, in the US are slated to be added by the end of 2007. The system has the capacity to receive up to 72 million messages a day that must be analyzed and posted to users within 2 hours, demanding a scalable solution for analysis and routing.

Messages may contain data coded with standard Consolidated Health Informatics (CHI) code systems or may contain free text or local codes in some cases which require conversion to standardized code systems for further analytical processing. The BioSense Messages are of 4 basic domain types; 1) ADT (Admission, Discharge and Transfer) which captures data about patient presentation and disposition including Chief Complaint, 2) Laboratory Requests and Results, 3) Radiology Orders and Results, and 4) Pharmacy Orders. Messages are analyzed based on content to determine syndromic classification and routing. Messages are correlated to build specific event profiles, refine classification accuracy, provide situational awareness and develop situational assessment.

OTR Ontology

The OTR Ontology was designed using Protégé 3.1 frames. The decision to use frames was based on the requirement to model the layered meta-classes common to the structure of many HL7 V3 artifacts such as the RMIM used in the Public Health and Emergency Response messages. The HL7 Version 3 Case Notification Message (PORR_RM100001UV01) was used as the basis for the ontology since it most closely mimics the structure of a case definition for a reportable disease. To meet the computational reasoning requirements for case classification purposes, additional metaclasses were added to model incubation period, case frequency and other attributes not included in the Case Notification message. All concepts including those added that were not part of the Case Notification Message were modeled in the appropriate HL7 data types and class structures for the dual purpose of analysis of HL7 data streams and producing HL7 V3 artifacts for use in message generation.

Concept representations in the ontology are structured in a hierarchical relationship native to the code system or if required to allow for reasoning at the concept level. SNOMED is used as the core clinical code system and all syndromes and sub-syndromes in ICD9-CM were mapped to the appropriate synonymous SNOMED ConceptID. Additionally, a lexicon particular to BioSense was also mapped to SNOMED but using a non-synonymous matching slot since the semantics relationship was of an is_related_type that was coarsely mapped to a BioSense syndrome.

In order to process the BioSense messages, HL7 Version 3 attributes in the OTR Ontology were mapped to HL7 2.5 message segments and fields in order to analyze the simpler 2.5 messages in the context of the richer semantics of HL7 Version 3 objects.

The ontology concept representations are then structured in the appropriate HL7 data type including the CD data type that allows post-coordination of concepts to represent more complex concepts. These data types are then organized into the metaclass objects that correspond to the HL7 V3 abstract classes and finally organized into an overarching RMIM represented as a metaclass in the ontology, representing the expected artifacts of a clinical encounter with a patient given a specific disease entity. This constellation of clinical acts and entities related to each disease is thus a superset of a CDC Case Definition for that specific disease, allowing a matching set of necessary and sufficient conditions to be evaluated for confirmation or exclusion.

Reasoning Platform

The reasoning platform supports a series of collaborative expert systems that implement various artificial intelligent constructs each using the ontological content as facts within their knowledgebase. The platform allows for the configuration of various reasoning components to solve complex analysis problems while operating on a consistent, ontology driven knowledge model. The ontological concepts are used to provide traditional pattern matching and differential diagnosis. The expert system implementations leverage correlation factors derived from the ontology in conjunction with domain expertise represented in the ontological model. For example, our case classification reasoner uses specific ontology case definition contents structured in JESS facts and combined with algorithms that adjust the confidence weighting based on factors such as geographic and seasonal disease occurrence and frequency estimates to instantiate a disease condition, sub-syndrome or syndrome classification.

In addition, the ontological model provides the ability to implement pattern matching from inexact concepts based upon the ontological representation that is inherent to the concept hierarchy structure. The model supports generalized concept matching at the more primitive levels of the hierarchy and specialized matching at the more sophisticated concept levels in the hierarchy. The ontology extraction and message processing is depicted in figure 1**.**

Visualization Platform

We have built an integrated message visualization platform that constructs a directed acyclic graph of the message observations linked to the message segments for viewing by the Biointelligence Center BioSense Monitor. This allows rapid visual analysis and comparison and is updated as additional messages from the same patient are received, parsed and analyzed. A confidence threshold slider control allows the analyst to dynamically reconfigure the graph based on the level of confidence selected (see Figure 2).

Discussion

Using the OntoReason Public Health Ontology as the core reference model, we have produced a demonstration system that evaluates incoming HL7 2.5 BioSense messages, providing parsing, semantic validation, syndromic classification, and case definition in real time, as a service add-on to a generic HL7 interface. The ontology classifications can also provide a means to generate public health application value sets linked in context to a particular disease.

The analytical framework is a multithreaded application and scales to high message volumes for runtime analysis as an HL7 interface listener and utilizes the HL7 standard terminology CTS API for vocabulary maintenance requirements.

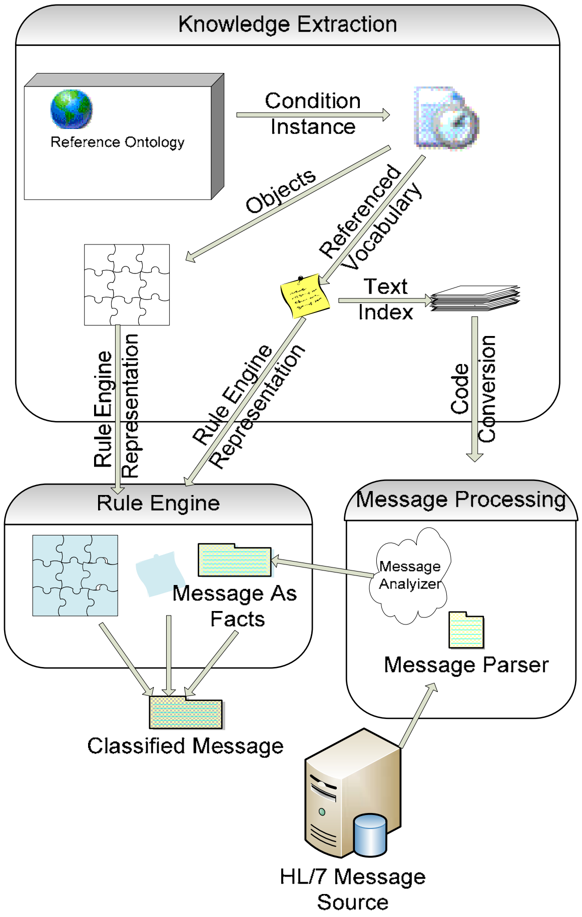

Figure 1

Buckeridge et al, were one of the first to demonstrate the use of an ontological framework for biosurveillance as part of the BioStorm project[3], defining the benefits of abstracting knowledge from applications as a means of lowering system costs and improving system flexibility. Additional ontological based biosurveillance work has been done by others, notably Mirhaji et al on chief complaint data analysis[4] and modeling of LOINC for laboratory surveillance[5]. Each of these approaches developed an ad hoc model for building the ontology that suited the purposes of the specific system implementations described.

Our approach differs from the previous work through the instantiation of a standard object model to generalize application functionality and in using both the model and content as a basis for reasoning. This HL7 based ontology approach has the major advantages of standardizing content exchange in a clinical messaging environment to enable standards based Model Driven Architecture software solutions that can be distributed in an Services Oriented Architecture environment as objects and persisted in a more efficient object database therefore maximizing throughput and flexibility for domain processes such as message fragment generation for visualization in context.

The major limitations of this approach include the complexity of the model which requires significant domain expertise both in the ontological modeling and the messaging environment which increases the time required to instantiate the ontology.

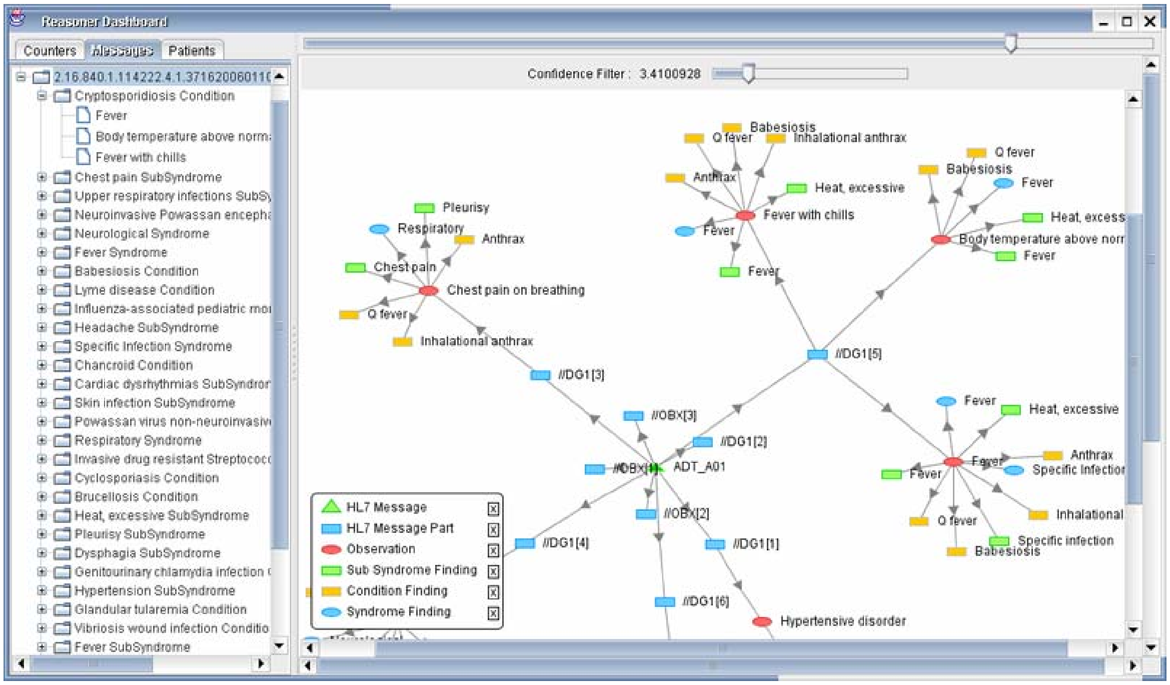

Figure 2

1. Loonsk, J.W., *BioSense--a national initiative for early detection and quantification of public health emergencies.* MMWR Morb Mortal Wkly Rep, 2004. **53 Suppl**: p. 53-5.

2. Steele, L. *BioSense: Integrating Local, Regional, Nationwide Biosurveillance Capabilities*. in *ISDS Annual Conference*. 2006. Baltimore, Maryland.

3. Buckeridge, D.L., et al., *Knowledge-based bioterrorism surveillance.* Proc AMIA Symp, 2002: p. 76-80.

4. Mirhaji, P., et al., *Semantic approach for text understanding of chief complaints data.* AMIA Annu Symp Proc, 2006: p. 1033.

5. Srinivasan, A., et al., *Semantic Web Representation of LOINC: an Ontological Perspective.* AMIA Annu Symp Proc, 2006: p. 1107.