Automated Information Integration from Heterogeneous Data Sources: A Semantic Web Approach

Parsa Mirhaji, Narendra Kunapareddy, Arunkumar Srinivasan, Sean Byrne, S. Ward Casscells

The University of Texas Health Science Center at Houston

OBJECTIVE

This paper proposes the use of Semantic Web technologies to integrate heterogeneous data generated by disparate systems for public health use.

BACKGROUND

Integration of information from multiple disparate and heterogeneous sources is a labor and resource intensive task. Heterogeneity can come about in the way data is represented or in the meaning of data in different contexts (semantics) [1]. Semantic Web technologies have been proposed to address both representational and semantic heterogeneity in distributed and collaborative environments [2]. We introduce an automated semantic information integration platform for public health surveillance using RDF and the Simple Knowledge Organization Standard (SKOS) [3] developed by the Semantic Web community.

Public health use of the information in systems participating in surveillance settings is not primarily anticipated. Further, decentralized and autonomous design and implementation of information systems has made it possible to extend the reach of surveillance systems to a variety of contextually disparate domains.

The task of integration in such environments is to provide a single and uniform query interface to underlying data. The Semantic Web is a framework specifically designed to foster information sharing and multidisciplinary use of informational resources in collaborative and distributed environments such as the World Wide Web. The Resource Definition Framework (RDF) provides a general purpose framework for representing information, with a schema language RDF(S) that provides the basic ingredients for a shared representation. The Simple Knowledge Organization System (SKOS) is a set of specifications and standards to support the use of knowledge organization systems (KOS) such as thesauri, classification schemes, subject heading lists, taxonomies and other types of controlled vocabulary, and possibly terminologies and glossaries, all within the framework of the Semantic Web.

METHODS

Schematization: In our methodology, an automated agent evaluates any incoming XML message and compares information element by element and attribute by attribute against an existing dynamically-built thesaurus.

New items are indexed using the SKOS:broader, SKOS:narrower, and SKOS:definition relationships among SKOS:Class(es). The schematizer ensures that every data instance has one and only one SKOS:Class representing it before submitting it to the data repository.

Integration: Each data instance is stored in the repository as an instance of an Observation class. Observations have properties to represent time stamps, source data, the SKOS:Class representing the observation, and any RDF:value associated with the observation.

RESULTS

An integrated RDF repository has been built from data submitted by eight community hospitals, consisting of all triage and medical records entries from emergency department visits, a combination of structured, semi-structured and non-structured data. All changes in the schema or vocabularies used in the source are captured automatically and reflected in the repository. The original context and schema of the data is computationally available for query planning and execution in the same repository.

CONCLUSIONS

The integrated repository lends itself to application of any kind of rule-based or ontology-based queries, natural language processing, Bayesian classification, and subgraph and link analysis algorithms.

A standard SPARQL [4]-based query language can be employed for information retrieval. External rule-base or inference engines can be interfaced with the repository for ontology-based classification and reasoning while querying for requested information.

REFERENCES

[1] Sujansky W. Heterogeneous Database Integration in Biomedicine. Journal of Biomedical Informatics. 2001;34,:285–98.

[2] Vdovjak R, Eindhoven G-JH. RDF Based Architecture for Semantic Integration of Heterogeneous Information Sources. Workshop on Information Integration on the Web; 2001; Eindhoven University of Technology; 2001.

[3] Miles A, Brickley D. Simple Knowledge Organisation System (SKOS). Nov 2005 [cited; Available from: http://www.w3.org/TR/2005/WD-swbp-skos-core-guide-20051102

[4] Prud'hommeaux E, Seaborne A. SPARQL Query Language for RDF. W3C. FEB2006.