A Semantic Framework for Global Disease Surveillance

Arunkumar Srinivasan MS [1, 2], Herman Tolentino MD [2], Asha Krishnaswamy MS [2,3]
[1] Division of Integrated Surveillance Systems and Services, NCPHI
[2] Public Health Informatics Fellowship Program, OWCD
[3] Division of Knowledge Management, NCPHI
The Centers for Disease Control and Prevention, Atlanta, GA 30033

Abstract

Global infectious disease surveillance allows epidemiologists to describe patterns of outbreaks of infectious diseases among human and animal populations across a spatio-temporal continuum [1]. Since disease surveillance involves reasoning tasks, it is essential to model semantic representation of collected data for effective interpretation and analysis by epidemiologists [2]. In this paper, we propose a framework to model public health data from distributed electronic sources in a global health surveillance context to support the emerging infectious disease surveillance functions of EpiSPIDER (Semantic Processing and Integration of Distributed Electronic Resources for Epidemiology) [3].

Today, Really Simple Syndication (RSS) feeds using Semantic Web technology have gradually become a de-facto Internet standard for content distribution and information exchange [4]. Type 1 RSS feeds are written using Resource Description Format (RDF), which, as the language of the Semantic Web, can potentially merge knowledge from different domains [2]. The proposed framework will support the use of RDF in a global health surveillance context to support named-entity recognition to extract geospatial, temporal and disease/event information from news and email information resources and utilize it for effective information representation (Figure. 1).

Since spatio-temporal and disease information are embedded in free text in emerging infectious disease reports and news feeds, we propose to extract them using Natural Language Processing (NLP) techniques like word tokenizing, part-of-speech tagging and named-entity recognition. The processed text will be harmonized by mapping them to appropriate UMLS Semantic Net concepts [5]. The disease conditions will subsequently be uniquely represented using SNOMED concepts [6]. Temporal concepts in the message like outbreak date and epidemic duration will be represented using time ontology concepts.

In determining risk for spread of emerging disease infections, we need to build knowledge about proximity of one location to another through a geographic proximity ontology. We also propose to represent the CIA Fact book [7] as an RDF knowledge base for geospatial, demographic, and socio-economic information. The country identity from nationality descriptors found in the incoming messages will be mapped to the relevant concepts in the fact book knowledge source. We also propose to capture other relevant information like reporting agencies (which are also named entities) and their related information. These RDF datasets are then seamlessly integrated and stored in a common RDF repository. Using Resource Description Query Language (RDQL), we perform test queries on the discrete knowledge bases and analyze their outputs, comparing them with information retrieval done by humans to evaluate performance.

We will use a convenience sample of ProMED [8] reports from EpiSPIDER to test the queries. We will test the merged output using an RSS reader and determine whether they were read correctly.

References

1. R. L. Berkelman and J. M. Hughes, Ann. Intern. Med. 119, 426 (1993).
2. P. Mirhaji,RL Richesson,A Srinivasan,J Zhang,J Smith. Public Health Surveillance; a Semantic Approach.SPIE; 2004. p. 339-50.
3. EpiSPIDER last accessed from http://epispider.org/ on 11 Sept 2006.
4. RSS 1.0 last accessed from http://web.resource.org/rss/1.0/spec on 11 Sept 2006.
5. D. Lindberg et al, Unified medical language systems. Methods of Information in Medicine, 32(4):281-291, 1993.
6. Côté R, Rothwell D. SNOMED-3.Chicago: College of American Pathologists, 1993.
7. CIA World Fact Book last accessed from https://www.cia.gov/cia/publications/factbook/index.html on 11 Sept. 2006.
8. ProMED Mail last accessed from http://www.promedmail.org on 11 Sept. 2006.

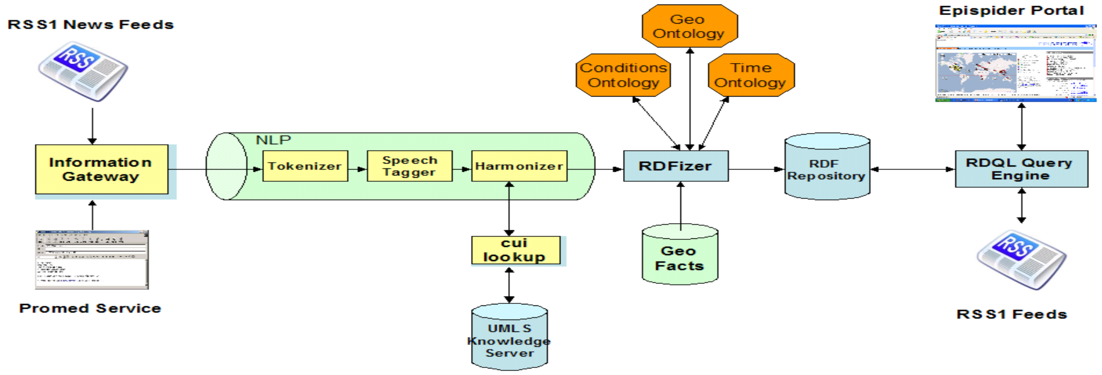

Figure 1. Knowledge extraction framework for emerging infectious disease surveillance