ISDS 2015 Conference Abstracts

ISDS INTERNATIONAL SOCIETY FOR DISEASE SURVEILLANCE

Developing the Scalable Data Integration for Disease Surveillance (SDIDS) Platform

David Buckeridge*[1], Maxime Lavigne[1], Kate Zinszer[2], Anya Okhmatovskaia[1], Samson Tu[3], Csongor Nyulus[3], Mark Musen[3], Wilson Lau[4], Lauren Carroll[4] and Neil Abernethy[4]

[1]Epidemiology and Biostatistics, McGill University, Montreal, QC, Canada; [2]Harvard University, Boston, MA, USA; [3]Stanford University, Stanford, CA, USA; [4]University of Washington, Seattle, WA, USA

Objective

To develop a scalable software platform for integrating existing global health surveillance data and to implement the platform for malaria surveillance in Uganda.

Introduction

Electronic data that could be used for global health surveillance are fragmented across diseases, organizations, and countries. This fragmentation frustrates efforts to analyze data and limits the amount of information available to guide disease control actions. In fields such as biology, semantic or knowledge-based methods are used extensively to integrate a wide range of electronically available data sources, thereby rapidly accelerating the pace of data analysis. Recognizing the potential of these semantic methods for global health surveillance, we have developed the Scalable Data Integration for Disease Surveillance (SDIDS) software platform. SDIDS is a knowledge-based system designed to enable the integration and analysis of data across multiple scales to support global health decision-making. A 'proof of concept' version of SDIDS is currently focused on data sources related to malaria surveillance in Uganda.

Methods

SDIDS is a web-based, ontology-driven software platform that automates the integration of heterogeneous data from multiple sources, and supports the discovery, retrieval, visualization, and analysis of these data. A data set is first "mapped" or linked explicitly to the ontologies used within SDIDS, and then the data are ingested into the system and stored in a manner that supports complex queries based on the concepts and relationships defined in the ontologies. Data in the system can be accessed via the SDIDS application program interface (API) for data visualization and analysis.

Results

Annotation and ingestion of data: A software client was created to guide a user through the semi-automated process of mapping a dataset to the SDIDS ontologies. This mapping process identifies the correspondence of each column: to a domain concept (e.g., 'Age', 'Symptom'), to a data type (e.g., 'Integer', 'Categorical), and possibly to a unit of measurement. For some data types, such as categorical variables, each unique value (e.g., 'Female', 'Male') must also be explicitly linked to a concept in the ontology. Natural language processing methods facilitate the identification of ontology concepts that are likely to correspond to elements of a dataset. Once the linkages are identified, mapping rules are automatically generated in the W3C standard language R2RML. When executed, these rules transform the data into RDF triples expressed in terms of the SDIDS ontologies. The R2RML mapping files are saved so that the provenance of all data is always accessible.

Retrieval and analysis of data: External applications can connect directly to SDIDS via an API to request data for further processing or to request the results of analyses applied to the integrated data. Within SDIDS, other server components facilitate the retrieval of data using SPARQL (a query language for semantic data), the transformation of data (e.g., aggregation, projection, joins, unions), and the calculation of indicators. Two software clients have been developed to demonstrate the functionality of SDIDS. One client addresses the needs of a malaria monitoring and evaluation manager tasked with following indicators of disease and control activities. The second client addresses the need of a funding officer in assessing malaria research activity and disease control measures in different geographical areas.

Conclusions

Our vision for global health surveillance is that data are easily shared and analyzed across diseases, countries, organizations, and data sources by a variety of users and client applications. SDIDS is a scalable platform that offers an initial step towards this vision. It is not a replacement for current systems, but a bridging technology that can help to integrate existing data now and encourage convergence of data models in the future. The next stage of the project will focus on scaling-up SDIDS to integrate surveillance data for the leading causes of under-5 mortality in Africa.

Keywords

surveillance; global health; ontology; data integration; malaria

Acknowledgments

Research supported by the Bill and Melinda Gates Foundation.

***David Buckeridge**
E-mail: david.buckeridge@mcgill.ca