OJPHI

ISDS 2016 Conference Abstracts

ISDS INTERNATIONAL SOCIETY FOR DISEASE SURVEILLANCE

Epi Archive: automated data collection of notifiable disease data

Nicholas Generous*, Geoffrey Fairchild, Hari Khalsa, Byron Tasseff and James Arnold

Los Alamos National Laboratory, Los Alamos, NM, USA

Objective

LANL has built a software program that automatically collects global notifiable disease data—particularly data stored in files—and makes it available and shareable within the Biosurveillance Ecosystem (BSVE) as a new data source. This will improve the prediction and early warning of disease events and other applications.

Introduction

Most countries do not report national notifiable disease data in a machine-readable format. Data are often in the form of a file that contains text, tables and graphs summarizing weekly or monthly disease counts. This presents a problem when information is needed for more data intensive approaches to epidemiology, biosurveillance and public health as exemplified by the Biosurveillance Ecosystem (BSVE).

While most nations do likely store their data in a machine-readable format, the governments are often hesitant to share data openly for a variety of reasons that include technical, political, economic, and motivational issues [1]. For example, an attempt by LANL to obtain a weekly version of openly available monthly data, reported by the Australian government, resulted in an onerous bureaucratic reply. The obstacles to obtaining data included: paperwork to request data from each of the Australian states and territories, a long delay to obtain data (up to 3 months) and extensive limitations on the data's use that prohibit collaboration and sharing. This type of experience when attempting to contact public health departments or ministries of health for data is not uncommon.

A survey conducted by LANL of notifiable disease data reporting in 52 countries identified only 10 as being machine-readable and 42 being reported in pdf files on a regular basis. Within the 42 nations that report in pdf files, 32 report in a structured, tabular format and 10 in a non-structured way.

As a result, LANL has developed a tool-Epi Archive (formerly known as EPIC)-to automatically and continuously collect global notifiable disease data and make it readily accesible.

Methods

We conducted a survey of the national notifiable disease reporting systems notating how the data is reported in two important dimensions: date standards and case definitions.

The development of software to regularly ingests notifiable disease data frand makes this data available involved four main steps scraping, extracting, parsing and persisting.

For scraping: we would examine website designs and determine reporting mechanisms for each country/website as well as what varies across the reporting mechanisms. We then designed and wrote code to automate the downloading of report pdf files, for each country. We stored report pdfs along with appropriate metadata for extracting and parsing.

For extracting: we developed software that can extract notifiable disease data presented in tabular form from a pdf file. We combined the methodology of figure placement detection with the in-house developed table extraction and annotation heuristics.

For parsing: we determined what to extract from each pdf data set from the survey conducted. We then parsed the extracted data into uniform data structures correctly accommodating the dimensions surveyed and the various human languages. This task involved ingesting notifiable disease data in many disparate formats extracted from pdf files and coalescing the data into a standardized format.

For persisting: We then store the data in the Epi Archive PostgreSQL database and make it available through the BSVE.

Results

The EpiArchive tool currently contains subnational notifiable disease data from 10 nations. When a user accesses the EpiArchive site, they are prompted with four fields: country, region, disease, and date duration. These fields allow the user to specify the location (down to the state level), the disease of interest, and the duration of interest. Upon form submission, a time series is generated from the users' specifications. The generated time series can then be downloaded into a csv file if a user is interested in performing personal analysis. Additionally, the data from EpiArchive can be reached through an API.

Conclusions

LANL as part of a currently funded DTRA effort so that it will automatically and continuously collect global notifiable disease data—particularly data stored in pdf files—and make it available and shareable within the Biosurveillance Ecosystem (BSVE) as a new data source. This will provide data to analytics and users that will improve the prediction and early warning of disease events and other applications.

Keywords

notifiable disease; data source; standards; scraping; data sharing

Acknowledgments

This project is supported by the Chemical and Biological Technologies Directorate Joint Science and Technology Office (JSTO), Defense Threat Reduction Agency (DTRA).

References

[1] van Panhuis WG, Paul P, Emerson C, et al. A systematic review of barriers to data sharing in public health. BMC Public Health. 2014. 14:1144. doi:10.1186/1471-2458-14-1144

***Nicholas Generous**
E-mail: generous@lanl.gov