OJPHI

ISDS 2018 Conference Abstracts

ISDS INTERNATIONAL SOCIETY FOR DISEASE SURVEILLANCE

Epi Archive: Automated Synthesis of Global Notifiable Disease Data

Hari S. Khalsa, Sergio Cordova*, Nicholas Generous, Prabhu S. Khalsa, Byron Tasseff and James Arnold

A-1, Los Alamos National Laboratory, Los Alamos, NM, USA

Objective

LANL has built software that automatically collects global notifiable disease data, synthesizes the data, and makes it available to humans and computers within the Biosurveillance Ecosystem (BSVE) as a novel data stream. These data have many applications including improving the prediction and early warning of disease events.

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.

While most nations likely store incident data in a machine-readable format, governments are often hesitant to share data openly for a variety of reasons that include technical, political, economic, and motivational issues[1].

A survey conducted by LANL of notifiable disease data reporting in over fifty countries identified only a few websites that report data in a machine-readable format. The majority (>70%) produce reports as PDF files on a regular basis. The bulk of the PDF reports present data in a structured tabular format, while some report in natural language.

The structure and format of PDF reports change often; this adds to the complexity of identifying and parsing the desired data. Not all websites publish in English, and it is common to find typos and clerical errors.

LANL has developed a tool, Epi Archive, to collect global notifiable disease data automatically and continuously and make it uniform and readily accessible.

Methods

We conducted a survey of the national notifiable disease reporting systems notating how the data are reported and in what formats. We determined the minimal metadata that is required to contextualize incident counts properly, as well as optional metadata that is commonly found.

The development of software to regularly ingest notifiable disease data and make it available involves three or four main steps: scraping, detecting, parsing and persisting.

Scraping: we examine website design 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 the data for each country. We store all artifacts presented as files (PDF, XLSX, etc.) in their original form, along with appropriate metadata for parsing and data provenance.

Detecting: This step is required when parsing structured non-machine-readable data such as tabular data in PDF files. We combined the Nurminen methodology of PDF table detection with in-house heuristics to find the desired data within PDF reports[2].

Parsing: We determined what to extract from each dataset and parsed these data into uniform data structures, correctly accommodating the variations in metadata (e.g., time interval definitions) and the various human languages.

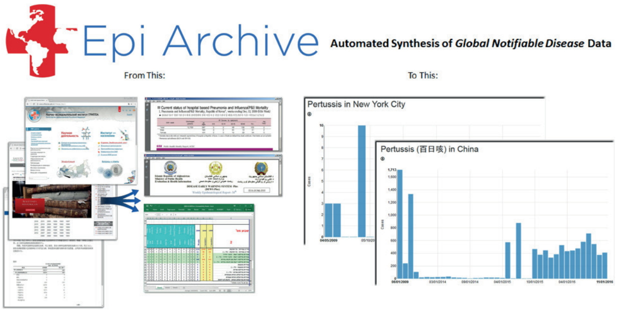
Persisting: We store the data in the Epi Archive database and make it available on the internet and through the BSVE. The data is persisted into a structured and normalized SQL database.

Results

The Epi Archive tool currently contains national and/or subnational notifiable disease data from twenty nations. When a user accesses the Epi Archive site, they are prompted with four fields: country, subregion, disease of interest, and date duration. Upon form submission, a time series is generated from the users' specifications. The generated graph can then be downloaded into a CSV file if a user is interested in performing personal analysis. Additionally, the data from Epi Archive can be reached through a REST API (Representational State Transfer Application Programming Interface).

Conclusions

LANL, as part of a currently funded DTRA effort, is automatically and continually collecting global notifiable disease data. While 20 nations are in production, more are being brought online in the near future. These data are already being utilized and will have many applications including improving the prediction and early warning of disease events.



Keywords

notifiable disease data; pdf tables; document extraction; data sharing; web scraping

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

[2] Nurminen, Anssi. "Algorithmic extraction of data in tables in PDF documents." (2013).

***Sergio Cordova**
E-mail: sergioc@lanl.gov