Epi Archive: Automated Synthesis of Global Notifiable Disease Data

Hari S. Khalsa, Sergio R. Cordova, Nicholas Generous, Prabhu S. Khalsa

A-1, Los Alamos National Laboratory, Los Alamos, New Mexico, United States

Objective

Automatically collect and synthesize global notifiable disease data and make it available to humans and computers. Provide the data on the web and 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

Government reporting of notifiable disease data is common and widespread, though most countries do not report in a machine-readable format. This is despite the WHO International Health Regulations stating that "[e]ach State Party shall notify WHO, by the most efficient means of communication available." [1] 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 can be hesitant to share data openly for a variety of reasons that include technical, political, economic, and motivational [2].

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 or graphical charts.

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

A survey of the national notifiable disease reporting systems is periodically conducted 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 to 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 design and write code to automate the downloading of 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 combine the Nurminen methodology of PDF table detection with in-house heuristics to find the desired data within PDF reports [3].

Parsing: We determine what to extract from each dataset and parse 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

Epi Archive currently contains national and/or subnational notifiable disease data from thirty-nine nations. When a user accesses the Epi Archive site, they are able to peruse, chart and download data by country, subregion, disease and time interval. Access to a cached version of the original artifacts (e.g. PDF files), a link to the source and additional metadata is also available through the user interface. Finally, to ensure machine-readability, the data from Epi Archive can be reached through a REST API. http://epiarchive.bsvgateway.org/

Conclusions

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

Acknowledgement

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

References

1. WHO International Health Regulations, edition 3. http://apps.who.int/iris/bitstream/10665/246107/1/9789241580496-eng.pdf
2. van Panhuis WG, Paul P, Emerson C, et al. 2014. A systematic review of barriers to data sharing in public health. *BMC Public Health*. 14, 1144. doi:https://doi.org/10.1186/1471-2458-14-1144. PubMed
3. Nurminen A. 2013. Algorithmic extraction of data in tables in PDF documents (Master's thesis). Retrieved from https://dspace.cc.tut.fi/dpub/bitstream/handle/123456789/21520/Nurminen.pdf.