SPA

The SPA biosurveillance application explores geographic and temporal trends in the reports published by ProMED-mail.

SPA scrapes and parses posts on the ProMED-mail website into a structured format and stores processed posts for future and further analysis. A toponym resolution algorithm extracts locations mentioned from the article bodies. The extracted data is presented in a user interface which allows visitors to filter and view data geographically.

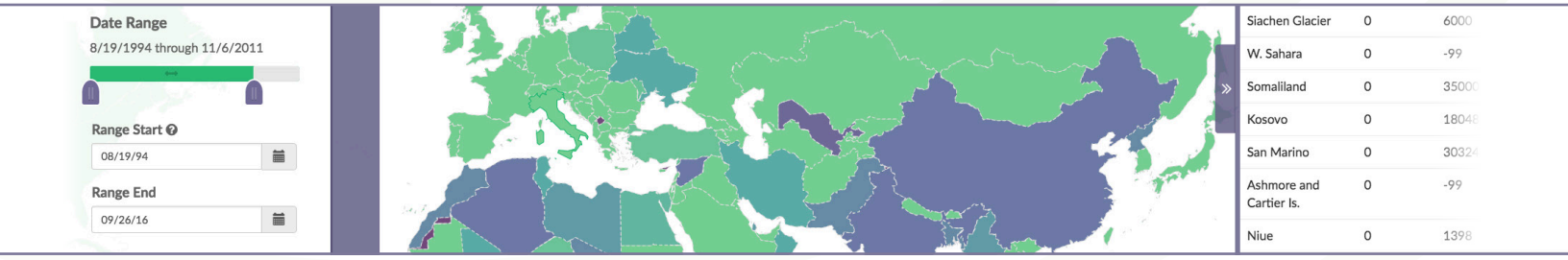
DATA ANALYSIS

The web-scraper periodically retrieves new posts from the main ProMED-mail feed, then using a number of heuristics based on the formatting conventions, it parses metadata such as publication dates and source names.

Place names mentioned within the bodies of articles in a ProMED post are identified by searching for them in the geonames.org database, then ambiguous or false-alarm place names are resolved based on characteristics of the location and document. Ambiguous place names are names that have multiple, distinct physical locations corresponding to them, like Columbia City. False-alarm names, are place names that in the context of the document are not actually place names. For instance, a document mentioning George Washington is not referring to Washington state when the word Washington is used. Both types of errors are mitigated by scoring candidate names across a number of criteria and only using those above a given threshold. Our scoring criteria includes features like the population of the location, the number of ways it is referred to, whether NLTK's named entity identifier identifies it as a location using its part of speech tag and context, and the proximity to other resolved locations.

INTERFACE

SPA's interactive choropleth map visualizes the number of mentions per capita of each country on ProMED-mail feeds. The interface includes a time range selection control that sets the time range data is shown form. By dragging the time range selection slider one can examine how the amount of reporting changes across the world over various intervals. The interface also include a panel where the precise number of mentions for each country can be viewed in a tabular format and downloaded in a spreadsheet.



RESULTS

SPA has analyzed the over 100,000 articles that appeared in ProMED-mail posts from the time of ProMED-mail's inception in 1994 to the current date. South America, Asia and Africa are mentioned less frequently in proportion to their populations than those in North America, Europe and Australia. All countries with a population over one million have some coverage. Uzbekistan has the least number of mentions per capita for an existing country with a population of over ten million.

spa.eha.io

For more information, contact:
Toph Allen
allen@ecohealthalliance.org

EcoHealth Alliance