Public Health Surveillance Using Global Health Explorer

James P. McCusker, Jeongmin Lee, Chavon Thomas, and Deborah L. McGuinness

Tetherless World Constellation
Department of Computer Science
Rensselaer Polytechnic Institute
110 8th Street Troy, NY 12180, USA
{mccusj,leej35}@rpi.edu, chavon.thomas@hws.edu, and dlm@cs.rpi.edu
http://tw.rpi.edu

Abstract. We demonstrate an early version of a semantic web tool, Global Health Explorer (GHX), that can be used to conduct public health surveillance using Twitter. Our infrastructure can use any controlled vocabulary to extract term uses in Twitter and supports hypothesis formation and exploration of data sets using visual analysis. The resulting data, gathered in RDF, makes it possible to analyze term usage through both temporal and spatial dimensions. GHX uses the qb.js framework to visualize and explore these data across dimensions, initially time and location. This allows users of GHX to monitor terms from pre-existing ontologies to conduct public health surveillance. We have prototyped the use of GHX to monitor terms from the NCI Thesaurus related to in uenza-like illnesses.

1 Introduction

Analysis of word usage relating to disease outbreaks is a growing eld within public health. Early success has been recognized in some visible projects including the Google Flu project. Google has been able to show that certain search terms are highly correlated with reports from the Center for Disease Control's In uenza Like Illness (ILI) Index [1]. However, Google Flu did not publish which terms they have identi ed, and it is not clear if the terms used are direct or indirect indicators of ILI. Lacking the term set makes it di cult for others to replicate results. We are exploring a more transparent approach and use Twitter as an open platform for mining and identifying indicators that can provide a similar kind of discussion. Monitoring twitter mentions from entire existing vocabularies may allow researchers to identify new terms that can serve as indicators for other diseases as well.

2 Architecture

The Global Health Explorer is composed of two components: qb.js[1] and Skitter[2]. qb.js is a general-purpose visualization and exploration environment for annotated RDF data. It facilitates visualization of multidimensional datasets expressed using the RDF Data Cube Vocabulary (QB). [2] This vocabulary supports descriptions of measures, annotations, and attributes of measured entities. We create what we call a Semantic Data Dictionary (SDD) using the QB vocabulary, types of measures (scalar, ordinal, nominal, or binary)[3], units of measure[4], and provenance [3] about how the measure is used. This SDD is what drives the use of data in a qb (pronounced cube) and allows the qb to understand the datasets that it is given to visualize and navigate. The overall architecture is shown in Figure 1.

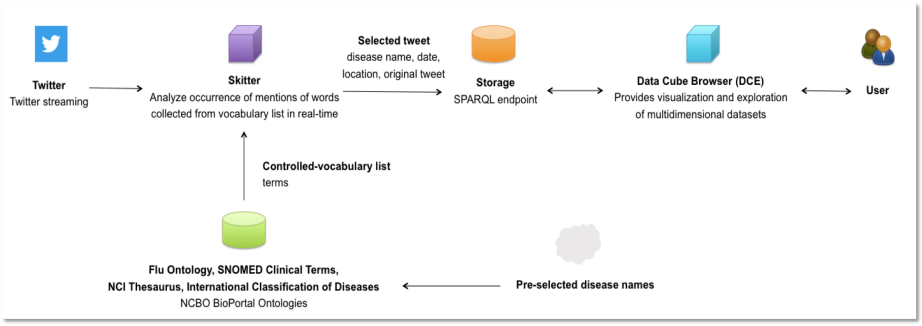

Fig. 1. The overall architecture of the Global Health Explorer. Skitter gathers tweets with terms matched from a controlled vocabulary and then stores them in a SPARQL endpoint. qb.js then queries the endpoint to provide a visualization of the data to users

qb.js uses a technique that we are beginning to call ontology-as-API, similar to the use of the term in Kohlhase et al [4]. We use RDFa in an HTML page to encode the con guration of the visualization with minimally su cient detail to be able to reconstruct it. In Figure 2 we show a starting con guration of qb.js using a particular dataset that is a named graph in a particular endpoint. This architecture allows us to embed useful information, including new SDDs for the data and provenance of visualizations and data that the current visualization has been derived from, into the HTML for the visualization itself. Any further manipulations of the visualization that are then saved and carry the same in-

[1] https://github.com/jimmccusker/qb.js
[2] https://github.com/leej35/Global-Health-Explorer.git
[3] Ontology of Experimental Variables and Values (OoEVV): http://purl.bioontology.org/ontology/OoEVV
[4] Measurement Units Ontology (MUO): http://forge.morfeoproject.org

formation along with it, simply by having it embedded in the RDFa that gets generated on export from qb.js.

(a)

```
<div typeof="http://semanticscience.org/resource/statistical graph"
      about="#dce" id="dce">
  <span class="config" style="display: none;">
  <a rel="http://www.w3.org/ns/prov#wasDerivedFrom"
      href="http://purl.org/twc/skitter"></a>
  <span typeof="http://rdfs.org/ns/void#Dataset"
        about="http://purl.org/twc/skitter">
    <a rel="http://rdfs.org/ns/void#sparqlEndpoint"
       href="http://localhost:3030/datacube/sparql"></a>
  </span>
  </span>
</div>
```

(b)

```
@prefix prov: <http://www.w3.org/ns/prov#>.
@prefix sio: <http://semanticscience.org/resource/>
<#dce> a sio:statistical graph;
     prov:wasDerivedFrom <http://purl.org/twc/skitter>.
<http://purl.org/twc/skitter> a void:Dataset;
     void:sparqlEndpoint <http://localhost:3030/datacube/sparql>.
```

Fig. 2. A minimum set of RDFa (a) needed to con gure a datacube. DCE uses SIO classes to refer to graphical elements, VOID classes and properties to describe datasets, and PROV-O to show how each uses the other. The id of the DOM node in the HTML de nes the element that the DCE visualization will be rendered into. (b) the same abstract RDF graph in Turtle.

Skitter is a tool that searches for mentions of words collected from potentially very large vocabularies. It accesses Twitter to analyze occurrences of terms from a controlled vocabulary in real-time. These concept mentions are then saved into a RDF Data Cube-compatible dataset that is then pushed at regular intervals to any SPARQL endpoint that supports SPARQL UPDATE. The time, user, original text, concepts mentioned, and if available, the location of the tweet is saved in the dataset. Skitter has successfully scaled up to more than 70,000 concepts from the NCI Thesaurus on the sample stream with the only delays occurring on loading and indexing the ontology, which only occurs when the program is started. Figure 3 shows the RDF representation of a tweet that has been extracted by Skitter.

Skitter is able to handle thousands of terms at once by building an in-memory search index of the ontology using Lucene. [5] We use the Snowball stemmer to allow for near matches of terms. Each concept becomes a Document in Lucene

with a URI eld and as many label elds as are asserted for the concept. As the hash-based index is only built once, on-line searches of terms in the index are constant-time relative to the size of the index. The only limiting factor is the size of the query. Since tweets are limited to 140 characters, this gives a very low upper bound on the time it takes to annotate a tweet. We have been able to successfully process and capture data from the statistical sample Twitter API using a modest laptop.

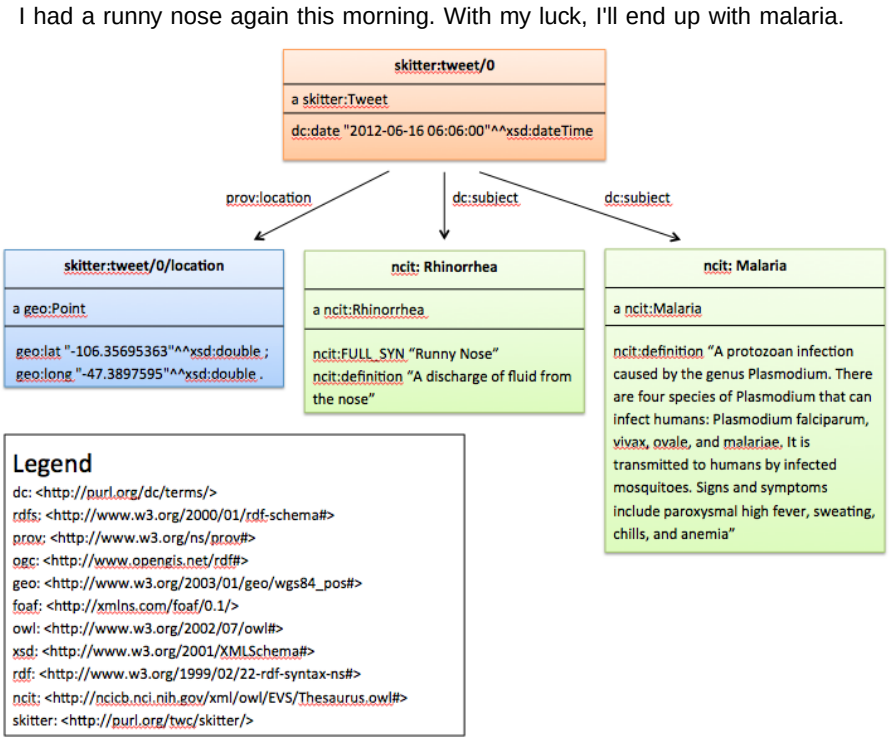

Fig. 3. A tweet (skitter:tweet/0) has been identi ed as having two terms as a subject, Rhinorrhea (runny nose) and Malaria. The tweet has a date (via Dublin Core Terms Date) and a location (prov:location and geosparql).

The Global Health Explorer uses qb.js, skitter, and terms from the NCI Thesaurus [6] and stores them into a single dataset that gradually aggregates in real time. Users can use the qb.js to see how certain terms, at any level, are mentioned in a particular area or time frame.

3 Discussion

The Global Health Explorer allows users to extract any mention of terms from very large vocabularies, such as the NCI Thesaurus, and build detailed datasets. Because many vocabularies are expressed as Linked Data, tools like qb.js can provide additional, on-demand information about the concepts mentioned in tweets simply by dereferencing their URIs. Additionally, use of QB and RDF makes it simple to view datasets in tools like qb.js This is applicable to research areas other than just in uenza research, as it can be used whenever discussion of a topic uses more concepts than can t in a conventional Twitter search. Research into the use of social networks like Twitter for public health surveillance has been very active. The Informatics Research and Development Laboratory (IRDL) recently published research about a web application called PHTweet that examined Twitter streams for a limited number of health terms.[5] Other research has introduced a technique that analyzes large quantities of the queries from the google search engine in order to target the regions of vast populations with in uenza epidemics and distinguish the communities with health-seeking habits. [7] In Figure 4 we show a GHX visualization using qb.js.

4 Conclusions

We have developed a prototype in which we can demonstrate how users can gather and analyze large data streams for occurrences of particular words (or phrases). These analyses may be useful to spot trends such as the emergence and growth of a disease outbreak. We mine social networks such as Twitter for term mentions and then map those terms into concepts that interpret the meaning of the tweet. We are working to allow researchers and members of the general community who would not ordinarily have access to public health surveillance data to use and access it. Although our Global Health Explorer does not act as a diagnostic system, we hope that it can serve as a new tool for researchers to access and analyze information that would be di cult to gather otherwise.

References

1. J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, Detecting in uenza epidemics using search engine query data. Nature, vol. 457, no. 7232, pp. 1012 4, 2009. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19020500
2. R. Cyganiak, D. Reynolds, and J. Tennison, The RDF Data Cube Vocabulary. [Online]. Available: http://www.w3.org/TR/vocab-data-cube/
3. T. Lebo, S. Sahoo, and D. McGuinness, PROV-O: The PROV Ontology. [Online]. Available: http://www.w3.org/TR/prov-o/

[5] http://edemo.phiresearchlab.org/PHTweet2

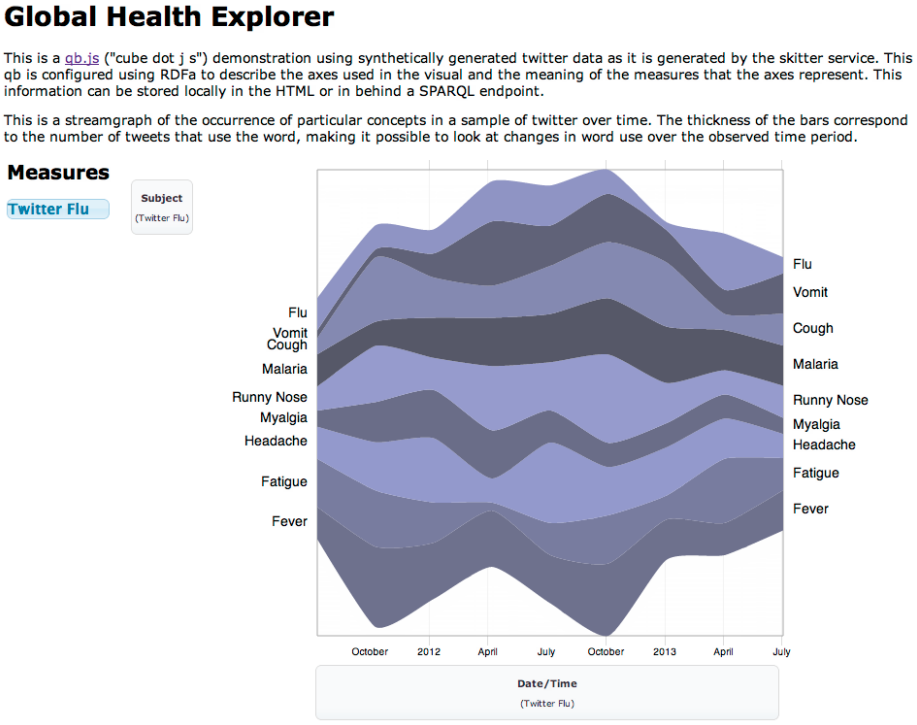

Fig. 4. This is a streamgraph of the occurrence of particular concepts in a sample of twitter over time. The thickness of the bars correspond to the number of tweets that use the word, making it possible to look at changes in word use over the observed time period. This demonstration is available at http://doppio.med.yale.edu/~jpm78/tw/qb.js/examples/twitter.html. This qb is using synthetically generated twitter data as it is generated by the skitter service. It is configured using RDFa to describe the axes used in the visual and the meaning of the measures that the axes represent. This information can be stored locally in the HTML or in a SPARQL endpoint.

4. M. Kohlhase, J. Corneli, C. David, D. Ginev, C. Jucovschi, A. Kohlhase, C. Lange, B. Matican, S. Mirea, and V. Zholudev, The Planetary System: Web 3.0 & Active Documents for STEM, Procedia Computer Science, vol. 4, pp. 598 607, 2011. [Online]. Available: https://svn.mathweb.org/repos/planetary/doc/epc11/paper.pdf
5. The Apache Software Foundation, Apache lucene, 2006. [Online]. Available: http://lucene.apache.org/
6. G. Fragoso, S. De Coronado, M. Haber, F. Hartel, and L. Wright, Overview and utilization of the nci thesaurus, Comparative and Functional Genomics, vol. 5, no. 8, pp. 648 654, 2004. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2447470&tool=pmcentrez&rendertype=abstract
7. G. Eysenbach, Infodemiology and infoveillance: Framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the internet, Journal of Medical Internet Research, vol. 11, no. 1, p. e11, 2009. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2762766&tool=pmcentrez&rendertype=abstract