Semantic Web Methodologies Provide Access to FLU Data Without Getting You Sick of Searching

Timothy Lebo, BS, MS[1], Joanne S. Luciano, BS, MS, PhD[1]
[1]Tetherless World Constellation (TWC), Rensselaer Polytechnic Institute, Troy, NY

Abstract

Epidemics and pandemics can place sudden and intense demands on health systems. One high-priority concern is that the highly pathogenic H5N1 avian influenza will spread to humans. This paper describes a RDF-based curation methodology to enable semantic search of influenza strain data.

Introduction

The influenza virus is responsible for the deadliest pandemic in human history, claiming 50 million lives worldwide in the “Spanish Flu” pandemic of 1918[1]. In the United States alone, 36,000 people die from seasonal flu each year[2]. Each year, public health researchers worldwide engage in diligent surveillance of emerging strains. For example, during the 2009 pandemic, North American H1N1 (swine flu) researchers needed to compare novel H1N1 strains to existing USA swine, USA human and USA avian strains. While a single geographic region is specified, the data did not support queries to address the hierarchical nature of geographic regions, forcing investigators to carry out several queries to identify emerging pandemic strains.

Open Linked Data, RDF and SPARQL

A principle benefit of semantic technologies is its explicit, actionable connections among data elements. While the health community may provide data resources for investigators, there is a steep learning curve to “know what is connected”, i.e., recognizing identifiers and knowing which site will return useful results. To demonstrate how semantic technologies ameliorates this “connection learning curve”, we demonstrate a search capability for a use case involving avian surveillance and nucleotide sequence data. First, data was manually extracted from fludb.org's website and stored to a data curation infrastructure. This data was organized according to a triad of provenance facets: the source (fludb-org), dataset identifiers (avian-surveillance, nucleotide sequence), and version identifiers (2010-Nov-30 and 2010-Dec-06). TWC's csv2rdf4lod[1] conversion utility was used to provide both verbatim and enhanced interpretations of the schema-less files from fludb.org and published in programmatically-accessible forms: 1) SPARQL endpoint, 2) Raw RDF dump files, and 3) Linked Data.

Once the initial semantic web representation for the fludb.org data is established, techniques can be used to provide augmentations beyond that available in fludb.org. A SPARQL query to the TWC's endpoint returns existing URIs for fludb.org's entities and the "strain name" property that contains an agglomeration of several entities and values (e.g., "A/American green-winged teal/California/HKWF-609/2007"). Using the same provenance-based organizational scheme described earlier (with source, dataset identifiers, and versions), this query response is incorporated as an additional dataset, converted using the same csv2rdf4lod converter, and hosted using the same three publishing forms as before.

During the conversion, the string “California” was promoted to a local URI, our:California, and asserted to be owl:sameAs California’s URIs in GeoNames, DBPedia, and GovTrack. This links the augmentation dataset to common identifiers and allows linking across datasets. A set of "lod-link" files enable this direct linking to existing URIs.

Relations from a previously accumulated dataset provide the final connection to identify strains within the US. The dataset included geospatial hierarchy data ("California", and "United States") that led to a geonames:parentFeature relation that can be obtained using the following query:

```
SELECT distinct(?parent) WHERE {
GRAPH <http://logd.tw.rpi.edu/source/fludb-
org/dataset/animal-surveillance/version/2010-Nov-30> {
   ?region  owl:sameAs ?link . }
GRAPH ?dataset {
   [] owl:sameAs ?link;
   geonames:parentFeature ?parent . }
```

Conclusion

A single SPARQL query leverages three data sources to select the strain information according to the regions associated within a geospatial hierarchy.

References

1. Tumpey, T.M. , Basler, C., Aguilar, P., Zeng, H. 2005. Characterization of the reconstructed 1918 spanish influenza pandemic virus. Science. 310:77–80 (2005).
2. CDC. Influenza: the disease. http://www.cdc.-gov/flu/about/disease/index.htm on 2/2/2009.

[1] https://github.com/timrdf/csv2rdf4lod-automation/wiki