Semantic Composition of Disparate Data in GeoHealthUS for Navigation, Display and Analysis

David Wood

GeoHealth US Corp, Arlington, VA 22209, USA
{david}@geohealth.us

Abstract. GeoHealth US Corp produces environmental monitoring data in the United States, and aggregates that data with historical environmental data from US Government sources. Data sources include five programs that track pollution history at the US Environmental Protection Agency, and air quality data collected by GeoHealth US Corp itself. Environmental monitoring data is combined with information from the US Department of Health & Human Services, including the US National Library of Medicine, to automatically associate environmental conditions with diseases and symptoms. Semantic Web techniques are used to perform the data integration, navigate the data for analysis, and to drive the display of contextually relevant data in a Web user interface.
GeoHealth US Corp is a Virginia Benefit Corporation (or "B-corporation"), a for-profit corporate entity that includes positive impacts on society and the environment in its core mission.

Keywords: Environment, Health, Environmental Data, Linked Data, Linked Data Platform, LDP, RDF, Linked Data, W3C, Callimachus, Semantic Web, US EPA, GeoHealthUS

1 Introduction

Human health is impacted by at least three classes of information: Lifestyle, genetics, and environment. Of these, environmental conditions remain the least poorly integrated in the US healthcare system. Environmental health information in the US is sparse, scattered, and in forms that make it difficult to find, to combine, and to relate to individual patients. In some cases, it is of questionable quality. We set out to address those problems in order to provide a more complete picture of the environmental impacts on US public health. The potential benefits are significant given the size of the US healthcare market: Approximately 3 trillion USD is spent annually on healthcare[2], with more than 1.3 trillion USD spent by government entities[1].

GeoHealthUS approached this problem by relating multiple US Government datasets in Linked Data formats (mostly from the US Environmental Protection Agency) with pseudo-real-time environmental air quality data collected by the company. Additional Linked Data was created to describe diseases and chemical

substances related to them. Such data was previously available only in traditional (XML) formats from the US Department of Health & Human Services.

The evolving GeoHealthUS Web site is available at `http://geohealth.us`. Individuals may currently view historic pollution report data while the newer environmental data is being integrated[1]. The site complies with the Linked Data Platform v1.0 specification[3], and is based on the Callimachus Project's Linked Data platform[4]. A SPARQL endpoint is available that allows for querying both RDF and non-RDF data, although it is currently restricted to authenticated users.

2 Current Air Quality Data

GeoHealthUS has created mobile sensor packages called GeoHealthBoxes™ to rapidly collect air quality information. GeoHealthBoxes are typically mounted on vehicles and measurements taken. GeoHealthBoxes use a combination of Open Hardware, software written by GeoHealth US Corp and a number of proprietary environmental sensors.

Newly collected environmental data from mobile sensors was considered too voluminous for RDF modeling to make sense. Instead, RDF summaries were created to facilitate the location, and querying of such data from larger relational databases. RDF and non-RDF data is presented in a single Web user interface. Contextual subsets of data are also presented for download in various reports and formats (such as Turtle, JSON-LD, and CSV). Contextual subsets of non-RDF data are converted to RDF at download time, as requested.

3 Historic Environmental Data

Historic environmental data from the US Environmental Protection Agency (EPA) was available from the programs summarized in Table 1.

Historical data from government sources represents a 25-year history of environmental pollutants, and some of their effects. All government data was either available in RDF formats or converted into RDF models for the purpose of simple composition. Vocabulary mapping of government data sets was undertaken as necessary.

The US EPA operates a Linked Open Data service in a quality assurance mode. This data service is not yet publicly available. The source data for the programs is generally available via `http://data.gov`, but in some cases EPA must be contacted directly to acquire information. This situation highlights the relative immaturity of US Government data sources when users desire to combine arbitrary data sets for further analysis and/or repurposing.

Table 2 list the namespaces of some of the common Semantic Web vocabularies used to represent the RDF portion of the data. Core vocabularies included the rdf, rdfs, owl, skos, and xsd namespaces.

[1] A prototypical, and temporary, interface for a portion of air quality data is available at `http://geohealth.us/home/pages/livedata.xhtml?view`

Table 1. EPA Programs Available as Linked Data

Abbreviation	Program Name	Purpose
FRS	Facilities Registry System	Facilities and locations
SRS	Substance Registry System	Chemical substances
TRI	Toxics Release Inventory	Air, water pollution reports
RCRA	Resource Conservation and Recovery Act	Solid & hazardous waste
CDR	Chemical Data Registry	Manufacturing, importation

Table 2. Common Vocabularies

Namespace	URI	Purpose
foaf	http://xmlns.com/foaf/0.1/	Nearness, depictions
geo	http://www.w3.org/2003/01/geo/wgs84_pos#	Locations
place	http://purl.org/ontology/places#	Locations
dbpedia	http://dbpedia.org/resource/	Units of measure, companies
vcard	http://www.w3.org/2006/vcard/ns#	Addresses

We made use of a number of RDF vocabularies specific to the EPA datasets, which include definitions such as the classification of facilities, substances, and reports. These vocabularies are under the base URI of http://opendata.epa.gov, but have not yet been formally published by the US EPA. We know of them through our prior work with the agency, and look forward to them being publicly available soon. Location information in EPA datasets was augmented by the use of a custom vocabulary created to represent US postal (ZIP) codes to facilitate the creation of maps and other geographic displays.

4 Relating Diseases and Symptoms

Additionally, a number of custom vocabularies were developed to represent information specific to the GeoHealthUS application, such as the description of diseases (extracted from traditional data formats available from the US Department of Health & Human Services). This data was mapped to chemical substances via SRS identifiers and to clinical findings using SNOMED-CT identifiers.

5 Data Architecture

Figure 1 presents a simplified architecture diagram for the GeoHealthUS cloud-based service. Updates to the historic data are periodically uploaded to the RDF databases. Air quality data collected directly by GeoHealthUS is uploaded to relational databases for summarization, analysis and querying. Callimachus acts as a data hub to dynamically generate both Web pages for human consumption and RDF- and CSV-formatted data extracts upon demand. The service is naturally distributed, and spread across many virtual machines.

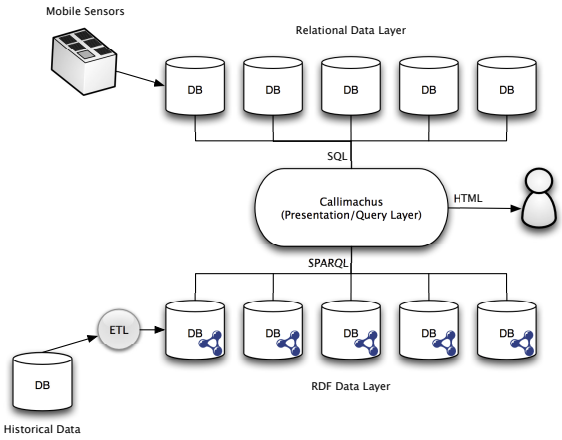

Fig. 1. Simplified architecture, showing dynamic data aggregation at Callimachus.

6 Conclusions

The GeoHealthUS poster illustrates three benefits of a Semantic Web approach to data integration:

1. Conversion to RDF formats facilitated data integration across many data sets by the simple mechanism of identifier alignment.
2. Presentation to end users is via a small number of SPARQL v1.1 queries, simplifying maintenance requirements, and reducing maintenance costs over traditional approaches.
3. The approach was shown to function in a large, real-world use case with significant economic potential.

References

1. Chantrill, C. US Health Care Spending. `http://www.usgovernmentspending.com/us\_health\_care\_spending\_10.html` accessed 30 June 2015.
2. Centers for Medicare & Medicaid Services. National Health Expenditures 2013 Highlights. `http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/Downloads/highlights.pdf` accessed 30 June 2015.
3. Speicher, S., Arwe, J., and Malhotra, A. (eds). Linked Data Platform 1.0. W3C Recommendation, 26 February 2015. Retrieved 19 April 2015 from http://www.w3.org/TR/2015/REC-ldp-20150226/.
4. Wood, D. and Leigh, J.: The Callimachus Project. Retrieved 19 April 2015 from http://callimachusproject.org.