Big Data Processing and Semantic Web Technologies for Decision Making in Hazardous Substance Dispersion Emergencies

Athanasios Davvetas[1], Iraklis A. Klampanos[1], Spyros Andronopoulos[2],
Giannis Mouchakis[1], Stasinos Konstantopoulos[1],
Andreas Ikonomopoulos[2], and Vangelis Karkaletsis[1]

[1] Institute of Informatics and Telecommunications,
NCSR 'Demokritos', Aghia Paraskevi 153 10, Greece
{tdavvetas,gmouchakis,iaklampanos,konstant,vangelis}@iit.demokritos.gr
[2] INRASTES, NCSR 'Demokritos', Aghia Paraskevi 153 10, Greece
{sandron,anikon}@ipta.demokritos.gr

Abstract. Emergencies that involve the release of hazardous substances into the atmosphere affects life and nature for several years. The timely and reliable estimation of the expected consequences on people and the environment facilitates informed decision making and timely response. Here, we demonstrate a tool that leverages Big Data and Semantic Web technologies to estimate the source location and the expected dispersion of the plume and to link this against geo-located data about people, infrastructure, industry and other production units, and any other information relevant to potential effects on the population and the environment.

1 Introduction and Motivation

Emergencies that involve the release of hazardous substances into the atmosphere may affect life and nature for several years. Such releases have occurred in the past, with the Chernobyl accident of 1986 being one of the most notable example. In the Chernobyl case, the accident was announced to the global community with significant delay, and was inferred due to readings and analysis by neighbouring countries. It subsequently affected most of Europe.

It is self-evident that in such an emergency, as well as in many less notable but sometimes equally hazardous cases, the timely and reliable estimation of the release origin and of the expected consequences facilitates informed decision making and timely response. The demonstration presented focuses specifically on the problem where no information is known about the release itself except from readings at monitoring stations. Under these circumstances, decision makers need a tool that uses measurements and atmospheric conditions to estimate the source location and the expected dispersion of the plume. Information about the dispersion is used to link against geo-located data about people, infrastructure, industry and other production units, and any other information relevant to potential effects on the population and the environment.

2   Demonstrated Technologies

The demonstration is deployed on the BDE Platform, a data management and processing environment that leverages Semantic Web technologies to handle the integration of heterogeneous data [1]. For the application demonstrated here, we used the platform's tools for executing models on HPC infrastructures and managing the resulting datasets and their provenance metadata, and its tools for integrating heterogeneous data and, specifically here, having a single SPARQL endpoint that federates multiple RDF and GIS data stores.

2.1   Managing Inverse Dispersion Modelling Datasets

Atmospheric dispersion models are computational codes that simulate the processes of transport and diffusion of air pollutants, as well as other physical processes that occur during dispersion, such as deposition on the ground and transformations (chemical reactions, radioactive decay, etc.) Dispersion model calculations are based on meteorological data. We use the NOAA HYSPLIT atmospheric dispersion model, which is known to work well with our weather data produced by the WRF atmospheric model [7]. Air concentrations are calculated on a 2-dimensional grid of $15km \times 15km$ with a temporal step of $1hr$.

The core idea is to use historical weather data in order to create a set of climatological regimes that represent the European climate. These are subsequently employed to pre-compute the dispersion patterns for a number of known locations of nuclear stations in Europe and to store these dispersion patterns in a Big Data infrastructure along with provenance metadata about the conditions and parameters used for the computation. In the event of an emergency, the application searches for pre-computed dispersion patterns computed under conditions and parameters that match the current weather data. These patterns are then used to estimate the source location based on pollutant concentration measurements and to predict future dispersion. This has been shown to be accurate enough for immediate response in the case of an emergency, before more accurate results can be computed several hours later by executing HYSPLIT on actual weather data and pollutant concentration measurements [3].

In the work demonstrated here, we build upon previous work on using the HDFS and Hive components of the BDE Platform to store and access weather data and the Cassandra component to store metadata [4]. That work is complemented with new BDE Platform components that implement pattern matching methods for identifying similar weather patterns.

2.2   Data Integration through Federated Querying

Semagrow is one of the core semantics-aware components of the BDE Platform, providing client applications with a uniform perspective of heterogeneous data stored in heterogeneous data management and processing infrastructures. Semagrow is a federated SPARQL query processing system that transparently selects relevant data sources, optimizes query plans, and applies the appropriate
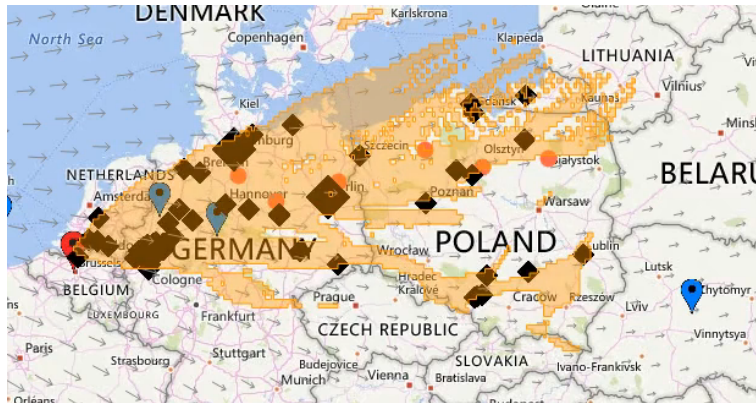
Fig. 1. Part of the UI of the demonstrated application. The dots indicate locations of radiation readings. The diamonds represent the populated areas affected by the plume. The arrows represent the current weather.

vocabulary transformations to hide schema heterogeneity [2]. The Semagrow execution engine also supports multiple query languages, including CassandraQL and SQL, again transparently serializing the query plan to the target store's query language and translating and joining the partial responses into the overall SPARQL query response [5].

In the application demonstrated here, Semagrow is used to link the cells of the modelling grid access via an stSPARQL endpoint [6] with population information from the Geonames dataset access via a SPARQL endpoint.

3 Description of the Demonstrated Application

A screen-cast of the application can be seen on https://vimeo.com/227245883 and the source code is is at https://github.com/big-data-europe/pilot-sc5-cycle3

The user can simulate the input from the fixed radiation detection network as well as from portable radiation detection devices. Then the user is asked to

Listing 1. SPARQL query template used in the demo. `?dispersCell` is binded with the dispersion model grid cell for which we retrieve more information.

```
SELECT ?lat ?long ?name ?population WHERE {
  ?dispersCell  <http://strdf.di.uoa.gr/ontology#hasGeometry> ?cell .
  ?populatedLoc <http://www.opengis.net/ont/geosparql#asWKT> ?point ;
                <http://www.w3.org/2003/01/geo/wgs84_pos#lat>  ?lat ;
               <http://www.w3.org/2003/01/geo/wgs84_pos#long> ?long ;
                     <http://www.geonames.org/ontology#name>  ?name ;
         <http://www.geonames.org/ontology#population> ?population .
  FILTER <http://strdf.di.uoa.gr/ontology#within>( ?point, ?cell )  }
```

select weather data, accessing real-time weather data is feasible. The application visualizes the weather using arrows. Based on this input, the application extracts predominant weather patterns and identifies the closest-matching pre-calculated dispersion modelling results to immediately show dispersion results for two major pollutants (Cs-137, I-131). The user can then choose the pollutant detected and the method for estimating source location. Given these choices, the application estimates and reports the three most likely pollution sources. The concentration plumes of each station is drawn on the map.

The map is also enriched with numerical information about the areas affected by the plume, and the user can filter these results by moving a slidebar. When showing population, in the example in the video, the slidebar sets the minimum population for showing an affected place on the map.

Acknowledgements

Bibliography

[1] Auer, S., Scerri, S., Versteden, A., Pauwels, E., Charalambidis, A., Konstantopoulos, S., Lehmann, J., et al.: The BigDataEurope platform — supporting the variety dimension of Big Data. In: Proc. 17th Intl Conference on Web Engineering (ICWE 2017), Rome, Italy. LNCS 10360. Springer (June 2017)

[2] Charalambidis, A., Troumpoukis, A., Konstantopoulos, S.: SemaGrow: Optimizing federated SPARQL queries. In: Proc. 11th Intl Conference on Semantic Systems (SEMANTiCS 2015), Vienna, Austria (September 2015)

[3] Klampanos, I., Pappas, C., Andronopoulos, S., Davvetas, A., Ikonomopoulos, A., Karkaletsis, V.: Identifying patterns in the weather of Europe for source term estimation. EGU General Assembly Conference Abstracts 19 (2017)

[4] Klampanos, I., Vlachogiannis, D., Andronopoulos, S., Cofiño, A., Charalambidis, A., Lokers, R., Konstantopoulos, S., Karkaletsis, V.: Towards supporting climate scientists and impact assessment analysts with the Big Data Europe platform. EGU General Assembly Conference Abstracts 18 (2016)

[5] Konstantopoulos, S., Charalambidis, A., Mouchakis, G., Troumpoukis, A., Jakobitch, J., Karkaletsis, V.: Semantic Web technologies and Big Data infrastructures: SPARQL federated querying of heterogeneous Big Data stores. In: Proc. of ISWC 2016 Demos and Posters Track, Kobe, Japan (2016)

[6] Kyzirakos, K., Karpathiotakis, M., Koubarakis, M.: Strabon: A semantic geospatial DBMS. In: Proceedings ISWC 2012, Boston, USA (2012)

[7] Stein, A.F., Draxler, R.R., Rolph, G.D., Stunder, B.J.B., Cohen, M.D., Ngan, F.: NOAA's HYSPLIT atmospheric transport and dispersion modeling system. Bulletin of the American Meteorological Society 96(12) (2015)