Implementation of the Fukushima Radiation LOD Framework

Isao KOJIMA, Yoshio TANAKA, Akihito NAKAMURA, Akiyoshi MATONO
Information Technology Research Institute
National Institute of Advanced Industrial Science and Technology(AIST)
Umezono1-1-4, Tsukuba-City, Ibaraki, 305-8568, JAPAN
{isao.kojima, yoshio.tanaka, nakamura-akihito, a.matono}@aist.go.jp

I. Overview of the Project

AIST has a project to construct a database service which provides radiation monitoring data of Fukushima area after the nuclear plant accident. This project is a part of a national project initiated by NRA(Nuclear Regulation Authority) and JAEA (Japan Atomic Energy Agency) manages the entire R&D project. AIST's role in this project is to develop database federation framework with open international standards. Radiation monitoring data is a huge set of various geospatial sensor data so that we are aiming to provide the data as geospatial LOD(Linked Open Data) along with popular OGC(Open Geospatial Consortium) standard services.

II. Problems

We had following development issues when constructing the framework.

A. *Providing Separate Standard Intefraces*

In geospatial community, OGC standards[1], such as WMS(Web Map Service) and SWE(Sensor Web Enablement) are very popular and useful. They provide a good set of specifications to handle geospatial data. However, they are not useful for federating with other non-spatial data such as statistics and other web data which is mostly handled as RDF/LOD environment. Providing these 2 kinds of standard specs for one single data is an issue.

B. *Supporting GeoSpatial Query and APIs*

We need to support various types of the radiation data such as real-time monitoring posts, car-survey and airborne survey. In order to access the data efficiently, query languages which can support geospatial search is indispensable. Simple APIs to access the data is also important.

C. *Converting Legacy CSVs to Linked Data*

Original radiation data[2] is provided with CSV format and the number of the CSV files is 239 which has approximately 17million records. These CSVs are categorized into slightly different 31 record types and integrating these CSV files into single database with single schema is also an issue. The real time monitoring data which we are trying to import now is quite larger than that and the efficient workflow to convert the data to the RDF/LOD is very important.

III. Approach and the Prototype System

A. *Software Configuration*

Our prototype system is public-domain based and we used Sesame[3] for triple stores. For query language, we constructed GeoSPARQL interface using uSeekM[4]. Elda LOD API[5] is also supported. Since Elda cannot support geometry operation, we are extending to support simple geospatial operations. We want to discuss the spec of the extension at the workshop.

For importing large number of CSV files, we created a web based batch processing tool called SOS importer. The tool crawls the CSVs of some URL into postGIS tables as an intermediate working database. This relational table is used to create OGC SOS XML data for 52north SOS server [6] and RDF triples to insert Sesame for radiation LOD. The feature of the importer is to make simple inference to unify the column names and store the mapping results for further integration with other CSVs. Interactive edit of the mapping can be also possible to adjust the automatic mappings.

B. *Application and the Schema Design*

Using this database, we constructed one showcase application to compute collective dose as in Fig.1. This application integrates one of our radiation database(airborne survey) with population statistics database provided by Ministry of Public Management(which is also provided by CSV format).

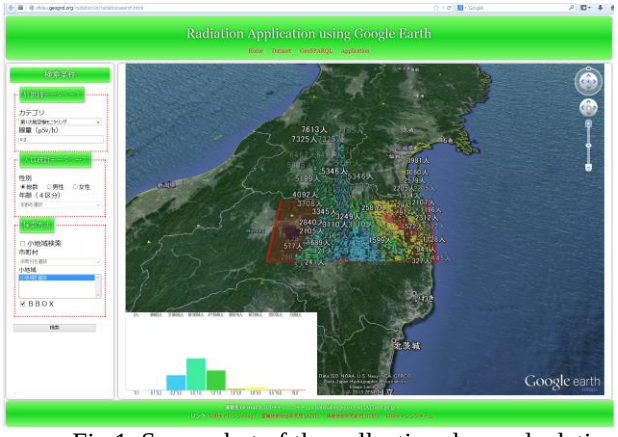

Fig.1. Screenshot of the collective-dose calculation application prototype. BBox search is performed with GeoSPARQL polygon operation.

Both radiation and population database is designed with RDF schema. These two databases have geometry information such as shape of the measured area and they can be integrated using the region class which has geometry classes as in Fig.2.

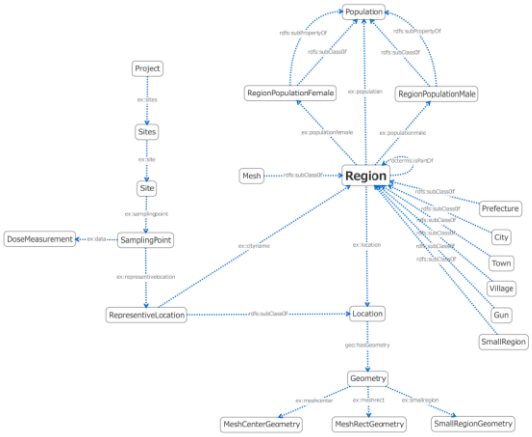

Fig.2. RDF Schema for Radiation Data and Population Statistics.

IV. ISSUES

Through the development of a prototype, following discussion issues are obtained. We want to discuss them in the workshop if possible.

- Slow Query Performance: The response time of GeoSPARQL queries is not sufficient for the interactive use. Our rough evaluation shows that it takes about 20 seconds for simple bbox search and over 3 minutes for complex geometry operations(details can be presented in the workshop). One reason for the slow is joining multiple RDF datasets(radiation and population) and we are trying to create additional triples to index the relationship between them. As in this example, we observe that the schema of many governmental dataset is not designed for the purpose of federating with other datasets. Thus, converting CSVs to RDF without any additional structure will not be suitable for the distributed LOD search.
- Large Number of Triples: Current radiation database includes 134,378 measurement points and the population database has 149,050 areas. However, total number of triples is 6,446,148 just to store the 1st monitoring set of airborne survey which is done at 2011.05.(Currently we have 6 monitoring datasets for airborne survey) The reason why number of the instance is very large will be investigated at the workshop but one reason is considered to have too many "part-of" relationships. This may also be the reason for slow response time.
- Dual Database Environment: We also support OGC SWE(Sensor Observation Service) for the same data and it is constructed with yet another public-domain software. Therefore, we cannot share the database between 2 database software although our intermediate database for importing CSVs is the same.

REFERENCES

[1] http://www.opengeospatial.org/
[2] http://radb.jaea.go.jp
[3] http://openrdf.org/
[4] https://dev.opensahara.com/projects/useekm/
[5] http://www.epimorphics.com/web/tools/elda.html
[6] http://52north/sos