SW at the department 'Environment' of the Flemish Government

Paul Hermans[1]

[1] ProXML, 3140 Keerbergen, Belgium

Abstract. This project proves that the semantic technology stack and related tooling allow to do data integration in a fast and agile way. Several technologies have been utilised: Ontology-Based Database Access, federated SPARQL, federation middleware. The solution is using the RDF Data Cube vocabulary for capturing the emission observations done. The additional 5 star LOD publishing was easily achieved at a minimal cost.

Keywords: Data integration, OBDA, federated search, LO(S)D.

1 Background

The 'Omgeving' Department is the environmental administration of the government of Flanders. It is responsible for preparing, following up and evaluating the Flemish environmental policy. In Belgium, companies that want to emit polluting substances in the air or water must have an environmental permit. Some of them are also obliged to report annually about the emissions of the previous year. In fact, data regarding the amount of substances per location have been collected since 2004.

The idea was to integrate these data with all kinds of other data sources (internal and external) to offer:

- the general public an application showing the emitted substances over the years in the area that they live
- companies the ability to benchmark themselves against other, similar, organisations
- public servants analytics dashboards to gain insights to steer the policy making.

The ultimate aim was that by integrating the data (research) questions can be addressed not being able to be answered on the separate data silos.

The solution preferably needed to be based on open source software or using open source libraries.

The project serves also as a pilot site of the OpenGovIntelligence project[1], receiving funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 693849, which aims to modernize Public Administration by connecting it to Civil Society through the innovative application of Linked Open Statistical Data (LOSD).

2 Solution

2.1 Solution outline

Dataintegration. The actual status is that more than 10 data silos have been integrated. These datasets contain data managed by the department itself (archival systems, RDBMS) and datasets published by other departments (Flemish addresses database, Belgian company register) together with well-known classification systems such as NACE (economical activities) and NIS (administrative geographical entities).

The solution takes a hybrid approach. For some datasets, the data are transformed via dedicated ETL processes into triples and these are loaded in a triple store. Other datasets, mainly from existing relational databases, are virtualized via OBDA[2] (Ontology-Based Data Access). The own ontology and controlled vocabularies will be managed with VocBench V3[3]. There are also connections to external SPARQL endpoints. In front of all these different SPARQL endpoints we offer a federation layer.

Data model. The most important datasets are in fact observations. Hence our use of the RDF Data Cube vocabulary[4]: a vocabulary explicitly made to capture statistical data to allow OLAP operations such as slice and dice, roll-up and drill-down. The code lists are encoded in XKOS[5], an extension of SKOS for statistical classifications.

LOD Publishing. All entities have their subject pages published using dereferenceable URL's. Next to this there are also public SPARQL endpoints available. But the important point is that the LOD publishing was not the first aim of the project: it was a nice free add-on.

General Public Dedicated Application. An application has been build showing the emitted substances over the years in the area that they live. This is a traditional web application using Polymer [6] web components, which is the standard in the department. This application is talking to the integrated set of data via the available SPARQL endpoints.

Business Intelligence. A connector has been developed for exploratory [7], a Data Science environment based on RStats [8], to connect with triple stores, so that BI reporting can be done using the integrated datasets.

2.2 Business Benefits

The approach used has clearly proven that semantic technologies allow to do data integration in a fast and agile way. Answers can now be given to questions which involve data from multiple datasets. The fact that 5 star LOD publishing is only one step away is a nice add-on. The system is used internally for the moment and will be become open to the public early November.

References

1. OpenGovIntelligence, http://www.opengovintelligence.eu/
2. Diego Calvanese, Giuseppe De Giacomo, Domenico Lembo, Maurizio Lenzerini, Antonella Poggi, Riccardo Rosati, Ontology-based database access, http://www.dis.uniroma1.it/~degiacom/papers/2007/sebd07.pdf
3. VocBench V3, http://vocbench.uniroma2.it/
4. RDF Data Cube Vocabulary, https://www.w3.org/TR/vocab-data-cube/
5. XKOS, Extended Knowledge Organization System, http://www.ddialliance.org/Specification/RDF/XKOS
6. Polymer Project, https://www.polymer-project.org/
7. Exploratory, https://www.exploratory.io/
8. RStats, https://www.r-project.org/