QBOAirbase: The European Air Quality Database as an RDF Cube

Luis Galárraga, Kim Ahlstrøm, and Katja Hose

Department of Computer Science, Aalborg University
{kah|galarraga|khose}@cs.aau.dk

Abstract. The Airbase is the European air quality dataset maintained by the Environmental European Agency. The dataset is available on the Web, and contains air quality monitoring data for 40 European countries. The multidimensional nature of the data makes it a good fit for OLAP (Online Analytical Processing) systems. Moreover, by linking the data to the Semantic Web, we can magnify its value, allowing for more sophisticated data analytics. In this paper, we introduce and describe QBOAirbase, a multidimensional provenance-augmented version of the Airbase dataset. QBOAirbase models air pollution data as an RDF cube, which has been linked to the YAGO and DBpedia knowledge bases.

1 Introduction

The Airbase[1] is the European air quality dataset maintained by the EEA (Environmental European Agency). The dataset is available on the Web, and contains air quality monitoring data for 40 European countries. The numerical and multidimensional data contained in the Airbase dataset can be efficiently handled by Online Analytical Processing systems (OLAP). Such systems are common in data warehousing or business intelligence scenarios, and are optimized to handle complex aggregation queries on multidimensional data with rare updates. Multidimensional datasets, known as data cubes, consist of a set of observations, e.g., measurements of the concentration of an air pollutant. These observations are described in terms of coordinates in a set of dimensions, e.g., time or location. Observations are the target of OLAP applications.

OLAP applications can benefit considerably from RDF and Linked Data [1–3, 8]. Thanks to the Linked Open Data initiative[2], resources from different datasets have been interlinked when they refer to the same real-world concept. Such a network of datasets constitutes what we call the *Semantic Web*, and allows us to see the Web as a giant knowledge base that can be queried, "understood", and analyzed by software agents. The interest in OLAP on the Semantic Web has been thrusted by the support for aggregation queries introduced in SPARQL 1.1 –the query language for RDF–, and the publication of the QB vocabulary [12] to model multidimensional data in RDF. This has motivated the publication of several datasets as RDF cubes[3].

[1] `https://www.eea.europa.eu/ds_resolveuid/3c756b2021754f6bba40447397d67fdf`
[2] `http://linkeddata.org/`
[3] `https://www.w3.org/2011/gld/wiki/Data_Cube_Implementations`

Keeping track of the origin of the data is crucial in a setting with multiple independent data providers. The provenance of a fact in an RDF data collection is metadata about the source and the data transformations that led to the publication of that fact. Such metadata finds application in scenarios such as data fusion or access control [4, 9]. While some RDF datasets have been augmented with provenance information [6], these still constitute a minority.

We present QBOAirbase, a multidimensional, linked and provenance-augmented version of the Airbase dataset in RDF. QBOAirbase represents air pollution data as a three-dimensional multilevel cube modeled with QB4OLAP [5] –an extension of QB that allows for multilevel dimensions. By linking QBOAirbase to YAGO [10] and DBpedia [7], we enrich the Airbase dataset with further information about the cities and countries of the monitoring stations, and the air pollutants. This opens the door to more sophisticated use cases in the analysis of air pollution data. We elaborate on the design of QBOAirbase in the following.

2 QBOAirbase

2.1 The Airbase Dataset

QBOAirbase is built upon version 8 of the Airbase dataset[1]. The original dataset is a collection of CSV and XML files containing annual concentration measurements for 238 air pollutants (e.g., SO_2, PM_{10}) in 40 European countries from years 1969 to 2012. Besides the actual measurements, the files also contain information about the data providers and the monitoring stations. For the latter, this includes the station's geographic location and technical details about the sensors' configurations. The data is accessible via a SPARQL endpoint[4]. Unlike QBOAirbase, this dataset is modeled with QB [12] instead of QB4OLAP [5] and provides neither provenance information nor links to existing knowledge bases. Furthermore, the SPARQL endpoint does not offer detailed documentation about the RDF schema and how to query the data.

2.2 Cube Structure

Observations. In QBOAirbase an observation maps to a measurement in the original Airbase dataset, that is, the aggregation of a set of measurements for a single air pollutant in an annual time span. QBOAirbase includes measurements for a list of 15 pollutants out of the 238 present in the original dataset. This is the minimal list of pollutants that a country must measure according to EU regulations. The original dataset considers 20 aggregation functions such as the annual mean, the maximum, the 50th percentile, among others. They are all considered in QBOAirbase.

Dimensions. An observation is characterized by its coordinates in the year, station, and sensor dimensions as Figure 1 shows. The edges in the figure define the *schema* properties, i.e., the RDF properties that connect the different levels of the cube structure. The station dimension contains three levels: station, city, and country. For some stations we did not have access to the information of the

[4] http://semantic.eea.europa.eu/sparql

city, hence those stations can only be rolled-up to the country level. We have manually linked the cities and countries in QBOAirbase to YAGO [10] and DBpedia [7]. The sensor dimension was artificially introduced and consists of two levels: sensor and component. The sensor level represents a sensor configuration and is described by a measurement unit, a type of equipment, a technique principle of the equipment, an aggregation function, etc. A sensor can be rolled-up to a component, which corresponds to an air pollutant (e.g., NO_2). The pollutants associated to the components have been manually linked to their corresponding YAGO and DBpedia resources as we did for cities and countries. Both the sensor dimension and the distinction between sensor and component allow us to model the fact that a station can provide measurements for the same pollutant under different sensor configurations, e.g., using a different aggregation function or measurement unit. Sensors and pollutants are instances of the classes *Sensor* and *Property* defined in the Semantic Sensor Network Ontology (SSN) [13]. SSN is a W3C candidate recommendation to describe sensors and measurement procedures.

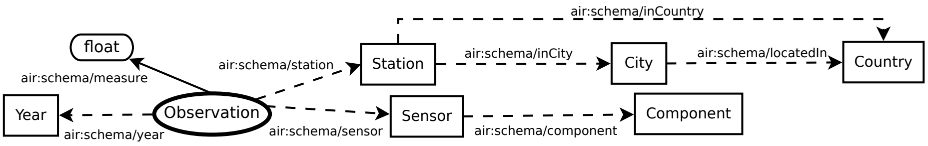

Fig. 1. QBOAirbase's cube structure. The prefix *air* is a shorthand for `http://qweb.cs.aau.dk/airbase/data/`.

2.3 Provenance

Each RDF triple in QBOAirbase is augmented with its workflow provenance [11]. In this model each triple is assigned an RDF resource, which we call its *provenance entity*. Such an entity models the processes that led to the generation of the triple, and is described with the W3C specification PROV-O [14]. A provenance entity can represent the source of a given measurement (e.g., a file), a schema mapping, or the result of a data operation such as a data extraction or a join. A data operation is modeled as an *activity* in PROV-O. Activities can produce or depend on provenance entities and they can directly or indirectly be carried by *agents*. These can be people, organizations, or even software agents.

3 Applications and Outlook

The Airbase's website provides an extensive list of reports in the form of figures, tables, interactive maps, and visualizations. Most of the reports can be generated by OLAP operations on the available data. However, some reports require additional data not present in the dataset. One example is the *percentage of urban population resident in areas where pollutant concentrations are higher than the recommend limit values*[5], which requires the population of the cities and

[5] `https://www.eea.europa.eu/data-and-maps/figures/urban-population-resident-in-areas-pollutant-limit-target`

the recommended concentration values of the pollutants. Such information can be obtained from another knowledge base. This justifies our decision of linking QBOAirbase to other sources in the Semantic Web. Moreover, the workflow provenance gives users more control on the data by e.g., restricting reports to data coming from particular institutions.

QBOAirbase is publicly available at `http://qweb.cs.aau.dk/qboairbase`. We also provide a SPARQL endpoint and a detailed documentation of the cube and provenance design. As Airbase, QBOAirbase is released under the terms of the ODC Open Database License (ODbL).

Acknowledgments

This research was partially funded by the Danish Council for Independent Research (DFF) under grant agreement No. DFF-4093-00301.

References

1. A. Abelló, O. Romero, T. B. Pedersen, R. Berlanga, V. Nebot, M. J. Aramburu, and A. Simitsis. Using Semantic Web Technologies for Exploratory OLAP: A Survey. *IEEE Transactions on Knowledge and Data Engineering*, 27(2), 2015.
2. Kim Ahlstrøm, Katja Hose, and Torben Bach Pedersen. Towards Answering Provenance-Enabled SPARQL Queries Over RDF Data Cubes. In *JIST*, 2016.
3. Alex B. Andersen, Nurefşan Gür, Katja Hose, Kim A. Jakobsen, and Torben Bach Pedersen. Publishing Danish Agricultural Government Data as Semantic Web Data. In *JIST*, 2015.
4. Tyrone Cadenhead, Vaibhav Khadilkar, Murat Kantarcioglu, and Bhavani Thuraisingham. A Language for Provenance Access Control. In *CODASPI*, 2011.
5. Lorena Etcheverry and Alejandro A. Vaisman. QB4OLAP: A Vocabulary for OLAP Cubes on the Semantic Web. In *COLD*, 2012.
6. Olaf Hartig. Provenance Information in the Web of Data. In *LOWD*, 2009.
7. Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web Journal*, 6(2), 2015.
8. Adriana Matei, Kuo-Ming Chao, and Nick Godwin. OLAP for Multidimensional Semantic Web Databases. In *BIRTE*. 2015.
9. Pablo N. Mendes, Hannes Mühleisen, and Christian Bizer. Sieve: Linked Data Quality Assessment and Fusion. In *EDBT/ICDT Workshops*, 2012.
10. Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. YAGO: A Core of Semantic Knowledge. In *WWW*, 2007.
11. Yannis Theoharis, Irini Fundulaki, Grigoris Karvounarakis, and Vassilis Christophides. On Provenance of Queries on Semantic Web Data. *IEEE Internet Computing*, 15(1), 2011.
12. Word Wide Web Consortium. The RDF Data Cube Vocabulary. `https://www.w3.org/TR/vocab-data-cube/`, 2014.
13. Word Wide Web Consortium. Semantic Sensor Network Ontology, W3C Candidate Recommendation. `https://www.w3.org/TR/vocab-ssn/`, 2017.
14. World Wide Web Consortium. PROV-O: The PROV Ontology. `http://www.w3.org/TR/2013/REC-prov-o-20130430/`, 2013.