QBOAirbase: 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. AirBase is the European air quality dataset maintained by the EEA (Environmental European Agency). The dataset is publicly 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. These systems are optimized for complex queries with grouping on multidimensional data, which are common in data analysis. Moreover, by linking the data to existing knowledge bases in the Semantic Web, we can magnify its value, allowing for more sophisticated data analytics. In this paper we introduce and describe QBAirbase, a multidimensional provenance-augmented version of the Airbase dataset. QBOAirbase represents air pollution information as an RDF data cube, which has been linked to the YAGO and DBpedia knowledge bases.

1 Introduction

AirBase[1] is the European air quality dataset maintained by the EEA (Environmental European Agency)[2]. The dataset is publicly available on the Web, and contains air quality monitoring data for 40 European countries, including all members of the European Union. The numerical and multidimensional nature of this dataset makes it a good fit for Online Analytical Processing systems (OLAP). Such systems are common in data warehousing and business intelligence scenarios, and are optimized to handle very complex aggregation queries on multidimensional data with rare updates. Multidimensional datasets, often referred to as data cubes, consist of a set of observations, e.g., measurements of the concentration of an air pollutant. These observations are the target of OLAP applications and are described in terms of coordinates in a set of dimensions, e.g., time or location. Data cubes and OLAP systems find applications in data analytics tasks such as reporting and data mining.

OLAP applications can highly benefit from RDF and Linked Data [1,2,4,9]. RDF [14] (Resource Description Framework) is the W3C standard to describe resources, i.e., concepts, in a domain of knowledge. RDF describes those concepts via facts in the form of triples such as ⟨`air:Denmark`, `air:capital`, `air:Copenhagen`⟩. There are thousands of large RDF data collections available in

[1] http://www.eea.europa.eu/data-and-maps/data/airbase-the-european-air-quality-database

[2] http://eea.europa.eu

the Web, about a wide range of domains such as general knowledge, life sciences, media, governmental data, etc. Thanks to the Linked Open Data initiative[3], resources from different datasets have been interlinked whenever they refer to the same real-world concept. Such a network of datasets constitute 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 computer programs. The interest in OLAP on the Semantic Web have been thrusted by the support for aggregation queries introduced in SPARQL 1.1 [15]—the query language for RDF—, and the publication of the QB vocabulary [16] to model multidimensional data in RDF. This has motivated the publication of some datasets as RDF cubes[4].

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 applications in scenarios such as data fusion or access control [3, 11]. While some RDF datasets have been augmented with provenance information in the Web [6], these still constitute a minority.

In this paper we introduce QBOAirbase, a linked, provenance-augmented, and multidimensional version of the Airbase dataset using RDF. QBOAirbase models air pollution data as a three dimensional cube, where each pollution measurement corresponds to a cell in the time, space and sensor configuration dimensions. By linking QBOAirbase to the YAGO and DBpedia knowledge bases, 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. In addition, QBOAirbase provides information about the sources of the data, as well as the corresponding ETL (Extract-Transform-Load) processes that led to the generation of the RDF triples. The remainder of the paper is structured as follows. Section 2 provides a deeper look at RDF, linked data, provenance, and data cubes. Section 3 describes our contribution, QBOAirbase. Section 4 sketches some use cases for QBOAirbase. Section 5 concludes the paper.

2 Preliminaries

2.1 RDF and Linked Data

At the conceptual level an RDF data collection or RDF dataset is a set of triples $t = \langle s, p, o \rangle \in (\mathcal{I} \cup \mathcal{B}) \times \mathcal{I} \times (\mathcal{U} \cup \mathcal{B} \cup \mathcal{L})$ where s is the subject, p is the predicate, and o is the object, e.g., ⟨`ex:Denmark, ex:area, "43,094"^^xsd:decimal`⟩. The sets $\mathcal{U}$, $\mathcal{B}$ and $\mathcal{L}$ are countably infinite sets of IRIs, blank nodes and literals. IRIs, e.g., *ex:Denmark*, identify concepts (also called resources) uniquely across datasets. A concept is associated to a set of classes, e.g., *ex:Denmark* is an instance of the class of countries, i.e., ⟨`ex:Denmark, rdf:type, ex:Country`⟩. Blank nodes are anonymous resource identifiers defined at the scope of a dataset. Literals represent constants such as numbers, e.g., *"43,094"^^xsd:decimal*, strings and dates. Since an RDF dataset can be naturally represented as a labeled graph

[3] http://linkeddata.org/

[4] https://www.w3.org/2011/gld/wiki/Data_Cube_Implementations

where the subjects and objects are nodes, and the predicates are labeled edges, a dataset is usually referred as an *RDF graph*.

2.2 Provenance in RDF

The provenance of a unit of information is metadata about its origin. There exist multiple provenance models for RDF data in the literature [13], however, we focus on *workflow provenance* in this paper. This is provenance about the source and the processes that led to the generation of an RDF triple. In this model each triple is assigned an RDF resource, which we call its *provenance entity*. Such an entity is described with the PROV ontology [10], the W3C specification to describe provenance entities. In PROV-O a provenance entity can represent a data source such as a file, a web service or the intermediate result of a data transformation process. A data transformation is modeled as an *activity* in PROV-O. Activities can be directly or indirectly carried by *agents*. These can be people, organizations, or even computer programs. The addition of provenance information to the triples of an RDF dataset produces a set of quadruples. The collection of triples that describe the provenance entities (the fourth component of the quadruple) is called the *provenance graph* of the RDF dataset.

2.3 RDF Cubes and OLAP

An RDF cube is a traditional data cube described in RDF, and consists of a set of observations—cells in the cube metaphor—. An observation is an RDF resource that represents a measurement. An observation can have zero, one or many measurements, each one described by an RDF predicate, which we call a *measure*. In QBOAirbase each air pollutant corresponds to a measure, and a cell stores a single measurement from a given monitoring station, at a given time, and under certain sensor configurations. These attributes define the coordinates of the cell in the location, time and sensor dimensions. Moreover, these dimensions can be described at different levels of granularity. For example, a measurement of 52.371 $\mu g/m^3$ of PM_{10}[5] particles can be described by its measurement station at coordinates 57°05'34"N 9°50'57"E, but also by its city, i.e., the city of Nørresundby in Denmark. It follows that in this example, the space dimension has three levels: station, city and country, where each level has a many-to-one relationship with its succesor. This hierarchy allows us, for example, to provide a summarized report per city by aggregating the values reported by the stations located in the same city. This operation is known as a *roll-up* in the OLAP literature [7]. In the cube metaphor, this is equivalent to merging sets of cells along one dimension whenever they are at the same level. In addition, each dimension level can have attributes that further describe the level. For example, the city of Nørresundby has a population and a set of postal codes. Level attributes allow for further constraints in OLAP. One may want to calculate the maximum concentration of PM_{10} particles in Denmark for cities with more than 80,000 inhabitants in 2012. Such an analysis can be defined in terms of (1) a *roll-up* from station to city, (2) a *slice* operation that removes the time dimension by

[5] Inhalable coarse particles

fixing it to the year 2012, and (3) a *dice* operation that picks those cities with more than 80,000 inhabitants.

3 QBOAirbase

In this section we describe the structure of QBOAirbase along the lines of its cube structure and provenance.

3.1 The Airbase Dataset

QBOAirbase is built upon the version 8 of the Airbase dataset[6]. The original dataset is a collection of concentration measurements for 238 air pollutants in 40 European countries from years 1969 to 2012. Examples of such pollutants are sulfur dioxide (SO_2), inhalable coarse particles (PM_{10}), and lead in PM_{10} (Pb). The dataset is split by country. For a single country the data is spread in three CSV files: stations, statistics, and measurement configurations. The file *stations* contains the complete list of monitoring stations in a given country. The file *statistics* contains the values of the measurements delivered by the sensors in the station. The stations provide measurements at heterogeneous time intervals, e.g., some provide values every hour, others provide a value every three days, and some others do not even have a fixed time interval. Due to this heterogeneity, the dataset contains aggregated values for measurements across a calendar year using different functions of central tendency. These functions include the annual mean, the 50th percentile, the maximum, etc. Finally, the file *measurement configurations* describes technical details of the methods and configurations of the sensors such as the measurement unit, e.g., $\mu g/m^3$, or the measurement technique principle, e.g., x-ray emission. The schema of the files is available at the dataset's website. The dataset is also available as an RDF cube using the QB vocabulary [16] via a SPARQL endpoint[7]. Unlike QBOAirbase, this dataset provides neither provenance information nor links to existing knowledge bases.

3.2 Cube Structure

We resorted to the QB4OLAP vocabulary [5] to describe the cube structure of QBOAirbase. QB4OLAP [5] is an extension of the QB vocabulary [16] that provides support for multilevel dimensions and user-defined aggregation functions. QBOAirbase uses the namespace *http://qweb.cs.aau.dk/airbase/data/* (abbreviated *air:*) for the IRIs of the RDF resources in the cube. IRIs have prefixes of the form *air:[type]/*, where *type* can be replaced by any of the following values: observation, year, station, city, country, sensor, or component.

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[8]. The original dataset considers 20 aggregation functions such as

[6] `https://www.eea.europa.eu/ds_resolveuid/3c756b2021754f6bba40447397d67fdf`

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

[8] `http://ftp.eea.europa.eu/www/AirBase_v8/airbase_v8_products.pdf`

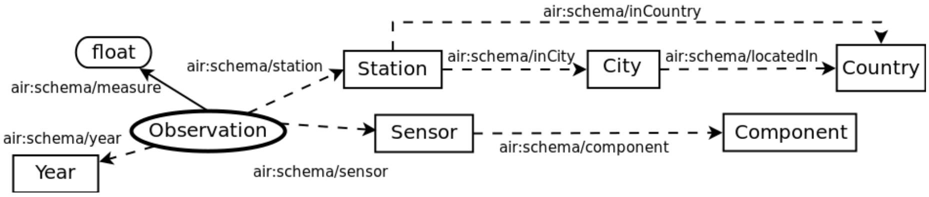

Fig. 1. QBOAirbase's cube structure.

the annual mean, the maximum, the 50th percentile, among others. They are all provided in QBOAirbase. Each pollutant is associated to a measure. For example, an annual average of 52.371 $\mu g/m^3$ of PM_{10} particles is represented as a triple ⟨`air:observation/`*CCx*`, air:schema/pm10, "52.371"^^xsd:decimal`⟩ where CC is the country's ISO code, and x is replaced by a unique integer identifier assigned by our data generation tool. Observations are instances of the class *qb:Observation* defined in the QB vocabulary.

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* predicates, i.e., the RDF predicates 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 city, hence the stations can be only rolled-up to the country level (via the predicate *air:schema/locatedIn* in Figure 1). We have manually linked QBOAirbase with the YAGO [12] and DBpedia [8] knowledge bases by reusing the YAGO and DBpedia identifiers for cities and countries. 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 its corresponding YAGO and DBpedia entities as we did for the cities and countries. Both the sensor dimension, and the distinction between sensor and component allows us to model the fact that a station can provide measurements for the same pollutant under different sensor configurations, e.g., using a different measurement unit or aggregation function. Table 1 shows the list of measures as well as the attributes of each dimension level of QBOAirbase. The underlined attributes define the primary key for the resources of each type and are used to construct the IRIs of the RDF resources. For example IRIs for instances of type sensor have the form *air:sensor/[id]* where id is the concatenation of the fields in the primary key, namely the station european code, the component code, and the measurement european group code.

3.3 Provenance

Each triple in QBOAirbase is augmented with a provenance entity that describes its provenance. This provenance is represented in RDF and PROV-O [10]. We

Dimension levels	**Attributes**
Measures	s:SO2, s:SPM, s:PM10, s:BS, s:O3, s:NO2, s:NOX, s:CO, s:Pb, s:Hg, s:Cd, s:Ni, s:As, s:C6H6, s:PM2.5
Station	p:europeanCode, p:station, p:localCode, p:establishedDate, p:shutDownDate, p:type, p:ozoneClassification, p:areaType, p:ruralSubType, p:streetType, p:longitudeDegree, p:latitudeDegree, p:altitude, p:localAdministrativeUnitLevel1Code, p:localAdministrativeUnitLevel2Code, p:localAdministrativeUnitLevel2Name, p:localAdministrativeUnitLevel1Code, p:isEuropeanMonitoringEvaluationProgramme
City	p:city
Country	p:isoCode, p:country
Year	p:yearNum
Sensor	p:europeanCode, p:code, europeanGroupCode, p:startDate, p:endDate, p:automaticMeasurement, p:measurementTechnique, p:equipment, p:samplingPoint, p:samplingTime, p:calibrationMethod
Component	p:component, p:code, p:caption, p:europeanGroupCode, p:unit

Table 1. Attributes of the dimension levels of QBOAirbase. The prefix *s:* is a shorthand for *air:schema/*, whereas *p:* means *air:property/*. Underlined characters define the dimension level's primary key.

distinguish two types of RDF triples in QBOAirbase: metadata and information triples. The *metadata* triples are facts generated by our data transformation tool in order to be compliant with the QB and QB4OLAP vocabularies. The metadata triples include *rdf:type* and *qb:memberOf* statements, as well as the triples with schema predicates (those with dashed lines in Figure 1). Such triples define the coordinates of an observation according to our proposed cube structure. We assign all metadata triples the single provenance entity *air:metadata/*. On the other hand, we call *information* triples those facts with predicates from Table 1, i.e., level attributes and measures. Information triples are assigned provenance entities with IRIs of the form *airprov:[sourceDType]2[targetDType]/[type]/[id]*, where *airprov:* is a shorthand for *air:provenanceIdentifier/*, *sourceDType* is the object's data type in the original Airbase schema, *targetDType* is the object's data type in QBOAirbase, *type* is the subject's type (e.g., observation, station, etc.) and *id* is an identifier. For triples about level attributes, this identifier is the subject's unique id defined in Section 3.2, whereas for observations the identifier is the concatenation of the country code and the tuple's line number in the source file. This source file is an intermediate file computed as the natural join of the original Airbase files, i.e., *stations*, *statistics*, and *measurement configurations*. Provenance entities model the data workflow carried out to generate the information triple in QBOAirbase, and are described in the provenance graph. Figure 2 shows part of the provenance of the triple ⟨`air:observation/DK1, air:schema/SO2, "49.582"^^xsd:decimal`⟩, a measurement of SO_2 in Denmark with provenance entity *airprov:Decimal2Decimal/DK2*. The provenance entity is highlighted, and is defined as the result of a conversion routine (decimal to decimal) from the source to the target schema. This conversion routine is modeled as an activity, i.e., *air:createTriple/DK60* in Figure 2. This activity

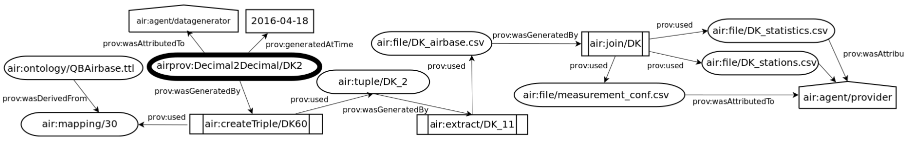

Fig. 2. Part of the provenance flow described by a provenance entity associated to an information triple in QBOAirbase. Entities are depicted as ovals, activities as rectangles and agents as pentagons.

relies on two intermediate provenance resources: *air:tuple/DK2* which models the tuple at line 2 in the source file, and *air:mapping/30* which defines one of the mappings from fields in the Airbase schema to RDF predicates in QBOAirbase. We wrote those mappings manually. The source file is constructed from the files provided in the Airbase website and are associated to the organizations that provided the data (not shown in Figure 2). The organizations are modeled as agents, as well as our data generation tool.

3.4 Publication

QBOAirbase is publicly available at `http://qweb.cs.aau.dk/airbase` as a set of N-quads files, one per country. We also provide a SPARQL endpoint, the data generation tool, the cube structure, and an extended version of Figure 2 with all the details of the triples' provenance. As the original Airbase dataset, QBOAirbase is released under the terms of the Open Data Common License.

4 Applications

The Airbase's website provides an extensive list of reports in the form of figures, interactive maps, visualizations and analyses. All those reports can be modeled as OLAP operations on cubes. They provide, for example, comparative analyses of the level of air pollution across different cities, rural areas, and countries throughout time. Most of the reports can be generated with the available data. Nevertheless, some reports require additional information not present in the dataset. One of those is the "Percentage of urban population resident in areas where pollutant concentrations are higher than the recommend limit values"[9], which requires to fetch information about the population of the cities and the recommended concentration values for the different pollutants. Such information can be obtained from a knowledge base. This justifies our decision of linking QBAirbase with YAGO and DBpedia. This could serve many other use cases, e.g., one could analyze the evolution of the concentration of certain pollutant in large urban areas, or spot correlations between air pollution and welfare indicators for countries. This opens the door for more advanced data analytics.

5 Conclusions

In this paper we have presented QBOAirbase, a multidimensional provenance-augmented version of the Airbase dataset using Semantic Web technologies such

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

as RDF, QB4OLAP, and PROV-O. QBOAirbase's multidimensional design facilitates the utilization of the data by OLAP systems. These are systems optimized for advanced data analytics. In addition, QBOAirbase has been linked to two knowledge bases in the Semantic Web, namely YAGO and DBpedia. As we showed, linked data broadens the applicability of the data to more sophisticated use cases in data analytics.

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. 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.
3. Tyrone Cadenhead, Vaibhav Khadilkar, Murat Kantarcioglu, and Bhavani Thuraisingham. A language for provenance access control. In *Conference on Data and Application Security and Privacy*, 2011.
4. Lorena Etcheverry and Alejandro Vaisman. Enhancing OLAP analysis with web cubes. *The Semantic Web: Research and Applications*, 2012.
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 *Linked Open Data Workshop*, 2009.
7. Christian S. Jensen, Torben Bach Pedersen, and Christian Thomsen. *Multidimensional Databases and Data Warehousing*. Synthesis Lectures on Data Management. Morgan & Claypool Publishers, 2010.
8. 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.
9. Adriana Matei, Kuo-Ming Chao, and Nick Godwin. Olap for multidimensional semantic web databases. In *Enabling Real-Time Business Intelligence*. 2015.
10. Deborah McGuinness, Timothy Lebo, and Satya Sahoo. PROV-o: The PROV ontology. W3C recommendation, W3C, April 2013. http://www.w3.org/TR/2013/REC-prov-o-20130430/.
11. Pablo N. Mendes, Hannes Mühleisen, and Christian Bizer. Sieve: Linked data quality assessment and fusion. In *EDBT/ICDT Workshops*, 2012.
12. Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. YAGO: A Core of Semantic Knowledge. In *WWW*, 2007.
13. Yannis Theoharis, Irini Fundulaki, Grigoris Karvounarakis, and Vassilis Christophides. On Provenance of Queries on Semantic Web Data. *IEEE Internet Computing*, 15(1):31–39, 2011.
14. Word Wide Web Consortium. RDF Primer (W3C Recommendation 2004-02-10). `http://www.w3.org/TR/rdf-primer/`, 2004.
15. Word Wide Web Consortium. SPARQL Query Language for RDF. `https://www.w3.org/TR/rdf-sparql-query/`, 2008.
16. Word Wide Web Consortium. The RDF Data Cube Vocabulary. `https://www.w3.org/TR/vocab-data-cube/`, 2014.