Publishing Skopje Air Quality Data as Linked Data

Kostadin Mishev, Angjel Kjosevski, Nikola Kalemdzhievski, Nikola Koteli, Milos Jovanovik,
Kosta Mitreski, Dimitar Trajanov
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, Macedonia

***Abstract*— Publishing raw data as Linked Open Data gives an opportunity of data reusability and data understandability for the computer machines. Today, the air pollution problem is one of the biggest in the whole world. Republic of Macedonia, especially its capital Skopje, has big problems with the PM2.5 and PM10 particles in the air approved by several measurement stations positioned on several locations in Skopje. In this paper, we demonstrate the process of centralizing of all the data collected from different measurement stations in one database. Also, we enable interpolation of collected data providing information about the current air quality state in the area between the measurement stations using previously implemented eco models. Interpolated data is saved in the same database providing interfaces that transform saved data into four-star and five-star data, by reusing the existing ontologies from the domain and linking them to the physical places where the measurements were taken and the interpolations were calculated. As a use case scenario, we provide and heat map about the values from various pollutants in the areas in Skopje providing information about the regions that have problems with air pollution.**

Keywords*— *air quality, indicator, measurements, measurement interpolation, ecoinformatics, open data, linked data

I. INTRODUCTION

Linked Open Data and Semantic Web principles are the main contributors in realizing the idea of data reusability and data scalability [1][2]. It gives the opportunity of linking the information from various fields and enabling simple access to them. With this approach, data becomes understandable not only for humans, but also for computer machines [3].

On the other hand, the air pollution in Macedonia, especially in its capital Skopje, is one of the biggest problems that the citizens of Skopje have. It can be described as the pollution of the atmosphere with gases, or dust of solid materials, particulate matter as other substances whose amounts are constantly increasing [4]. This information is approved by the multiple measurement stations positioned on several locations in Skopje and its environment. Some of them offer public domains for data access.

In this paper, we demonstrate the process of transforming the collected data from several measurement stations into four- and five-star Linked Open Data. We aggregate all the data collected from the services provided by stations of The Ministry of environment and physical planning, The Institute of Eco informatics at Faculty of Computer Science and Engineering, and the CO_2 measurements provided by the project Skopje Green Route, into one centralized database. They provide different air quality indicators. Afterwards, we transform the collected data into four- and five-star data by reusing the existing ontologies from the domain, necessary for the transformation and annotation process and linking them to the physical places where the measurements were taken [5].

The network of monitoring stations is very important in urban environment because it provides information of actual quantity of air quality indicators. The main problem is impossibility of obtaining appropriate values for all points of interest. There are several reasons and as most important we can mention the price of the measurement stations. Consequently, only the most important points could be monitored. Air dispersions models are solving this problem by providing estimations and predictions of the pollutants in the air using mainly emissions and meteorological data. These models include on mathematical algorithms based on combinations of physical and chemical simulating the spread of pollutants in the air. We describe the process of interpolation of the air quality data obtained from the measurement stations. This process of interpolation is repeating on constant time intervals to obtain approximate values of all air quality parameters [6]. The number of measurement stations is upgradable and it is directly proportional with the accuracy of the approximations.

In the final section, we propose and demonstrate several use-case scenarios querying published data set and represent the results on a heat map providing information about the current pollution from various pollutants in the area of Skopje.

II. RELATED WORK

The problem with air pollution is emerging almost all urban cities in the world. It is estimated that worldwide, 2 million people and more than half of them are in developing countries, die every year from air pollution. By releasing of the air indicators from measurement stations as open data provide contribution of the public understanding and dialogue around far-reaching and potentially data-rich aspects of life in the city. Consequently, air quality measurement datasets are already part of the LOD cloud. Home Weather ontology is intended for weather phenomena and exterior conditions providing property hasAirPollution which express an index of air pollution depending on the current air quality measurement values (Fig. 1).

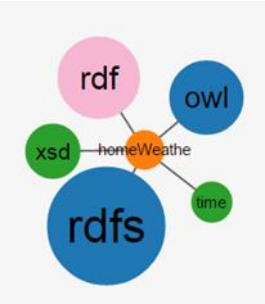

Figure 1. HomeWeather dataset

AirQuality+ is a project[1] which gathers real-time air quality measurements from different points in Sheffield, England, providing open licenses for communities and organizations to access and re-use. This includes near real-time data on pollutant levels in Sheffield collected by a network of monitoring stations, as well as data related to the issue of air quality, such as industrial activity, traffic and transport, public health, weather and land use.

The PESCaDO Ontology is a modular application ontology exploited[] for personalized environmental decision support, that enables to formally describe:

- the user decision support request
- the environmental data relevant to process the request
- the decisions and conclusions to be produced

The PESCaDO Ontology was thoroughly developed following state of the art best practices, and it is matched with a comprehensive and detailed documentation.

Air dispersion models are based on mathematical algorithms providing probabilistic values about the current air quality at each point of the city. They are related to the city infrastructure, the current weather conditions and the real-time measurements from several measurement stations. Currently, the Institute of Eco informatics has developed air dispersion models about the capital of R. Macedonia, Skopje. They use the measurements from The Ministry of environment and physical planning providing interpolation values for each point in the central region. These data is kept as 1 star data providing interpolated monochromatic visualizations (Fig. 2) on Skopje's map, generated from ArcGIS server. In this paper, we will convert the information from this visualization into 5 star data and we will provide useful information mining the gathered dataset.

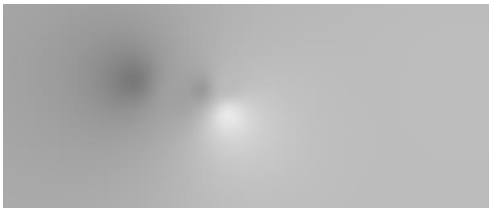

Figure 2. CO interpolated monochromatic visualization

[1] http://betterwithdata.co/portfolio/air-quality-plus/

III. Data from Skopje's Air Quality Measurement Stations

A. Centralizing data from multiple services

There are multiple measurement stations distributed in the region of Skopje providing different air quality indicators. Most of them, provide open access REST services which return real-time measurements. In our paper, we will use the services provided by:

- Ministry of environment and physical planning

The JSON service which provides measurement about CO, NO_2, PM10, PM2.5, SO_2, O_3 air quality indicators providing data refresh each hour during the day. There are multiple measurement stations over Macedonia, but for purposes of this paper, we will use only the stations located in Skopje: Centar, Karpos, Lisitche, Gazi Baba and Rektorat.

- Measurement CO_2 stations provided by the project Skopje Green Route

These measurement stations are placed on the most frequent crossroads in Skopje: Justice Palace, Red Cross and Faculty of Agriculture, providing measurement about CO2 air indicator refreshing the information each 5 minutes.

- Measurement station maintained by the "Laboratory of Eco informatics at Faculty of Computer Science and Engineering" providing information about the same air indicators like the measurements stations enabled by Ministry of environment and physical planning

All services are RESTful and provide open URL location which can be accessed with GET parameters. The log of all services is kept on our database which centralizes all information about all air quality indicators for all measurement stations (Fig. 3). It runs scheduled processes which poll the JSON services asking for new fresh data from sensors. It appends timestamp to the measurement information and saves in MySQL database whose EA diagram is represented on Fig. 4.

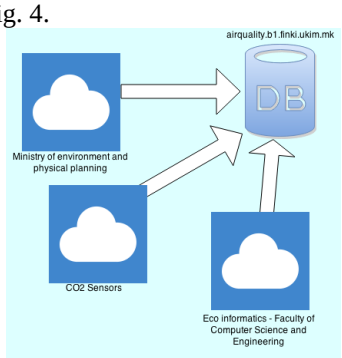

Figure 3 Centralized data polling architecture.

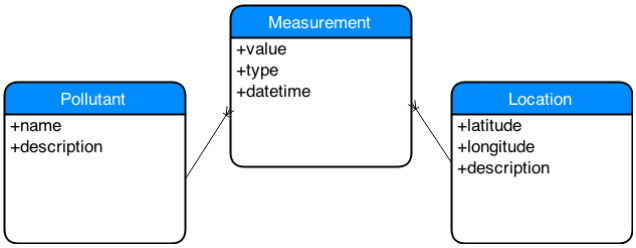

Figure 4. EA diagram of the centralized database.

B. *Interpolation of the measurements in the area of Skopje*

The process of interpolation is based on the newest "up to time" data as the average of the air parameter concentration per hour for each parameter. The data are provided by the network of measurement stations described in section A stored in one centralized database. They are used for generation of the grid raster layers by interpolating techniques. The model depends on the weather conditions and the infrastructure of the area taking as references the values from the nearest real weather and pollution measurement stations. The interpolation data is calculated on ArcGIS Simulation Server which implements the pollution model [6], gets the measurements from weather and pollution stations and provide interpolated information about the area of Skopje. The output of the model is a raster image, so we create algorithm for data transformation from raster monochromatic image to numerical format about the pollution state. The darker positions represent greater values of pollution. After the transformation process, we provide RESTful services which could be easily accessed by setting the latitude and longitude of the required position. This RESTful services is accessed by the interfaces of our application. They append appropriate timestamp and current weather conditions, and save in the centralized database.

C. *PESCaDO Ontology*

In order, to transform the measurement 3 star data, from the centralized database, into RDF, we need an ontology. Among multiple ontologies that we reviewed in our research, the PESCaDO Data ontology proved as the most useful for our needs. It is developed by Data & Knowledge Management research group [7] which is part of the Information and Technology Center in Fondazione Bruno Kessler and it is provided for mapping of measurement data from sensors of PM10, PM2.5, CO and other air pollutants. It also provides mapping of the weather conditions so we concluded that this ontology is satisfying our needs.

We divided the properties of the ontology in two main types: weather conditions properties (Table 1) and air pollution indices (Table 2).

Table 1. Weather condition properties

Property	Type	Description
HumidityValue	Datatype Property	Air humidity value
TemperatureValue	Datatype Property	Ambient Temperature value
WindSpeedValue	Datatype Property	Wind speed value

Table 2. Air pollution indices properties

Property	Type	Description
PM10IndexValue	Datatype	Provides information about the value of the PM10 particles in the air
PM2.5IndexValue	Datatype	Provides information about the value of the PM2.5 particles in the air
COIndexValue	Datatype	Provides information about the value of the CO concentration in the air
NO2IndexValue	Datatype	Provides information about the value of the NO_2 concentration in the air
SO2IndexValue	Datatype	Provides information about the value of the SO_2 concentration in the air
O3IndexValue	Datatype	Provides information about the value of the O_3 concentration in the air

D. *Geo Ontology*

To provide mapping of the geographical location of the measured or interpolated instance, we used the Geo ontology which is one of the most used (Fig. 5). We have the correct positions of the static measurement stations and we link the measurement with the exact position of the station. Afterwards, we divide the area of Skopje in zones providing interpolated information about each zone separately. Each zone has own latitude and longitude enabling linking to the appropriate instance from the Geo ontology [8].

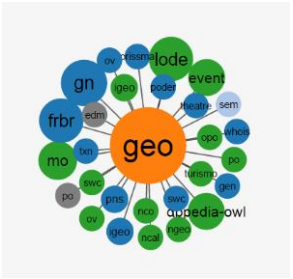

Figure 5. Geo ontology.

Table 3. Geo ontology

Property	Type	Description
Location	ObjectType	Description of the geographic entity
Lat	Datatype	Latitude of the mapped object
Lon	Datatype	Longitude of the mapped object

E. *Mapping the data from 3-star to 5-star data*

After defining the ontologies, we need to transform the data saved in database to RDF. In order to accomplish this, we

decided to use D2RQ server[2] which is compatible with MySQL databases and provides accessing relational databases as virtual, read-only RDF graphs without replicating into an RDF store. Using D2RQ we can query a non-RDF database using SPARQL, access to the content of the database as Linked Data over the Web, create custom dumps of the database in RDF formats for loading into the RDF store and provides access to a non-RDF database using the Apache Jena API.

The mapping process consisted of two steps. The first step provides wrapping of the relational database, in our case MySQL, with the interfaces provided by the D2RQ providing access to the stored data. Afterwards, we should define a mapping file using the D2RQ mapping language[3] to map the relational database schemas to RDF vocabularies and OWL ontologies. The mapping file defines a virtual RDF graph that contains information about the database. This graph contains RDF terms using d2rq:ClassMaps and d2rq:PropertyBridges. The class map specifies how URIs (or blank nodes) are generated for the instances of the class. It has a set of property bridges, which specify how the properties of an instance are created.

Our database, referencing to Figure 4, stores information about Pollutant, Location and Measurement. As the image represents, it is designed in 3rd normal form. As defined in section D, we use PESCaDO and Geo ontologies so we need to decompose the database in 2nd normal form providing the table pollutant be part of the measurement. To solve this problem, we change the mapping configuration using the D2RQ mapping language, so we need not to make any changes in the model of the relational database only by using the property d2rq:condition. The property d2rq:condition provides the SQL WHERE condition so an instance of this class will only be generated for database rows that satisfy the condition.

```
map:measurements_CO a d2rq:ClassMap;
        d2rq:dataStorage map:database;
        d2rq:uriPattern "measurement/@@T_MEASUREMENT.id@@";
        d2rq:class pescadoData:COIndexValue;
        d2rq:join "T_MEASUREMENT.pollutant_id =>
T_POLLUTANT.id";
        d2rq:condition "T_MEASUREMENT.pollutant_id = 3";
        d2rq:propertyDefinitionLabel "Measurements CO";
```

Transforming the data to 5 star data is provided by the geo:Location property linking the measurement information to specified Location where it is measured or interpolated. So, in the global graph, we provide air pollution indicator for the specified location.

F. USE CASE EXAMPLE

In this section we will demonstrate that transformation of data into Linked Data, can provide useful use-case scenarios. The result of use-cases gives opportunities for visual representation of the pollution on a heating map caused by all pollutants separately.

By following query, we can obtain information about CO measurements for the area of Skopje in determined time:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX pescado: https://ontohub.org/fois-ontology-competition/PESCaDO_Ontology/pescadoData.owl#
PREFIX prov: <http://www.w3.org/ns/prov#> .
SELECT DISTINCT ?lat ?lng  ?value WHERE {
?s rdf:type pescadoData:COIndexValue;
rdf:value ?value;
prov:atLocation ?location;
prov:generatedAtTime "2015-03-03T19:15:46"^^xsd:dateTime.
?location geo:lat ?lat.
?location geo:lng ?lng.
}
```

This query starts executing over the local RDF graph providing the measured and interpolated measurements in a determined time from the area of Skopje. This query returns similar data shown on the table 4:

Table 4. Partial result from the SPARQL query

Lat	Lng	Value
"42.05"^^xsd:fkiat	"21.32"^^xsd:fkiat	0.3
"41.96"^^xsd:fkiat	"21.31"^^xsd:fkiat	0.33
"41.94"^^xsd:fkiat	"21.29"^^xsd:fkiat	0.37
"42.03"^^xsd:fkiat	"21.3"^^xsd:fkiat	0.31

The result from the query could be used as input of a heat map obtaining the visual representation of the concentration of the CO in the air of the area of Skopje. The results from the previous query are represented on Figure 5.

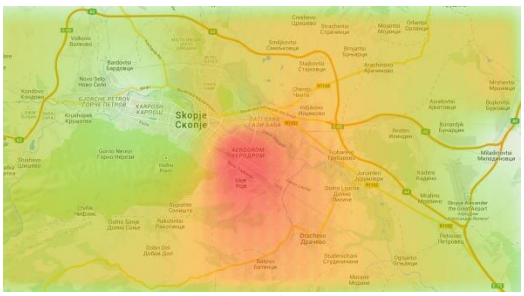

Figure 6. Visual heatmap representation of the CO measurement in the area of Skopje.

[2] http://d2rq.org/d2r-server
[3] http://d2rq.org/d2rq-language

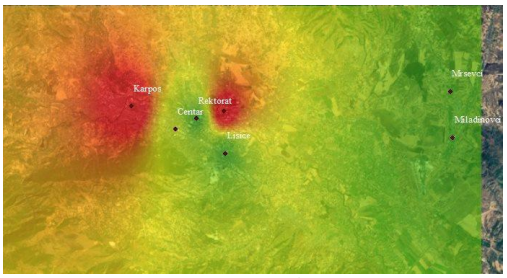

Figure 7 Visual heatmap representation of the PM10 measurements in the area of Skopje.

Analyzing the results on the heat map, we can conclude that the municipality of Aerodrom has the highest values of CO pollution.

The SPARQL endpoint for reviewing and analyzing the results of measurements is available on the following url:

http://airpollution.b1.finki.ukim.mk/

IV. Conclusion

The concept of Linked Data represents a big advantage in representation and retrieval of structured data from distributed parts of the Web. A large number of communities, companies and other interested stakeholders are taking part in the initiative and are contributing to the expansion of the LOD Cloud.

The type of the data that we contribute to open and to link is providing interesting analyzes about the current state of the air in the area of Skopje. We are allowing measurements about CO, CO_2, SO_2, O_3,NO_2, PM10 and PM2.5 air pollutants. The measurements are provided by 7 air pointer stations and 3 CO_2 measurement stations. We provide interpolated values for the areas that are not covered by the measurement stations. Interpolated values are created by sophisticated models of pollution spreading taking as parameters the infrastructure and the model of spreading of the appropriate pollutant. We save all of the data, measurements and interpolated values, in centralized database that is wrapped by D2RQ server providing mapping to RDF triples and linking to appropriate locations in Skopje. We provide a URL for accessing the data and reviewing the results using SPARQL query.

This type of data can help the citizens to find the best places for their activities or for living in Skopje. Also it can be used for retrieving the best eco routes for travelling in Skopje.

REFERENCES

[1] C. Bizer, T. Heath, K. Idehen, and T. Berners-Lee, “Linked data on the web,” 17th International conference on World Wide Web, ACM, 2008, pp. 1265-1266

[2] C. Bizer, T. Heath, and T. Berners-Lee, "Linked Data - the story so far," International Journal on Semantic Web and Information Systems 5, no. 3, 2009

[3] A. Naeve “The Human Semantic Web, Shifting From Knowledge Push To Knowledge Pull”, International Journal on Semantic Web and Information Systems (IJSWIS), 2005

[4] L. Barandovski, V. Urumov “Air pollution studies in Macedonia using the moss biomonitoring technique, NAA, AAS and GIS technology”, INIS, 2006

[5] M. Oprea “Mapping Ontologies in an Air Pollution Monitoring and Control AgentBased System”, Lecture Notes in Computer Science Volume 4265, 2006, pp 342-346.

[6] N.Koteli, K. Mitreski, D. Davcev “Monitoring, Modeling and Visualization System of Traffic Air Pollution – A Case Study for the City of Skopje”, International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies, 2014

[7] V. Epitropou, L. Johanson, K.D. Karatzas, A. Bassouckos, A. Karppinen, J. Kukkonen, M. Haakana, “Fusion of Environmental Information for the Delivery of Orchestrated Services for the Atmospheric Environment in the PESCaDO project”, International Environmental Modelling and Software Society (iEMSs), 2012

[8] A. Patil, S. Oundhakar, A. Sheth, K. Verma, “Meteor-s web service annotation framework”, Proceedings of the 13th international conference on World Wide Web, pp 553 - 562