Data Platform and Urban Data Services on Private Cloud

Vassil Vassilev[1], Bal Virdee[2], Karim Ouazzane[1], Dion Mariyanayagam[2], Viktor Sowinski-Mydlarz[1] Monika Rabka[2], Herbert Maosa[1], Sorin Radu[1]

[1] Cyber Security Research Centre, London Metropolitan University, UK
[2] Communication Technologies Research Centre, London Metropolitan University, UK
{v.vassilev,b.virdee,k.ouazzane}@londonmet.ac.uk

Abstract. This paper presents the result of a pilot project of London Metropolitan University, aiming at developing a set of urban data services in support of the local communities in several boroughs of the city of London. They are targeting the health and well-being of the citizens by analyzing different types of information from a number of data sources – environmental sensors, geolocation information, and models of the urban infrastructure. Unlike the complex government projects covering large urban areas, which require significant resources and typically involve large service providers operating public clouds, our project targets local communities with limited capabilities by utilizing the concept of a private cloud running on commodity infrastructure within their reach. By employing a number of proven data technologies, software tools, and AI methods the project delivers a comprehensive picture of urban life. The first phase of the project reported here focuses on outdoor and indoor pollution, which are the keys for addressing many local community activities such as environment protection, urban planning, local transport, communal housing, and social services within the area.

Keywords: Urban Infrastructure, Environmental Data, Private Cloud, Spatial and temporal reasoning for IoT, Hybrid AI.

1 Introduction

For several years two research centers of London Metropolitan University (Communication Technology and Cyber Security) have been working together to implement an affordable and comprehensive solution for monitoring and analyzing the dynamics of the outdoor and indoor environment and to explore its impact on the urban life using contemporary cloud technologies and methods of data analytics, Machine Learning and AI. Although already there are several public services, which provide information about the outdoor pollution in London [1] they are limited to monitoring the current levels of outdoor pollution by using publicly available data from sensor stations across London and provide only a basic statistical analysis of the pollution without considering its impact on the element of urban life due to its global nature. Our project aims to provide more elaborated analysis of the environmental factors, localized to particular areas of the city in order to serve the needs of the local community with timely and impactful analysis and potential advice. We are looking for

the impact of the air quality on the health conditions inside the living and working spaces and its dependence on various elements of urban life such as infrastructure, communications and neighborhood with the aim to suggest measures to limit it. In this paper we will present the prototype solution of a private cloud-based system which integrates data coming from both outdoor and indoor sensors and combines the pure data analysis with more informed hybrid methods which rely on the use of both ontological models and spatial visualization.

2 Technologies behind the Solutions

2.1 IoT and Real-time Data Processing

For a number of years, the Communication Technology Research Centre have been experimenting with sensors measuring physical data from motion to temperature, pressure and other factors of the physical environment [2,3]. On the other hand, in a previous project, dedicated to real-time security analytics, the Cyber Security Research Centre implemented a full pipeline for data management of real-time security data along its entire lifecycle from generation to elicitation, aggregation, transportation, ingestion and analysis [4]. The experience of the two centers converged in the current project and provided the technological basis for implementing statistical and correlation analytics of environmental data in real-time directly during its ingestion into the database on the cloud, in addition to the analysis of the collected data.

2.2 Private *vs.* Public Cloud

Our first project which utilized the cloud technologies was the Audio Beacon project, aiming at providing secure solution for online banking using voice-controlled devices such as **Amazon Alexa**. For prototyping of the solution, the public cloud of Amazon **AWS** was used [5]. Based on the success of this project and to expand its business opportunities the Cyber Security Research Centre created its own private cloud using **Kubernetes** and commodity hardware, which do not require significant resources and can be easily deployed to SMEs and public institutions preferring to keep the data on their own premises. During the pandemics, when the university was closed and all the premises were not operating, this solution was replicated on the premises of the GATE Institute for Big Data at Sofia University [6]. Considering the limited resources available at these two academic institutions this approach demonstrated viability and became the basis for project infrastructure at both organizations.

2.3 Ontologies and Data Analysis

Hybrid AI is one of the recent trends in data analysis which from fashion quickly becomes a necessity. The use of domain ontologies allows to combine the data analysis, based on discovering of patterns in data, with the possibility to account contextual relations within the data which remain hidden [7]. This way, it enables a deeper analysis of the data using domain-specific knowledge. The Cyber Security Research Cen-

tre has been pioneering such hybridization for analysing the vulnerability, for assessing the security risks and for designing solutions with guaranteed security of the financial and commercial systems in several of its projects [8,9]. As a first step towards similar hybridization for the purpose of analyzing the pollution in the current project the team enhanced the sensor readings with information about the location and the urban neighborhood. This would allow to initiate purely logical analysis of the potential causes and impact of the pollution on other areas of urban life (urban development, transport, logistic, etc.)

2.4 Sofia Air Pollution Pilot

During the work on a project dedicated to security analytics on a private cloud during the pandemics the team of researchers from London Metropolitan University worked in close collaboration with the researchers of GATE Institute of Sofia University, which resulted in a joint position paper presenting the concept of AI-based hybrid data platform on private cloud as a base for building Data Spaces [8]. As a technological feasibility study for the concept the joint team implemented on the private cloud of GATE Institute a pilot system for assessing the outdoor pollution in Sofia, which is one of the most polluted capitals in Europe suffering from extreme intensification of its private transport park in the recent years [10,11]. This system became the testbed for many ideas which made a way to the London pilot.

3 Cloud-based Platform of London Metropolitan University

Our data platform runs on a small commodity hardware cluster operating under Linux controlled by **Kubernetes** container management system. It has been equipped with well-proven community software which support most of the tasks for processing both streamlined data in real-time and large datasets, thus providing cheap and robust alternative to the big clouds like **AWS**, **Google Cloud** and **Microsoft Azure**. All four databases (Postgres, MongoDB, Neo4J and 3DCityGML) are deployed within Docker containers using yaml scripts to serve the need of this particular project (Fig. 1).

```
citydb:
        image: 3dcitydb/3dcitydb-pg:latest
        ports:
        - 5432:5432
        environment:
        - "CITYDBNAME=citydb"
        - "SRID=4326"
        - "SRSNAME=urn:ogc:def:crs:EPSG::4326"
        - "POSTGRES_USER=postgres"
        - "POSTGRES_PASSWORD=postgres"
```

Fig. 1. Containerization of **3dCityDB**

Despite the limited number of off-the-shelf products the platform demonstrated convincingly the advantages of the big cloud-based platforms. This was achieved thanks to the careful selection of software and its full integration. At the same time, it shows that it is feasible to have an enterprise quality of services on a private cloud without substantial investment, significant resources and levels of management.

The software components of the platform which support the environment pollution system are shown on Fig. 2. All off-the-shelf software products are containerized

which allows independent use of the platform for other applications as well, although in the case of sharing they can be installed directly on the server. In the original position paper, which presented this concept [6] there was an additional dimension for automatic generation and configuration of data pipelines using an explicit ontology of the data processing on the platform themself, but in the final implementation if the platform it was left out in favor of simplicity and flexibility. Currently, the platform is hosted within the Cyber Security Research Centre of London Metropolitan University, supporting several projects of the university research centers.

4 Urban Data Services on Private Cloud

The pilot implementation of London Urban Data Space was developed by a joint team of the Cyber Security and Communication Technology research centers of London Metropolitan University using publicly available outdoor data [1] and indoor data coming from the main building of the university.

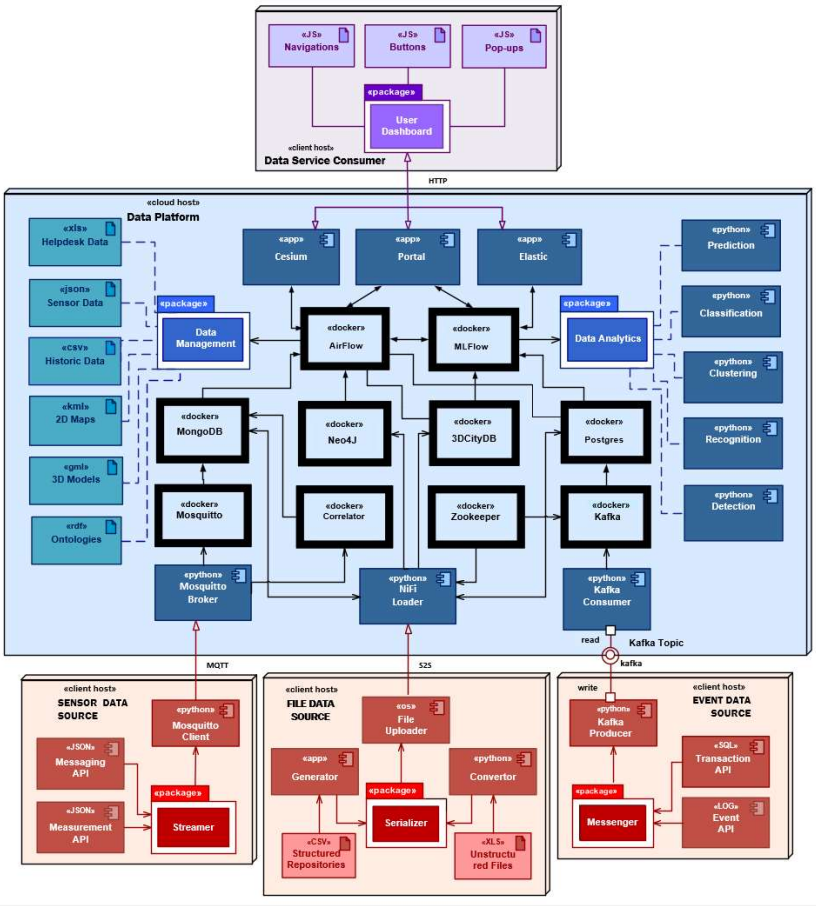


Fig. 2. Software Architecture for Data Services on the Private Cloud

4.1 Data Sources

Both the town municipalities and various environmental organizations are monitoring the air quality and collecting data about various pollutants in order to provide information, issue warnings, trigger actions and recommend decisions concerning the safety and health of the citizens. Sometimes such data can be obtained freely from public sources, but it is just the starting point for the analysis and require proper preparation. Our system collects environmental data from three different sources:

▼ 0 {16}
@LocalAuthorityCode : 27
@LocalAuthorityName : Richmond
@SiteCode : TD0
@SiteName : - National Physical Laboratory, Teddington
@SiteType : Suburban
@DateClosed : 2018-01-01 00:00:00
@DateOpened : 1996-08-08 00:00:00
@Latitude : 51.4243043441456
@Longitude : -0.345714576446947
@LatitudeWGS84 : 6696103.27675
@LongitudeWGS84 : -37808.8858115
@DisplayOffsetX : 0
@DisplayOffsetY : -200
@DataOwner : Richmond
@DataManager : King's College London
@SiteLink : http://www.londonair.org.uk/london/asp/publicdetails.asp?site=TD0

Fig. 3 JSON formatted information about the individual sensor stations

Outdoor sensor stations: Produce current readings from 107 sensor stations across London (temperature, humidity, pressure, gases, mechanical particles and biological agents). The readings are collected, formatted and sent to the cloud for accumulation and further analysis using suitable messaging protocols (MQTT). This process is repeated every hour so that the server receives new data on an hourly basis as it becomes available. All data is publicly available via **LondonAir** API from the sensor stations across London. The same API provides also information about the stations (Fig. 3) but since the call may contain information about stations that are no longer in operation, the returned information needs to be filtered to focus only on the running sensor stations. Once the filtering is complete, individual details can be extracted for each of the London stations in order to be used for subsequent visualization (Fig. 4). This API is a typical RESTful Web API which provide the data in JSON format, accessed via HTTP calls using URL-encoded parameters.

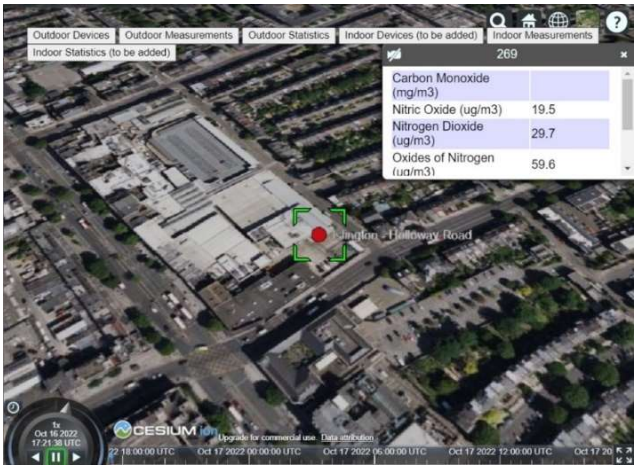


Fig. 4. Outdoor pollution around London Metropolitan University building

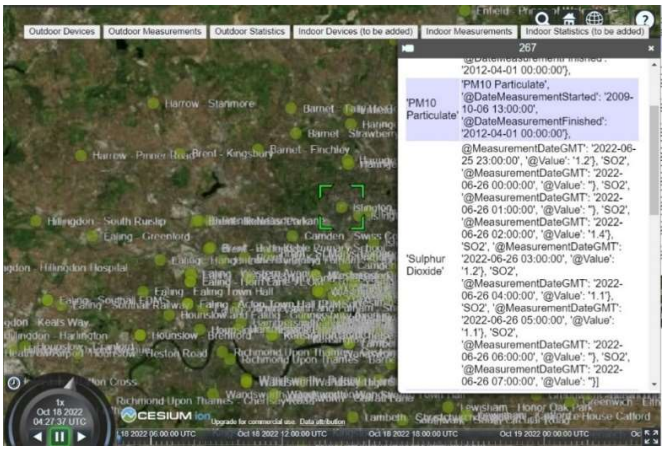


Fig. 5. Archived Data about Outdoor Pollution across London

Historical data archives: Deliver files and database exports containing accumulated data from the environment for certain periods of time over suitable streaming protocol (FTP, Kafka or SQS). This data is also publicly available and can be downloaded from the corresponding stations (Fig. 5). In our Sofia pilot we were also able to ingest text data from the helpdesk system of the local municipal authority, which contains citizen's complaints about the air quality in their living area, but when we migrated the system to London we had to limit ourself to sensor data only due to privacy and ownership constraints.

Fig. 6. Indoor sensor station

Indoor sensor stations: The Indoor Sensor Station (Fig. 6) is based on ESP32 microcontroller and sensors for temperature, humidity, pressure, particulate matters (PM1.0, PM2.5 and PM10), equivalent carbon dioxide (eCO2) and total volatile compounds (TVOC). It can provide data from public and communal buildings, offices and private houses. Fig. 7 shows the levels of pollution in the university building in the same area as the outdoor station shown on Fig. 2.

4.2 Hybridization

The information about outdoor and indoor pollution is certainly useful and can be used for different purposes, but it does not provide any interpretation of the causes and does not analyze the impact of the pollution on other elements of the urban life – urban infrastructure, transport networks, power and water supply, social services, healthcare, etc. This is especially important in projects aiming at Smart Cities [13]. But in order to go further in this direction, we need a whole set of analytics instruments for analysis of not only the sensor data, but also meta-data, contextual information and specialist knowledge.

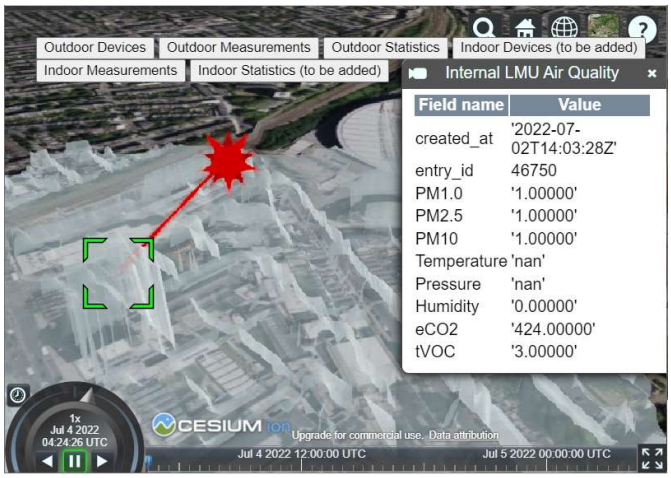
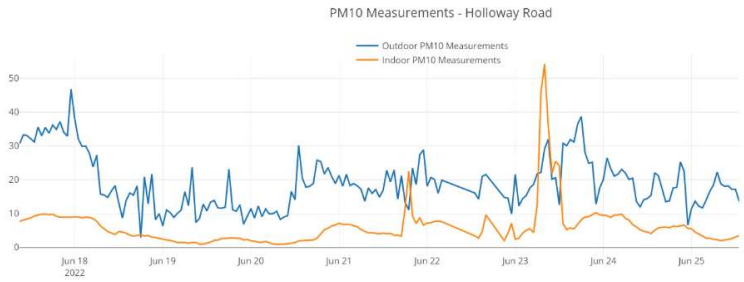


Fig. 7. Indoor pollution inside London Metropolitan University building at Holloway Road

In the recent years there has been a big push to standardize the models which support such hybridization [14,15]. The data platform we are using provides two major repositories for maintaining ontological and geospatial models of the environment – **Neo4J**, which allows to store ontologies, including urban ones, and **3DCityDB**, which supports both 2D and 3D graphical models in vector format. Neo4J is a graph database management system which has its own query language and in addition allows to make informed analysis of the sensor data based on its contextualization within the urban environment. In our pilot project in Sofia we used it to store a fragment of the urban ontology, which allowed us to use the knowledge about types to detect the potential polluters in the area. We are currently working on an ontological model of the buildings which would allow us to contextualize the indoor data as well. 3DCityDB, on the other hand, is a geospatial extension of the SQL DBMS Postgres, which we employ as a repository for structured data. It turns the database into a GIS-like system for storing 3D layers and 2D maps, which allow us to use it for visualization purposes. At the same time, it contains location data which can help the analysis by considering the location and the spatial distribution.

4.3 Correlation Analysis

The air pollution and the other environmental factors have detrimental effect on certain area of urban life – transport, housing, healthcare and social services. Standard methodology for investigating the effect of their presence is the correlation analysis, which allows us to discover dependencies and to estimate impact. In our pilot we have implemented both real-time data analytics for estimating the correlation between outdoor and indoor data, based on a fixed size temporal window using standard Pearson and Spearman methods. Fig. 8 shows the correlation between outdoor and indoor pollution in one particular rea of London where the university is. It is quite obvious that the level of pollution inside the building is heavily correlated with the level of pollution outside. The splash of pollution inside the building which does not correlate with the outside data is due to some building works taking place in the same period.



Correlation Matrix

Indoor Value	Outdoor Value
1	0.3634504520543431
0.3634504520543431	1

Fig. 8. Correlation between Outdoor and Indoor data in London area of Islington

4.4 Visualization

The visualization plays and important role in the provision of a comprehensive picture of the pollution across the city. This is why we paid special attention to it and used the best tool to provide high-quality visualization of the distribution of the pollution, **Cesium Ion**. This software supports most of the popular 2D and 3D formats for vector representation (KML, GML, WMS), supports integration of geolocation information with numerical, textual and symbolic data through GeoJSON and even allows extension of the visualization engines with custom-build templates. We first used it in the Sofia pilot [11] and after the successful completion of that project we continued using it in London (see Figs. 4, 5 & 7).

The visual representation of the city infrastructure for visualizing outdoor pollution is obtained from two different public sources: Open Street Map and ESRI World Imagery terrain level. In the case of indoor pollution we also used some Lidar data.

5 Conclusion and Future Work

The successful completion of our pilot demonstrates the feasibility of the private cloud solution. Its advantages compared to the public cloud include financial gains, operation isolation and privacy of the information. The disadvantages are in the need to deploy all necessary tools, to integrate them with the infrastructure of the cloud and to maintain the services guaranteeing security and safety of operations on enterprise level. One possible solution to address these issues is to host the cloud infrastructure in a data center, which can combine the advantages of both public and private cloud.

The system presented here allows to use the air pollution data for different purposes:

- Issuing warnings for avoiding the polluted areas to pedestrians, passengers and drivers
- Taking measures for reducing the impact of pollution by applying protective measures, putting on hold scheduled events and complete cancellation of regular work
- Making planning and design decisions concerning building works, organizing public spaces and establishing data services to prevent high impact of the pollution on people's health at work
- Addressing the challenges of urban life by devising suitable policies for improving the living and working conditions of the citizen

These options lead to different business scenarios for analyzing the environmental data, its causes and potential impact on the urban life. They serve as a guidance for developing data services for businesses, local councils, public organizations and governmental agencies. From a long-term strategic perspective our ultimate goal is together with our partners from EU to build an Urban Data Space for London, which involves our platform as a provider and a consumer of urban data services [14]. A simplified view of the possibility to integrate different data services which rely on information about the environment pollution across three of domains of the urban life is shown in Tab. 1 bellow. From purely research perspective we are interested in continuing our research in Hybrid AI by combining methods for data analysis and machine learning with methods for knowledge processing and logical inference [15, 16].

Tab. 1. Cross-domain Urban Data Services

	CITY MOBILITY CENTRE	ENVIRONMENT CONTROL AGENCY	URBAN PLANNING DEPARTMENT
Data Shared	routes, places, vehicles, waiting times	pollutions, standards, polluters	
Data Access Permissions	public (routes, places, waiting times); restricted (vehicles, locations)	public (pollution, standards); restricted (polluters)	
Operations Supported	locating, placing, timing	pollution, polluter determination	
Execution Rights	public (placing, timing); restricted (locating)	public (pollution); restricted (polluter)	
Data Consumed		routes, places, vehicles	vehicles, places, routes, pollution, polluters
Operations Executed		placing, locating	place, vehicle and route pollutions

Acknowledgements

The prototype of the system presented here is a result of several years of research which combines methods for data analysis, machine learning and AI with experiments with IoT and cloud technologies. It was initiated with the support of Lloyds Banking Group and was continued in a number of projects funded by UK Department of Culture, Media and Sport and UK

Higher Education Innovation Fund. During the pandemics the research was continued without interruption thanks to the additional support of GATE Institute of Sofia University, which also provided the testing ground for our concept and became an academic partner with which we are currently working on the preparation of an EU-UK project dedicated to the next level of digital services for the City of Future, Urban Data Space. We are grateful for their commitment and the support provided at the different stages of our long-term effort.

References

1. Imperial College [2022], LondonAir Network [Online: https://www.londonair.org. uk/]
2. M. Rabka, D. Mariyanayagam, and P. Shukla (2021) IoT based horticulture monitoring system. In: 5th World Conf. Smart Trends in Systems, Security and Sustainability (WS4 2021), Lecture Notes in Networks and Systems, Vol. 334, pp. 765-774, Springer (2021).
3. N. Mitu, M. Tabany, and V. Vassilev, Low Cost, Easy-to-Use IoT-based Real-Time Environment System (2021), Internet of Things and Web Services, Vol. 6, 202, pp. 30-44.
4. V. Vassilev, K. Ouazzane, H. Maosa, S. Nakarni, et. al., Security analytics on the cloud: Public vs. Private Case, in: Proc. 13th Int. Conf. on Cloud Computing, Data Science & Engineering (CONFLUENCE2023), 19-20 Jan 2023, Noida UP, India (to appear).
5. V. Vassilev, A. Phipps, M. Lane et al. (2020), Two-factor authentication for voice assistance in digital banking using public cloud services, In: Proc. Cloud Computing, Data Science & Engineering (CONFLUENCE2020), Noida UP, India, pp. 404-409, IEEE (2020).
6. V. Vassilev, S. Ilieva, D. Antonova, et al., AI-based Hybrid Data Platforms (2021), in: Curry, E., Scerri, S. and Tuikka, T. (eds.), Data Spaces, pp. 147-172, Springer (2022).
7. F. Vázquez, Ontology and Data Science [Online https://towardsdatascience.com/ontology]
8. V. Vassilev, V. Sowinski-Mydlarz, P. Gasiorowski, et al., Intelligence Graphs for Threat Intelligence and Security Policy Validation of Cyber Systems (2021), in: P. Bansal et al. (eds.), Adv. Intelligent Systems and Computing, Vol. 1164, pp. 125-140, Springer (2021).
9. V. Sowinski-Mydlarz, V. Vassilev, K. Ouazzane, and A. Phipps (2022). Security Analytics Framework Validation based on Threat Intelligence, in: Int. Conf. on Computational Science and Computational Intelligence (CSCI'22), Dec 14-16, 2022, Las Vegas, USA, IEEE.
10. S. Chow (2022), Sofia University Synthesizing Air Quality Data with Cesium [Online: https://cesium.com/blog/2022/03/15/sofia-university-synthesizing-air-quality-data-with/]
11. V. Vassilev, V. Sowinski-Mydlarz, et al. (2022), Towards first urban data space in Bulgaria, in: Proc. Smart Cities Conference (ISC2), 26-29 Sep 2022, Paphos, Cyprus, IEEE.
12. S. Zygiaris, Smart City Reference Model, J. of the Knowledge Economy (2013), vol. 4(2).
13. N. Tcholtchev, P. Lämmel, et al. (2018), Enabling the structuring, enhancement and creation of urban ICT through the extension of a standardized smart city reference model, in: Proc. Utility and Cloud Computing Companion (UCC Companion), IEEE, 2018.
14. Fraunhofer Institute (2018), Datenaustausch und Zusammenarbeit im urbanen Raum [online: https://www.fokus.fraunhofer.de/de/fokus/projekte/urbane_datenraeume]
15. M. Rasmussen, M. Lefrançois, M. Lefrançois, et al. (2020), BOT: the Building Topology Ontology of the W3C Linked Building Data Group, Semantic Web 1, IOS Press (2020).
16. C. Metral, G. Falquet, and K. Karatzas (2012), Ontologies for the Integration of Air Quality Models and 3D City Models [Online: https://doi.org/10.48550/arXiv. 1201.6511]