# Linking of open and private data in dataspace: A case study of air quality monitoring and forecasting

Alex Acquier[1], Andy Donald[1], Edward Curry[1], Ihsan Ullah[1,2] and Umair ul Hassan[1,3]

[1]Insight SFI Research Centre for Data Analytics, University of Galway, Ireland
[2]School of Computer Science, University of Galway, Ireland
[3]JE Cairnes School of Business and Economics, University of Galway, Ireland

Abstract

Air pollutant monitoring and its efficient visualisation can support accurately assessing the air quality and harmful emissions; and it can guide us towards potential mitigation strategies to reduce its impact on public health and our environment. This paper presents a case study of employing dataspaces and proposing an ontology for modelling mobility to address the challenges posed by the heterogeneity of data sources in environmental monitoring, as well as using machine learning for forecasting pollutants. We employ Linked data as a powerful paradigm for harmonising and interlinking diverse and publicly available environmental data with private company data to create a dataspace for environmental monitoring. By applying semantic technologies and ontological modelling to integrate heterogeneous data, our approach fosters data interoperability and facilitates enhanced data exploration and decision support. For decision support, we demonstrate the utility of integrated data for forecasting air pollutants with the help of models developed using machine learning. Finally, a spatio-temporal visualisation platform harnesses the power of semantic relationships and contextual enrichment to support data exploration.

**Keywords**

linked data, dataspaces, machine learning, spatiotemporal data, environmental monitoring, air pollution

## 1. Introduction

Advances in digital technologies have ushered in an era of unprecedented data generation across various domains. Environmental monitoring, in particular, has significantly improved data collection from diverse sources such as remote sensors, satellite imagery, social media platforms, commercial databases and government databases [1]. This wealth of information provides invaluable insights into our environment and weather. It helps in creating information systems that aid in decision-making processes, policy formulation, and resource allocation [2]. The breakthroughs in machine learning and artificial intelligence have led to renewed interest in forecasting environmental and weather conditions [3]. However, the heterogeneity of data

*The Second International Workshop on Semantics in Dataspaces, co-located with the Extended Semantic Web Conference, May 26 – 27, 2024, Hersonissos, Greece*

alex.acquier@universityofgalway.ie (A. Acquier); andy.donald@universityofgalway.ie (A. Donald); edward.curry@universityofgalway.ie (E. Curry); ihsan.ullah@universityofgalway.ie (I. Ullah); umair.ulhassan@universityofgalway.ie (U. u. Hassan)
0000-0003-2672-4085 (A. Acquier); 0009-0007-3307-2800 (A. Donald); 000-0002-1234-0000 (E. Curry); 0000-0002-7964-5199 (I. Ullah); 0000-0002-3647-9020 (U. u. Hassan)

sources, characterised by differences in formats, structures, and semantics, poses significant challenges to their effective integration and utilisation [4].

To address these challenges, the concept of Linked data has emerged as a powerful paradigm for harmonising and interlinking data from disparate sources [5]. Linked data fosters a decentralised approach to data integration, wherein each data source is assigned a unique identifier and linked to related information using standardised ontologies and vocabularies [6]. This approach enables seamless data interoperability and facilitates discovering hidden relationships and patterns that might remain obscured when data sources are treated in isolation. While Linked data focuses on using the Web standards and standardised ontologies to make data available, a *dataspace* aims to develop the necessary set of services to enable a custom view, for data consumers, of heterogeneous data sources which different data controllers manage [7].

This paper presents a novel approach in environmental monitoring, which is to create a dataspace using the Linked data approach. Our proposed dataspace acts as a linked platform where heterogeneous environmental data sources are integrated, interconnected, and made accessible in an iterative yet coherent manner. By utilising semantic technologies and ontological modelling, the dataspace allows for the easy integration of data originating from sources as diverse as mobile weather platforms, remote sensing platforms, meteorological websites, and governmental statistical offices. Towards this end, the primary objective of this paper is to showcase the feasibility and advantages of employing Linked data principles in environmental monitoring. We will discuss the intricacies of data source integration, ontology development and data inter-linkage within the context of the environmental monitoring. Furthermore, we will demonstrate how the dataspace fosters enhanced data exploration and decision support by harnessing the power of semantic relationships and contextual enrichment using appropriate visualisations and forecasting models. A vital aspect of this dataspace is the use of spatio-temporal data features as the anchoring point of all information integration and utilisation processes.

In the subsequent sections, we will delve into the technical aspects of the Linked data approach, elucidating the methods used for data extraction, transformation, and integration. We will also highlight the challenges encountered during the process and the strategies to overcome them. Additionally, we will provide case studies and real-world examples to underscore the practical implications of the proposed dataspace in aiding environmental research, policy-making, and public awareness.

## 2. Background

### 2.1. Environment and Climate Data Sources

The internet hosts diverse open data sources for environmental monitoring, making it a valuable resource for researchers, policymakers, and the public [6]. These open datasets cover various environmental factors, including air quality, climate, water quality, biodiversity, etc. One prominent example is the European Environmental Agency (EEA), which provides extensive environmental data, including weather observations, climate records, and satellite imagery, all accessible to the public for various applications, from climate research to weather forecasting [8]. Besides the EEA, each European country has national organisations, such as the Environmental

Protection Agency (EPA) and the Irish Meteorological Service (Met Éireann). In addition to open data, private companies also contribute valuable information for environmental monitoring. For instance, technology companies like Google capture real-time data from their mapping and navigation services, which can be harnessed to create custom views of environmental data. By analysing people's movement and traffic patterns, this data can be used to assess air pollution, identify congested areas, or monitor urban heat islands. Such private data can enhance our understanding of urban environments and enable more customised approaches to address environmental challenges. It is a valuable complement to open datasets in pursuing sustainable and eco-friendly solutions [9].

### 2.2. Common European Dataspaces

Since 2020, the European Commission has proposed establishing several dataspaces for sector-specific data exchange, sharing and pooling [10]. Such dataspaces can allow organizations to create dynamic and on-demand custom views over heterogeneous and distributed data sources, including the organization's protected data, private data from its partners, and publicly available open data [4]. Concerning environmental monitoring, two dataspaces are particularly relevant: the Mobility dataspace and the Green Deal dataspace. The Mobility dataspace focuses on the transportation sector where data from transport systems, traffic monitoring systems, transport companies, etc. [11] is collected, and insights are provided. When combined with the data from EEA, such data can provide a more granular analysis of the impact of transport on climate. More importantly, the Green Deal dataspace aims to provide a set of common infrastructure and services that will facilitate easy access to interoperable data related to climate, environment and sustainability across Europe [12]. While both these dataspaces aim to provide technical and legal guidance regarding sharing and exchanging data, the core challenge organisations still face is building their local services and interfaces over such data sources.

## 3. Dataspace for Environmental Monitoring

A pay-as-you-go approach to creating a custom view of environmental data leverages the variety of available open and private data in an incremental, flexible, and cost-effective manner [4, 13]. This approach is particularly relevant in environmental monitoring, where the diverse data sources are often voluminous and specialised. Rather than maintaining extensive infrastructure and data repositories, organisations and individuals can tap into open data resources like the EEA's climate data or Google's real-time location data as needed. This approach allows users to access the precise data they require, paying only for the specific resources and processing power necessary to create custom views. For instance, if a researcher needs real-time information on air quality in an urban area to study the impact of traffic patterns, they can harness private data sources like Google's mobility data alongside open datasets. By doing so, they can tailor their data processing and analysis to their project's scope, optimising costs and resource utilisation. This flexibility empowers many stakeholders to engage in environmental monitoring, ultimately fostering sustainable practices and decision-making.

To establish a dataspace, a set of services plays a pivotal role in creating custom views of environmental data when following a pay-as-you-go approach [14]. As illustrated in Figure 1,

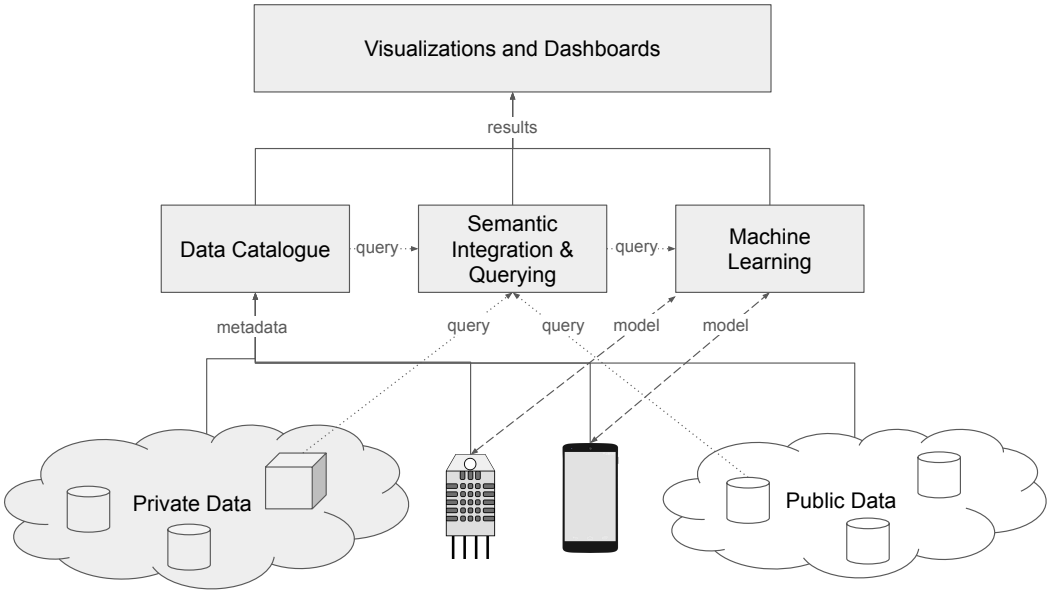


**Figure 1:** Overview of the dataspace defined using public data and private data of a company

these services encompass a range of capabilities and tools that facilitate data access, integration, analysis, and visualisation. First and foremost, data discovery services are essential, as they help users identify relevant datasets from the vast pool of public and private sources. These services provide metadata and cataloguing information, making finding the data that suits specific monitoring needs easier. In the absence of automated discovery services, a more practical approach is to create a data catalogue of known datasets and data sources. This cataloguing service serves as a canonical sources of metadata about data sources, and it is improved and updated overtime to facilitate current and future requirements for data discovery [15].

Once the correct set of data sources has been identified, a set of services for querying and semantic integration come into play. These services enable the harmonisation and blending of diverse data sources, ensuring compatibility and consistency across heterogeneous sources when accessing data. For instance, if one is creating a custom view of air quality data by combining open data from the EEA and EPA and private location data, data integration services help reconcile different data formats and units of measurement. Besides semantic integration, data querying services are equally vital to facilitate further processing and analysis. They allow users to apply various algorithms and statistical methods to extract valuable insights from the integrated datasets. In the context of environmental monitoring, this may involve calculating pollution trends, identifying hot-spots, or predicting future environmental conditions. Finally, dataspace services provide data visualisation and presentation tools to visualise and share the results effectively. These non-core services enable the creation of custom dashboards, maps, and reports to communicate findings to stakeholders, researchers, or the public in a user-friendly and actionable manner.

Overall, dataspace services bridge the wealth of available data and create custom views, facilitating the pay-as-you-go approach that allows users to tailor their data usage to their specific requirements while optimising efficiency, accuracy, and cost-effectiveness. To match the information requirements of custom views with the syntax and semantics of underlying data sources, applying some form of standardisation and semantic mapping across schemas and entities of data sources becomes imperative. Furthermore, any required statistical analytics will be used to present the data and its analysis better.

## 4. A Case Study of Air Pollutant Monitoring

This section presents a case study employing the dataspace approach and federated learning for air quality monitoring and forecasting. Table 1 shows a list of data sources used to create the environmental dataspace for pollutant monitoring. In addition, Pollutrack has emerged as a trailblazer in environmental data collection by implementing a sophisticated approach to monitoring air quality. Recognizing the critical importance of understanding and mitigating air pollution, Pollutrack has partnered with DPD to strategically deploy a combination of fixed and mobile platforms for comprehensive data collection in the city of Dublin. The fixed platforms, strategically positioned in urban centres and industrial zones, serve as constant monitoring hubs, capturing baseline air quality metrics over extended periods. Complementing this platform network, DPD introduced a fleet of mobile platforms equipped with the latest sensors on their delivery vehicles. These vehicles traverse diverse regions within Dublin, providing real-time data on the go. This dynamic approach offers a nuanced understanding of pollution dynamics influenced by various environmental factors.

In the following subsections, we will describe the proposed ontology and its integration, data quality and pre-processing, an overview of the machine learning-based pollutant forecasting module, and finally, our spatiotemporal visualisation.

### 4.1. Ontology Development and Integration

To enable linkages between data across multiple sources, the first step involved the definition of an ontology to define entities and their relationships [6]. The objective of the *Mobility in Cities* ontology, as illustrated in Figure 2, is to establish a standard that represents how information about environmental monitoring can be expressed via a set of classes and relationships. Following the Semantic Web standards, this ontology is designed to support various concepts associated with environmental monitoring and mobility across time and space in an urban environment.

**Core concepts:** At the core of the ontology, we have an observation that can be described, measured and observed. Each observation can have a location and time of when and where the observation occurred. We express a measurement of an observation as a phenomenon which needs to be observed. These phenomena can have multiple types of measurements, such as $PM_{2.5}$ (defined as particles that are 2.5 microns or less in diameter), traffic, weather or airport activity. Similarly, an observation can be observed by a sensor where a sensor is a device e.g. an

**Table 1**
List of data sources used to create the environmental dataspace for pollutant monitoring

| Data | Data Controller | Type | Data Source |
|---|---|---|---|
| Air pollution in Dublin | DPD | Private | Mobile & Fixed platforms |
| Weather Data in Ireland | Irish Meteorological Services | Public | www.met.ie |
| EPA Ireland's Open Data | Environmental Protection Agency (EPA) | Public | data.epa.ie |
| TII Open Data | Transport Infrastructure Ireland (TII) | Public | data.tii.ie |
| EEA Data Hub | European Environmental Agency (EEA) | Pubic | eea.europa.eu |
| Dublin Open Data | Smart Dublin (Local Authorities) | Pubic | data.smartdublin.ie |
| Weather Forecast Data | European Centre for Medium-Range Weather Forecasts (ECMWF) | Public | www.ecmwf.int |
| Atmosphere Data Store | Copernicus Atmosphere Monitoring Service (CAMS) | Public | ads.atmosphere.copernicus.eu |

IoT or edge sensor with a particular measurement capability with defined features. The sensor entity can implement sensing, which has a specific sensor output that can be considered an observation of a defined phenomenon.

**Linkage with other ontologies:** The ontology is designed to describe issues related to urban environments, such as air quality, water quality, traffic volume, etc. Following the principles of Linked data, some concepts within this ontology are defined using existing and well-known ontologies in the Semantic Web. For instance, the HealthRisk Ontology [16] is used to describe an observation of air quality. Similarly, the concepts related to time and location have been defined using existing Time and GEO ontologies, respectively. A key benefit of such linkages with existing ontologies is supporting the semantic integration of data fetched from multiple sources.

The ontology design was carefully developed to allow flexibility for future extension to cover different urban mobility observational phenomena. For instance, we detail air quality measurements from sensors capturing traffic volume at traffic lights. Adding sub-classes covering data sources from, for example, public transport or cycling infrastructure can significantly enhance the ontology's impact. With the development of this dataspace, we anticipate it will support several future implementations using public data sources, which will validate our hypothesis and support continual extensions to the ontology.

## 4.2. Data Quality & Pre-processing

The raw data about the air quality was not usable straightaway due to (i) the wide area covered by the mobile platforms, (ii) some readings recorded outside the normal times of business, (iii) issues with the sensor identification scheme, and (iv) differences of data recording intervals

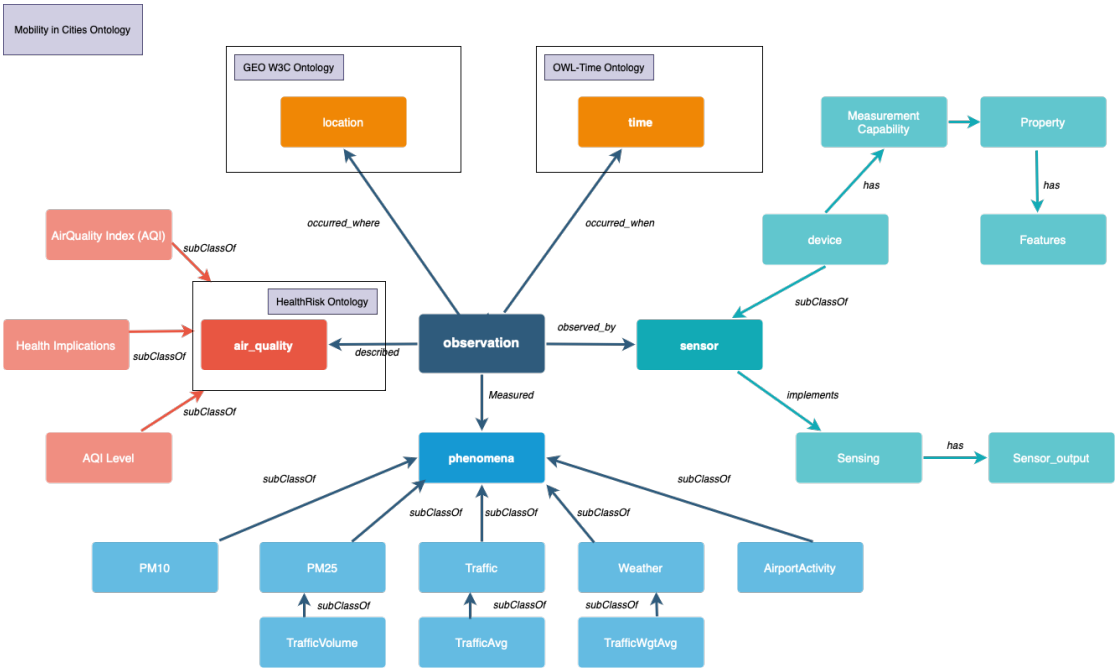


**Figure 2:** Ontology designed for modelling of mobility and environment-related data

between the different platforms (v) missing data due to breakdown of sensors or other reasons. To address these issues and improve data quality, a set of pre-processing steps were applied before any machine learning tasks. The air pollutant data is filtered through to keep only an area of $15x15$ kilometres that encompasses the centre of Dublin (latitude ranging from 53.2821 to 53.417, longitude ranging from -6.377 to -6.15065) for both fixed and mobile platforms data and sorted into three different spatial granularity. Each of the following datasets was used in a series of tests to allow the predictions of particulate matter at both $PM_{2.5}$ and $PM_{10}$ sizes for all spatial subdivisions.

- Global ($3x3$): 9 squares with 5 kilometres sides
- Local ($5x5$): 15 squares with 3 kilometres sides
- Hyper-local ($15x15$): 175 squares with 1-kilometre sides

**Data standardisation:** In the first step, new identifiers were assigned to the different platforms, which made it more easily human readable and helped for the display in applications and debugging. All data was homogenised with a window size of one hour. For the fixed data, the values of the readings were added up and divided by the number of records to give the final results per hour for each platform concerned. For the mobile data, each reading was kept untouched concerning the values and given the timestamp associated with the original reading time, which means that the same timestamps can appear several times for the platforms. Still, the uniqueness is ensured by the coordinates where the reading has been taken. It was found

**Table 2**
Results of air pollutant forecasting based on public and private data

| | $PM_{2.5}$ | | $PM_{10}$ | |
|---|---|---|---|---|
| | RMSE | MAE | RMSE | MAE |
| $3x3$ | 7.516558622 | 3.420987033 | 8.4745889 | 4.086204333 |
| $5x5$ | 6.970830328 | 3.384417792 | 7.721180158 | 3.778929738 |
| $15x15$ | 5.228858959 | 2.474564648 | 6.055090154 | 2.966958507 |

that the mobile platforms were used far outside the area where the fixed platforms are located, and those data points were removed. Also, only the readings were taken between 8am and 8pm, as these are the hours when mobile platforms should be out for data collection.

**Outlier detection:** The Local Outlier Factor (LOF) algorithm is an unsupervised anomaly detection method which computes the local density deviation of a given data point concerning its neighbours. It considers the samples with a substantially lower density than their neighbours as outliers. The LOF method was used by looking into the relation between $PM_{2.5}$ and $PM_{10}$ values using 1000 neighbours; between 1% to 3% of the data points were deemed outliers and were removed from the final datasets.

**Data imputation:** Dealing with missing values is one of the most common data quality requirements in environmental data processing, and previously, several data imputation methods for missing air quality are hyper-local level [17]. When the particulate matter data was mapped according to three levels of granularity, not all the squares created were necessarily populated (i.e., the square either had no fixed platform around it or was located outside the urban area). To overcome this problem, the values of the neighbouring squares were used to extrapolate the missing data by adding the values of all the neighbour-populated squares for a given timestamp. The result was divided by the number of neighbour-populated squares. After this data imputation, both fixed and mobile data for datasets of both types of particulate matter was integrated with weather data (e.g., dry and wet bulb temperatures, dew point, and atmospheric water content) for pollutant forecasting and visualization.

### 4.3. Pollutant Forecasting with Machine Learning

Machine learning, specifically its deep learning sub-branch, has shown promising results for various applications, e.g., autonomous vehicles and the medical domain. Similarly, it is used to predict and forecast air pollution. In this work, a hybrid model [18] resulting from the combination of a conv layer, LSTM layer, and an attention-based layer is adapted to be used in a federated learning approach [19, 20] to forecast pollutants ($PM_{2.5}$ and $PM_{10}$) in the air. The model is trained on data from Dublin City. It consists of both private pollutant data collected by DPD and public data collected by the Irish Meteorological Service (Met Éireann).

In time series data, the sliding window approach is normally adopted to select the input for predicting the next value/values. In this scenario, the window can comprise different time lags,

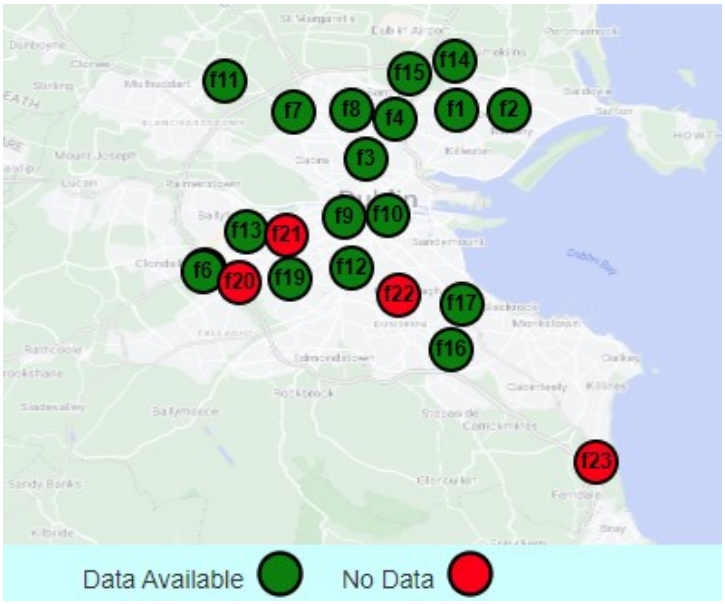


**Figure 3:** Location of the fixed sensors platforms

i.e., the number of previous hours of data necessary to generate the predictions for the next second/hour/day. We chose 6-time lags, i.e., 2, 4, 6, 8, 10, and 12 hours, to experiment at different granularities. The RMSE and MAE are used to assess the model's performance. In addition, to divide and select only specific region data, we divided the regions with in three regions called global $3x3$, local $5x5$, and hyper local $15x15$ as being defined before in section 4.2.

Table 2 shows that a decrease in the values of MAE and RMSE occurs as the number of divisions increases. When the data is grouped within a $3x3$ granularity, the best time lag found is 12 hrs for both $PM_{2.5}$ and $PM_{10}$. The $5x5$ granularity shows different optimum results for $PM_{2.5}$ (12hrs) and $PM_{10}$ (8hrs) but given the small increase for RMSE and MAE in the 12hrs results compared the 8hrs result for $PM_{10}$. The $15x15$ shows different optimum results for $PM_{2.5}$ (8hrs) and $PM_{10}$ (4hrs), but given the small increase for RMSE and MAE in the 8hrs results compared to the 4hrs result for $PM_{10}$. It was observed that for all time lags, as the granularity increases, both RMSE and MAE values decrease for both $PM_{2.5}$ and $PM_{10}$.

### 4.4. Spatio-temporal Visualisation

Besides forecasting, the objective is to create custom views so that any user would be able to visualise the clean and integrated raw data with the help of appropriate visualizations and graphing tools. For this purpose, a web-based visualization tool was created that allowed a user to select a geo-spatial area of interest and time frame for viewing data on a graph. To help with the selection of an area time frame, the location of each fixed platform is shown on a map (see Figure 3) and when the mouse hovers on one of the locations, the coordinates, first and last reading timestamps are shown.

To query the relevant Linked data, users are asked to provide the maximum and minimum

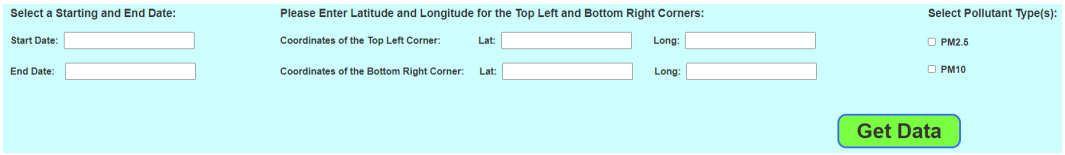


**Figure 4:** Visualisation tool's user interface

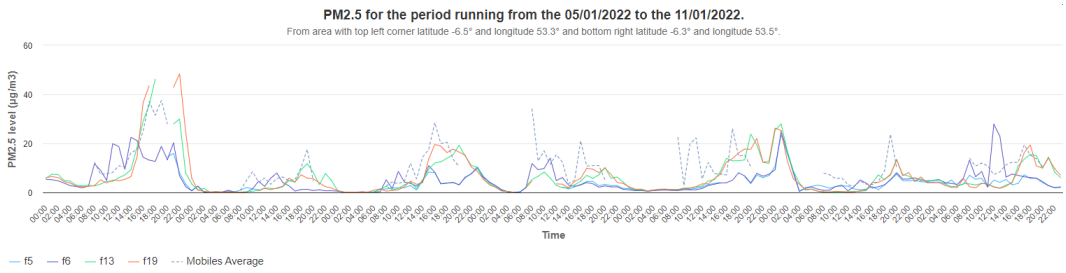


**Figure 5:** Plotted data for $PM_{2.5}$ pollutant.

latitude, longitude and dates as well as the type of pollutant (see Figure 4). The dates are used to first filter the mobile data, it is then further refined using the latitude and longitude coordinates and the values are averaged for each each hour of the designed time frame and over the selected area. The fixed platform identifiers within the search area's compound are determined using the metadata of fixed platforms. Once the identifiers are found, they are used in concert with the start and end dates to query data. These processes are used for any of the types of pollutant(s) selected ($PM_{2.5}$ and/or $PM_{10}$). Once this is over, the data of each fixed platform within the designated area and the associated averaged data are plotted (see Figure 5) and their associated mean and standard deviation are displayed in a table below the graph for each pollutant type requested.

As shown in Figure 5, the machine learning forecast can be added to this time-series visualisation. It can allow the user to see future forecasts of pollutants in a specific area. Similar visualizations for descriptive and predictive analytics will be added as part of the future work.

## 5. Conclusion and Future Research

This paper presents a novel approach combining public and private data to create an on-demand dataspace for the environment. Our proposed dataspace acts as a linked platform where heterogeneous environmental data sources are integrated using semantic technologies and ontological modelling. The utility of the proposed data is demonstrated with the help of a case study that combines pollutant data from a company with publicly available weather data to create interactive visualizations and a pollutant forecasting model. Both public and private data are relevant to pollution monitoring in Dublin city, and the outputs of this research work can help in achieving the Net Zero 2050[1] target of the Government of Ireland. Furthermore, this

[1]https://www.gov.ie/en/press-release/16421-climate-action-plan-2021-securing-our-future/

work is aligned with UN's SDG 13[2] on climate action.

To extend the work presented in this paper, several challenges still require further investigation. One of those challenges is the data quality. For instance, if there are significant time gaps in the data, how can we address these issues? Various imputation techniques can be adopted, including but not limited to averaging or EM algorithm [21]. Such techniques can be further extended with more intelligent mechanisms to fill the gap closer to the original data. In this regard, a proposed work could utilise public data to pre-train a base model and then use it to fill the gaps in private data. Another challenge is the integration of public and private data into data provenance. Providing details of the sources of data and its lineage can help improve the opacity of both the visualization and machine learning models. For this purpose, existing ontologies such as PROV-DM[3] and Dublin Core[4] can be used to generate provenance metadata [22].

Creating complex artificial intelligence services over combined public and private brings its own set of challenges. One of those challenges is to deal with different scopes of data privacy and protection applicable to each source dataset [23]. One approach to address this issue is to follow a federated learning approach and train models individually on each data source [20]. This approach allows the building of aggregate models without explicit integration of heterogeneous datasets.

## Acknowledgments

This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant Number 12/RC/2289_P2 - Insight SFI Centre at the University of Galway. For the purpose of Open Access, the authors have applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

## References

[1] H. Messer, A. Zinevich, P. Alpert, Environmental monitoring by wireless communication networks, Science 312 (2006) 713–713.

[2] N. P. Melville, Information systems innovation for environmental sustainability, MIS quarterly (2010) 1–21.

[3] I.-I. Prado-Rujas, E. Serrano, A. García-Dopico, M. L. Córdoba, M. S. Pérez, Combining heterogeneous data sources for spatio-temporal mobility demand forecasting, Information Fusion 91 (2023) 1–12.

[4] E. Curry, Real-time linked dataspaces: Enabling data ecosystems for intelligent systems, Springer Nature, 2020.

[5] E. Curry, W. Derguech, S. Hasan, C. Kouroupetroglou, U. ul Hassan, A real-time linked dataspace for the internet of things: enabling "pay-as-you-go" data management in smart environments, Future Generation Computer Systems 90 (2019) 405–422.

[2] https://sdgs.un.org/goals/goal13

[3] https://www.w3.org/TR/prov-dm/

[4] https://www.dublincore.org/

[6] T. Heath, C. Bizer, Linked data: Evolving the web into a global data space, Springer Nature, 2022.
[7] M. Franklin, A. Halevy, D. Maier, From databases to dataspaces: a new abstraction for information management, ACM Sigmod Record 34 (2005) 27–33.
[8] B. Schmidt, B. Gemeinholzer, A. Treloar, Open data in global environmental research: The belmont forum's open data survey, PloS one 11 (2016) e0146695.
[9] A. Zuiderwijk, M. Janssen, K. Poulis, G. van de Kaa, Open data for competitive advantage: insights from open data use by companies, in: Proceedings of the 16th annual international conference on digital government research, 2015, pp. 79–88.
[10] European Comission, A european strategy for data, 2020.
[11] S. Pretzsch, H. Drees, L. Rittershaus, Mobility data space, in: B. Otto, M. ten Hompel, S. Wrobel (Eds.), Designing Data Spaces: The Ecosystem Approach to Competitive Advantage, Springer International Publishing, 2022, pp. 343–361. doi:10.1007/978-3-030-93975-5_21.
[12] M. Gutierrez David, M. Dietrich, N. Raczko, S. Denvil, M. Santoro, C. Chatzikyriakou, W. Borejko, Towards the european green deal data space, in: EGU General Assembly Conference Abstracts, 2023, pp. EGU–8788.
[13] U. ul Hassan, S. Hasan, W. Derguech, L. Hannon, E. Clifford, C. Kouroupetroglou, S. Smit, E. Curry, Water analytics and management with real-time linked dataspaces, in: Government 3.0–Next Generation Government Technology Infrastructure and Services: Roadmaps, Enabling Technologies & Challenges, Springer, 2017, pp. 173–196.
[14] A. Halevy, M. Franklin, D. Maier, Principles of dataspace systems, in: Proceedings of the twenty-fifth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, 2006, pp. 1–9.
[15] U. ul Hassan, A. Ojo, E. Curry, Catalog and entity management service for internet of things-based smart environments, in: Real-time Linked Dataspaces: Enabling Data Ecosystems for Intelligent Systems, Springer, 2020, pp. 89–103.
[16] X. Meng, F. Wang, Y. Xie, G. Song, S. Ma, S. Hu, J. Bai, Y. Yang, An ontology-driven approach for integrating intelligence to manage human and ecological health risks in the geospatial sensor web, Sensors 18 (2018) 3619.
[17] H. Junninen, H. Niska, K. Tuppurainen, J. Ruuskanen, M. Kolehmainen, Methods for imputation of missing values in air quality data sets, Atmospheric environment 38 (2004) 2895–2907.
[18] S. Li, G. Xie, J. Ren, L. Guo, Y. Yang, X. Xu, Urban pm2.5 concentration prediction via attention-based cnn–lstm, Applied Sciences 10 (2020). doi:10.3390/app10061953.
[19] G. Zhang, S. Zhu, X. Bai, Federated learning-based multi-energy load forecasting method using cnn-attention-lstm model, Sustainability 14 (2022). doi:10.3390/su141912843.
[20] I. Ullah, U. ul Hassan, M. I. Ali, Multi-level federated learning for industry 4.0 - a crowdsourcing approach, Procedia Computer Science 217 (2023) 423–435. doi:10.1016/j.procs.2022.12.238.
[21] W. Junger, A. P. De Leon, Imputation of missing data in time series for air pollutants, Atmospheric Environment 102 (2015) 96–104.
[22] D. L. da Silva, A. Batista, P. L. Correa, Data provenance in environmental monitoring, in: 2016 IEEE 13th International Conference on Mobile Ad Hoc and Sensor Systems (MASS),

IEEE, 2016, pp. 337–342.
[23] S. Arora, P. Lewis, A. Fan, J. Kahn, C. Ré, Reasoning over public and private data in retrieval-based systems, Transactions of the Association for Computational Linguistics 11 (2023) 902–921.