FAIR and Quality-Aware Air Quality Data Management: A Knowledge Graph-Based Approach*

Martin Katzenstein[1,*,†], Lorena Etcheverry[2,†]

[1]Facultad de Ingeniería, Universidad de la República, Uruguay
[2]Instituto de Computación, Facultad de Ingeniería, Universidad de la República, Uruguay

Abstract

Air pollution poses significant environmental and public health challenges, necessitating effective air quality data management. However, current approaches face limitations in ensuring data quality, interoperability, and compliance with FAIR (Findable, Accessible, Interoperable, and Reusable) principles. In this work-in-progress, we present a prototype of a knowledge graph-based system designed to enhance air quality data management across its entire lifecycle, from collection and validation to publication. Our approach integrates semantic web technologies to explicitly represent data provenance, quality dimensions, and interoperability requirements. We apply our system to a case study in Uruguay, where air quality data is collected from multiple organizations, highlighting the challenges of cross-institutional data integration and validation. Preliminary results demonstrate improvements in data consistency, traceability, and usability. Future work will focus on refining scalability, enhancing data quality inference mechanisms, and integrating additional environmental datasets.

Keywords

Air Quality Data Management, Knowledge Graphs, FAIR Data Principles,

1. Introduction

Air pollution is a critical global challenge with significant environmental and public health implications. Monitoring and managing air quality data are essential for assessing pollution levels, identifying trends, and supporting policy decisions to mitigate harmful effects [1]. However, the effective use of air quality data is hindered by challenges related to data heterogeneity, quality assessment, provenance, and interoperability [2]. Data is often collected from diverse sources—such as ground-based sensors, satellite observations, and environmental agencies—each with varying formats, standards, and reliability. Ensuring that this data is properly validated, traceable, and accessible in compliance with the FAIR (Findable, Accessible, Interoperable, and Reusable) principles remains a pressing concern [3].

In addition to the above mentioned intrinsic complexities related to the management of air quality data, the processes involved in data management are generally built from a wide range of subprocesses that include policies and actors from many different organizations. Among others, these organizations are typically governments, private companies, nongovernmental organizations and independent laboratories. The roles of each of the organizations involved are often described in applicable laws, commercial contracts, international agreements between counties, etc. Sometimes, roles arise from ad hoc exchange between organizations with no clear responsibilities or accountability for the quality of the data.

In this work, we present a prototype of a data management system designed to address the entire lifecycle of air quality data, from collection and validation to publication, where multiple organizations are involved. Our approach leverages knowledge graphs and semantic web technologies to enhance data integration, representation, and usability. A key novelty of our system is the explicit representation

The 3rd International Workshop on Knowledge Graphs for Sustainability (KG4S2024) – Colocated with the 22nd Extended Semantic Web Conference (ESWC2025), June 1 2025, Portoroz, Slovenia.

*Corresponding author.
†These authors contributed equally.
martin.katzenstein@fing.edu.uy (M. Katzenstein); lorenae@fing.edu.uy (L. Etcheverry)
https://www.fing.edu.uy/~lorenae/ (L. Etcheverry)
0000-0002-0877-7063 (M. Katzenstein); 0000-0001-8121-8076 (L. Etcheverry)

of air quality data quality dimensions using RDF, which enables structured, machine-readable metadata about the reliability and accuracy of air pollution measurements. This approach ensures transparency in data provenance and facilitates traceability across the data lifecycle, enabling users to assess the trustworthiness of air quality information effectively.

As this is an ongoing research effort, we present a prototype implementation along with preliminary results that showcase the feasibility of our approach. The current system provides an initial framework for integrating and managing air quality data, and future work will focus on refining its scalability, improving inference mechanisms for data quality assessment, and expanding interoperability with external environmental datasets.

The remainder of this paper is structured as follows: Section 2 discusses related work on air quality data management and semantic technologies. Section 3 presents our proposed system, describing its architecture and data processing pipeline. Section 4 describes the application of our approach to a case study involving Uruguayan air quality monitoring, and Section 5 concludes with insights on further work and future research directions.

2. Related Work

Several studies focus on air quality management from a methodological perspective; however, they do not specifically emphasize the information systems required for effective implementation [1]. Other research reviews existing and potential applications of data management and analysis techniques throughout the air quality lifecycle, but these studies tend to avoid delving into detailed discussions [4, 5].

Previous work already explored using knowledge graphs and semantic web technologies for air quality data management. Still, none focuses on data provenance and/or modeling data quality and validation processes. In [6], the authors present an overview of a system that collects data from sensors, using linked data and SPARQL query to retrieve information. This work does not consider the data lifecycle, which deals with inconsistencies and data quality problems. Wu et al. [7] describe how to use the Semantic Sensor Network Ontology (SSN) [1] and custom vocabularies to represent air quality measures using semantic web technologies while Galarraga et al. [8] focus on the publication of air data quality measures as Open Data, in particular as multidimensional data cubes on RDF using QB4OLAP vocabulary [9]. Finally, in [10], the authors present the use of knowledge graphs in the case of Australia. Although this work has a broader perspective on the data lifecycle, it does not focus on tracking data provenance or dealing with data validation and data quality metadata.

3. Proposed Approach

Our approach integrates existing semantic web vocabularies to comprehensively represent the different aspects of the lifecycle of environmental data, specifically air pollutant data. By leveraging well-established ontologies, we ensure interoperability, provenance tracking, and structured data quality representation, aligning with FAIR principles. This section outlines the key components of our system, including the semantic vocabularies used, the system architecture, and the mechanisms for ensuring traceability across the data lifecycle. In particular we adopt concepts from the following ontologies and volcabularies

3.1. Semantic Vocabularies for Environmental Data Representation

To effectively model and manage air quality data, we adopt a combination of semantic web vocabularies, each addressing specific aspects of the data lifecycle:

[1]Semantic Sensor Network Ontology (SSN) https://www.w3.org/TR/vocab-ssn/

- **SSN** The Semantic Sensor Network (SSN) [2]ontology is an ontology for describing sensors and their observations, the involved procedures, the studied features of interest, the samples used to do so, and the observed properties, as well as actuators. Specifically, the **SOSA Module** (Sensors, Observations, Samples and Actuations) defines the core classes and properties, and provides textual definitions and other annotations. This ensures consistency in representing air pollutant concentrations, measurement units, and sensor specifications.
- **PROV-O**: The PROV Ontology[3] is employed to model provenance information, including dataset relationships and data generation processes. This enables traceability and accountability throughout the data lifecycle.
- **DQV (Data Quality Vocabulary)**[4]: Integrated to represent data quality dimensions, metrics, and values explicitly. These vocabularies allow us to encode quality metadata in a machine-readable format, facilitating automated quality assessment.
- **DataCube**[5] and **QB4OLAP[9]**: Utilized for structured and queryable data publication. These vocabularies support multidimensional analysis, enabling users to explore air quality data across dimensions such as time, location, and pollutant type.

By combining these vocabularies, our system provides a unified framework for managing air quality data while adhering to FAIR principles. In order to distinguish raw data from validated data we will use the following terminology: i) Registers represent data points in the data staging area, and represent the output of a sensor at a certain date and time, while ii) Measures represent data points in the validating area, which are derived from the information contained in a Register.

We organize our approach based on three ontologies that reuse and extended concepts from existent ontologies and vocabularies:

- AIRQorg - That allows to represent all relevant institutional information, including monitoring stations, agreements between institutions, and personnel involved in related tasks.
- AIRQreg - That models all the registers as they are generated from the information sent from the Sensors, including geographical information, pollutant, etc.
- AIRQmed - That models all the validated measures generated from the registers. It also models the quality requirements (e.g. the value is within the sensor operating range) and provides information on the validation process, in particular on which agents participated in it.

Figure 1 presents an overview of the AIRQMeas ontology. All the proposed ontologies and additional diagrams are accessible in this repository https://gitlab.fing.edu.uy/air-data-quality/vocabularies-and-ontologies

3.2. System Architecture

Our system follows a structured data management pipeline that ensures the traceability, quality, and FAIR publication of air quality data. It is designed around three main components: **Data Staging Area**, **Data Validation Area**, and **Data Publication**, each playing a critical role in the data lifecycle. Figure 2 depicts the data flow through these components, showing how the proposed ontologies take part in the different stages. In the following subsections we outline the main responsibilities of each component.

3.2.1. Data Staging Area: Ingesting and Organizing Raw Data

The **Data Staging Area** is the initial entry point for raw air quality data collected from measuring stations and third-party environmental agencies. This component handles:

[2]Semantic Sensor Network Ontology (SSN) https://www.w3.org/TR/vocab-ssn/
[3]PROV-O: The PROV Ontology https://www.w3.org/TR/prov-o/
[4]Data on the Web Best Practices: Data Quality Vocabulary https://www.w3.org/TR/vocab-dqv/
[5]The RDF Data Cube Vocabulary https://www.w3.org/TR/vocab-data-cube/

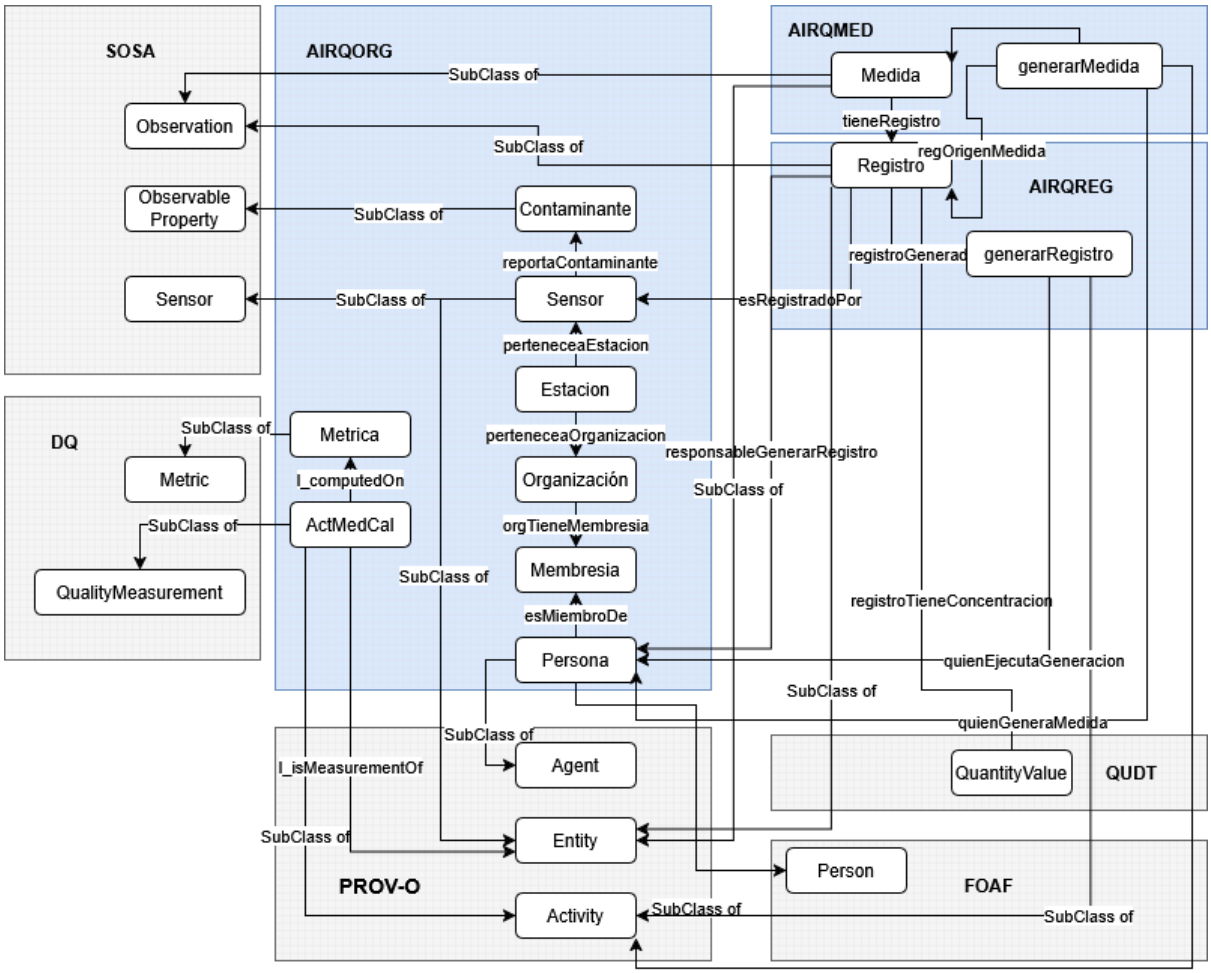

Figure 1: Overview on the proposed ontologies and their relation with existent ontologies and vocabularies.

Figure 2: System architecture overview and data flow.

- **Data Ingestion**: Automated pipelines retrieve sensor data in various formats (CSV, JSON, XML, RDF). These pipelines are designed to handle heterogeneous data sources, ensuring seamless integration into the system.
- **Preprocessing and Harmonization**: Raw data is standardized to ensure consistency in units, timestamps, and formats. The **AIRQorg** and **AIRQreg** vocabularies map diverse data representations into a unified model.
- **Initial Provenance Tracking**: Metadata about data sources, timestamps, and ingestion processes is recorded using features from **PROV-O** included in **AIRQreg**. This ensures traceability from the moment data enters the system, providing a foundation for end-to-end provenance tracking.

3.2.2. Data Validation Area: Assessing Data Quality

Once ingested, data is transferred to the **Data Validation Area**, where quality evaluation is performed based on predefined criteria. This component includes:

- **Data Quality Assessment**: Quality dimensions such as accuracy, completeness, consistency, and timeliness are evaluated using **DQV** and **BigOWL4DQ**. These vocabularies enable the explicit representation of quality metrics and their values, ensuring transparency in quality assessment.

- **Anomaly Detection & Correction**: Missing values, outliers, and sensor errors are identified using statistical and rule-based methods. Automated correction mechanisms are applied where applicable, while unresolved anomalies are flagged for manual review.
- **Provenance and Traceability**: All validation steps, including quality assessments and anomaly corrections, are logged using **PROV-O**. This ensures that transformations, quality evaluations, and modifications remain traceable throughout the data lifecycle.

3.2.3. Data Publication: FAIR-Compliant, Multidimensional Representation

Validated and enriched data is published in a structured and queryable format, supporting multidimensional analysis for decision-making. This component incorporates:

- **Multidimensional Modeling**: Using **DataCube** and **QB4OLAP**, air quality data is organized into multidimensional structures. This allows users to aggregate and analyze data by dimensions such as time, pollutant type, and geographical region.
- **Linked Data Integration**: Air quality data can be connected with external datasets (e.g., weather conditions, industrial emissions) to enable enriched analysis. This integration is facilitated by RDF-based linking mechanisms, ensuring interoperability across datasets.
- **FAIR Data Accessibility**: The system ensures **Findability** through metadata indexing, **Accessibility** via standard SPARQL endpoints, **Interoperability** through RDF-based vocabularies, and **Reusability** with well-defined licensing and quality metadata. These features make the data accessible to a wide range of stakeholders, including researchers, policymakers, and the public.

Finally, to ensure end-to-end traceability, we implement persistent identifiers by assigning unique URIs to datasets, observations, and validation results, ensuring long-term referenceability. Versioning and lineage tracking are managed using PROV-O to capture dataset versions, modifications, and transformations, allowing users to track changes and their impact. Additionally, provenance metadata is accessible through SPARQL queries, enabling stakeholders to verify data history and reliability, while access control mechanisms protect sensitive information while maintaining transparency for authorized users.

3.3. Case Study: Air Quality Monitoring in Uruguay

To evaluate our approach, we apply it in a case study based on the **national agency responsible for air quality monitoring in Uruguay**, focusing on data collected near **industrial regions**. This setting presents unique challenges regarding data validation, provenance, and cross-institutional data sharing, further highlighting the need for robust, FAIR-aligned data management solutions. As this is an ongoing research effort, we present a **prototype implementation** along with **preliminary results** that showcase the feasibility of our approach. The current system provides an initial framework for integrating and managing air quality data, and future work will focus on refining its scalability, improving inference mechanisms for data quality assessment, and expanding interoperability with external environmental datasets.

4. Case Study: Data Management in Uruguayan Air Quality Monitoring

Air quality monitoring in Uruguay is managed by national and local agencies, with the Ministry of Environment (MA) overseeing regulations and compliance. Several initiatives provide public access to air quality data but face challenges in standardization and data quality. The **Observatorio Nacional Ambiental** (OAN)[11] compiles environmental indicators, including air quality, from governmental and third-party sources. However, it lacks machine-readable formats, comprehensive metadata, and cross-institutional integration. At the local level, **Montevideo Air Quality**[12] offers real-time data

via an open-access platform, focusing on public awareness. Its limitations include restricted historical data and the absence of a standardized data model.

Air quality data collection in Uruguay is fragmented, with institutions using heterogeneous formats and varying management criteria. The lack of a shared ontological framework hinders interoperability, while sensor drift, missing values, and inconsistencies affect data reliability. No unified methodology ensures traceability, versioning, or comprehensive quality assessment. Existing platforms prioritize data publication but offer limited analytical tools for environmental professionals. Assessing pollution trends, detecting anomalies, and correlating industrial activity with air quality remain challenging. These issues not only impact decision-making but also overburden scarce human resources, leading to inefficiencies and compounding operational challenges.

4.1. Application of Our Approach on the Uruguayan Context

To implement a Proof of Concept, we utilized data provided by the MA for the years 2021 to 2023, sourced from twelve air quality measuring stations. The MA dataset includes information for 11.7M records on 10 different pollutants, collected from 12 different stations, all located on Uruguayan territory. It is relevant to notice that the dataset had been built on a few basic guidelines as was meant to be used internally in the MA in a particular ad hoc scenario and was not intended to be shared with third parties. These characteristics of the source dataset should make this implementation a relevant Proof of Concept. Much of the mandatory information needed to apply our approach is missing from the dataset, notably the nature of agreements between involved organizations, identification of agents (personnel, software, or other) who reported the register information, and the review processes the registers underwent. Some of this information was intentionally left blank for later inclusion, while default entities were created for other fields, such as personnel inserting data in the landing or validation stages.

4.2. Preliminary Results and System Prototype

We have implemented a prototype of our system, including: a data ingestion pipeline capable of processing real-world air quality datasets, a validation module that applies data quality checks to detect anomalies, and a semantic data model ensuring compliance with FAIR principles. Data Quality measures were implemented. Validity on the values inserted, basically verifying values are within sensor's range. Accuracy for valid measurements, checking outliers and completeness for sets of measurements, dividing the time window into a number of slots and verifying existence of measurements for each of the slots.

Preliminary results demonstrate improved data consistency and completeness compared to raw datasets, enhanced traceability of data sources and transformations that increase trust in the published data, and the ability to generate multidimensional reports for analyzing pollution trends over time. Upon loading the first subsets of the source dataset into the staging area, we issued SPARQL queries to identify registers that did not comply with basic requirements, such as values reported outside the sensor's operating range and multiple data points for the same sensor and time. For example, a data point representing the current valid reading for the pollutant PM10 (particulate matter under $10\mu m$) at the station DU_PDT2 (Paso de los Toros) with a reading time 2021-01-01T00:50:00 had previously been reported nine times over ten months. Morevoer, we can provide SPARQL queries to retrieve all the previous registers and the reasons why they where overuled by subsequent registers, a feature that is not currently available in the data management systems employed by the MA. Each register linked to the readings is also linked to an Agent (person, software, etc.), allowing for further analysis of the reasons behind these repeated observations. Our system is still in early-stage development, with ongoing efforts to scale the solution for real-time data streams, integrate external environmental datasets, such as meteorological information, and to improve visualization tools for technical staff and policymakers.

5. Conclusion

This work presents a knowledge graph-based approach to enhance air quality data management by integrating semantic web technologies. Our prototype demonstrates improvements in data consistency, traceability, and interoperability, addressing key challenges in standardization and quality assessment. Preliminary results highlight the feasibility of this approach in the Uruguayan context, showcasing its potential to facilitate FAIR-compliant data publication and enable better-informed decision-making. Future work will focus on scalability, refining quality assessment mechanisms, and expanding interoperability with additional environmental datasets.

Declaration on Generative AI

During the preparation of this work, the authors used Mistral, chatGPT-4 and Grammarly to do Grammar and spelling checks. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] S. Gulia, S. S. Nagendra, M. Khare, I. Khanna, Urban air quality management-a review, Atmospheric Pollution Research 6 (2015) 286–304.

[2] U. E. Program, Air quality monitoring and data management guidebook for the states of the gulf cooperation council, 2022. URL: https://www.ccacoalition.org/resources/air-quality-monitoring-and-data-management-guidebook-states-gulf-cooperation-council.

[3] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al., The fair guiding principles for scientific data management and stewardship, Scientific data 3 (2016) 1–9.

[4] W. Huang, T. Li, J. Liu, P. Xie, S. Du, F. Teng, An overview of air quality analysis by big data techniques: Monitoring, forecasting, and traceability, Information Fusion 75 (2021) 28–40.

[5] D. Singh, M. Dahiya, R. Kumar, C. Nanda, Sensors and systems for air quality assessment monitoring and management: A review, Journal of environmental management 289 (2021) 112510.

[6] M. U. H. Al Rasyid, I. Syarif, I. A. H. Putra, Linked data for air pollution monitoring, in: 2017 International Electronics Symposium on Knowledge Creation and Intelligent Computing (IES-KCIC), IEEE, 2017, pp. 65–70.

[7] J. Wu, F. Orlandi, I. Gollini, E. Pisoni, S. Dev, Uplifting air quality data using knowledge graph, in: 2021 photonics & electromagnetics research symposium (PIERS), IEEE, 2021, pp. 2347–2350.

[8] L. Galárraga, K. A. M. Mathiassen, K. Hose, Qboairbase: The european air quality database as an RDF cube, in: N. Nikitina, D. Song, A. Fokoue, P. Haase (Eds.), Proceedings of the ISWC 2017 Posters & Demoand Industry Tracks, volume 1963 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2017. URL: https://ceur-ws.org/Vol-1963/paper507.pdf.

[9] L. Etcheverry, A. A. Vaisman, Qb4olap: a new vocabulary for olap cubes on the semantic web, in: Proceedings of the Third International Conference on Consuming Linked Data, volume 905, CEUR-WS. org, 2012, pp. 27–38.

[10] A. Q. Gill, M. Bandara, Using knowledge graphs for architecting and implementing air quality data exchange: Australian context, in: Proceedings of the 25th Annual International Conference on Digital Government Research, 2024, pp. 534–541.

[11] Ministerio de Ambiente, Uruguay, Observatorio Nacional Ambiental, https://www.ambiente.gub.uy/oan/, 2024.

[12] Intendencia de Montevideo, Uruguay, Montevideo Air Quality, https://montevidata.montevideo.gub.uy/ambiental/calidad-del-aire, 2020.