Ingeniería y Competitividad
Vol 26 (3) 2024; doi: 10.25100/iyc.v26i3.13644

Vol 26 (3)

ISSN
0123-3033
e- 2027-8284

Original Article

Use of Model-Driven Architecture in the storage of PM 2.5 and public health data

Uso de Arquitectura Dirigida por Modelos en el almacenamiento de datos de PM 2.5 y salud pública

James A. Vergara-Correa[1] Jorge E. Giraldo-Plaza[1] Miriam Gómez-Marin[1] Juan P. Holguín-Marin[2] Nora A. Montealegre-Hernández[3] Juan G. Piñeros-Jiménez[3]

[1]Politécnico Colombiano Jaime Isaza Cadavid. Medellín, Colombia.
[2]Universidad Nacional de Colombia. Medellín, Colombia.
[3]Universidad de Antioquia. Medellín, Colombia.

Abstract

Introduction: this paper addresses the storage of data on health events and PM2.5 particles in the city of Medellín, Colombia. The consolidation of data from heterogeneous sources poses a significant challenge in this context.
Objective: the aim of this study is to propose a metamodel that facilitates the integration and storage of these data using a model-based approach.
Methods: a modeled approach was developed to identify common aspects for building a data warehouse. An abstraction layer was defined over the conceptual models of particulate matter and health events.
Results: the main result was the creation of a data warehouse prototype that allows for the efficient consolidation of data on PM2.5 and health events. This prototype demonstrates the effectiveness of the proposed approach in data integration.
Conclusion: it is concluded that using a model-based approach strengthens decision-making in public health policies and quality management strategies in the healthcare sector.

Keywords: metamodel, Data warehouse, Architecture, Air quality, Public health.

Resumen

Introducción: en este artículo, se aborda el almacenamiento de datos sobre eventos de salud y partículas PM2.5 en la ciudad de Medellín, Colombia. La consolidación de datos provenientes de fuentes heterogéneas representa un desafío significativo en este contexto.
Objetivo: el objetivo de este estudio es proponer un metamodelo que facilite la integración y almacenamiento de estos datos, utilizando un enfoque basado en modelos.
Métodos: se desarrolló un enfoque modelado que identifica aspectos comunes para la construcción de un data warehouse. Se definió una capa de abstracción sobre los modelos conceptuales de materia particulada y eventos de salud.
Resultados: como resultado principal, se obtuvo un prototipo de data warehouse que permite la consolidación eficiente de datos sobre PM2.5 y eventos de salud. Este prototipo demuestra la efectividad del enfoque propuesto en la integración de datos.
Conclusión: se concluye que el uso de un enfoque basado en modelos fortalece la toma de decisiones en políticas de salud pública y estrategias de gestión de calidad en el ámbito sanitario.

Palabras clave: metamodelo, Bodega de datos, Arquitectura, Calidad de aire, Salud pública.

How to cite?

Vergara-Correa, J.A., Giraldo-Plaza, J.E., Gómez-Marin, M., Holguín-Marin, J.P., Montealegre-Hernández, N.A., Piñeros-Jiménez, J.G. Use of Model-Driven Architecture in the storage of PM 2.5 and public health data. Ingeniería y Competitividad, 2024, 26(3) e-20513644

https://doi.org/10.25100/iyc.v26i3.13644

Recibido: 20-03-24
Evaluado: 16-05-24
Aceptado: 15-08-24
Online: 12-09-24

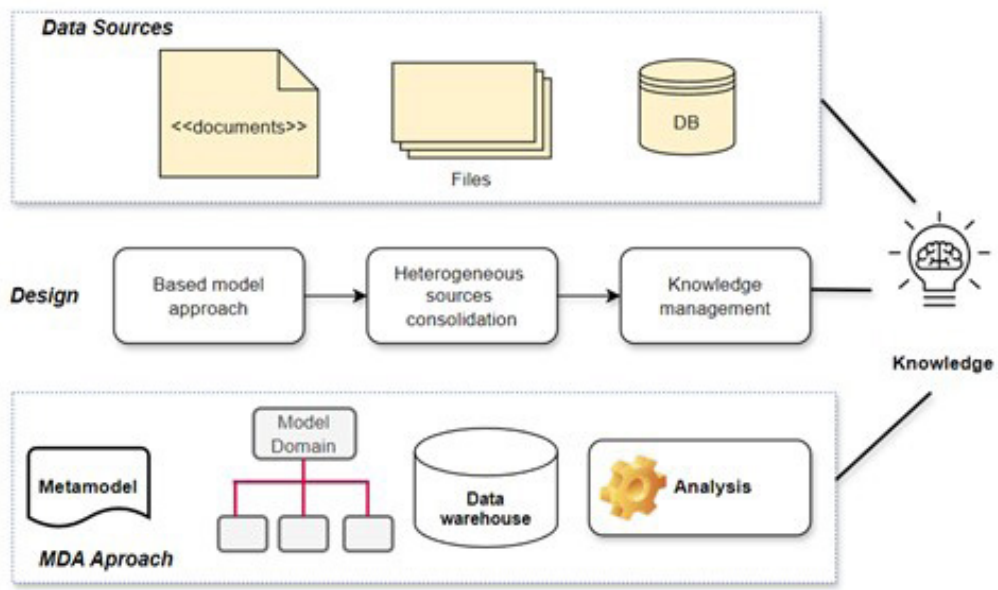
Correspondence
jegirado@elpoli.edu.co
Carrera 48 N° 7-151
Medellín El Poblado.

Why was it conducted?:
This research was made in a Interuniversity macro project financed by Colombian Government. The aim of this project is to improvement the innovation and research capabilities of states (Antioquia, Caldas) invoved.

What were the most relevant results?
The principal results of this project is a health and particular matter domain modelling. Additionally, we proposed a metamodel to consolidate the data of the domains. Too we get a first approach to analytic model.

What do these results contribute?
The results of this project can be used to generate a big data models of different domain related with public health.

Graphical Abstract



doi: 10.25100/iyc.v26i3.13644

Introduction

Data storage is a proccess that refers to the saving, preservation and organization of information, in a specific format, with accessibility capacity, for later retrieval and use by data users (1). In that order of ideas, data storage frames the technologies, instruments, processes and standards related to the management of information in digital format. This means that information is stored in different types of storage devices, whether hard disks, floppy disks or virtual disks in the cloud.

Storage also involves the use of databases, file systems, database management systems, use of external systems and/or dynamic data structure systems. This facilitates access to and processing of information from different sources (2).

Different approaches are used to address data warehousing, including: i) approaches based on domain ontologies (3), ii) based on business process management (4), and iii) based on data warehouses (***Datawarehouse***). These approaches have one characteristic in common, which is the prior analysis of the data, as well as the use of appropriate standards for data modeling.

One of these standards is the CWM (***Common Warehouse Metamodel***), which is a set of specifications for the standardization of the way database models (schemas), data transformation models, OLAP and data mining models, among others, are represented (5). The main objective of CWM is to facilitate the exchange of metadata, in a metamodel context, with a data warehouse in distributed heterogeneous environments.

Model Driven Architecture (***MDA***), is an architecture that unifies each step of the software development life cycle, using "Metamodels" to describe the functionalities and performance requirements of an application (6). MDA has use in the integration, transformation, translation and alignment of notations, formats and/or schemas with heterogeneous characteristics. For example, Saiji, et at. (7), who use MDA together with BPMN (***Business Process Managment Notation***) to semi-automatically transform conceptual models into specific models of a web-type platform.

Some related works in this area are those of Sun, et al. (8) and Azzaoui, et al. (9), where they propose the implementation of solutions based on MDA, for the construction of knowledge graphs from relational databases and for the generation of multidimensional schemas from data warehouses and with a metamodel. Also Belkadi (10), design a solution for the transformation of business rules described in SQL to a NoSQL database.

Another related work is presented by Xie et al. (11), which proposes a mechanism for mining data from heterogeneous sources, using MDA. Hanine et al. (12) also present a novel method for the construction of conceptual schemas from relational databases, using, as in the previous work, a multidimensional approach. Esbai et al. (13), for their part, have developed an approach based on MDA for the automatic generation of data warehouses from business rules and performance indicators.

These works reflect the potential that MDA has for data management, processing, storage and integration, mainly with the use of database architectures, specifically, data warehouses. The latter facilitate the work with MDA, since their internal processes and

doi: 10.25100/iyc.v26i3.13644

structures can be seen as part of the model-based approach.This is why the motivation of being able to integrate data for storage related to particulate matter and health vents, a key aspect for decision making in the air quality management cycle.

Given the magnitude and complexity of the data generated in different scientific areas, among them, the operation of air quality monitoring and surveillance networks, in relation to different pollutants such as particulate matter (PM), a criteria pollutant with high impact on health, the information generated in its measurement is robust, being feasible to analyze it under a model-based approach.

A area that is directly related to PM is the management of health events, since it is precisely the effects generated by the presence of PM that affect people's health. Health events are recorded from information on patient attendance at health services in the cities, whether they are emergency or basic care services (14). From the analysis of health events, it is possible to design surveillance systems that allow the detection of the behavior of a community and, from this, strategies for its improvement can be defined (15).

Health surveillance systems are characterized by a systemic approach to the collection, analysis and interpretation of health data, which can be accessed periodically. The main objective of these systems is the detection and monitoring of diseases and other important events for the health of a given population (15). With this it is possible to perform early detection of health problems, evaluation of the impact of public health interventions and the recording of evidence for decision making in the field of health.

However, despite advances in the management and processing of data related to particulate matter and health events, these are still characterized by high complexity, given the diversity of formats, heterogeneous structures, different access mechanisms and different storage architectures. Because of the above, it is complex to carry out processes such as data analysis, intelligent information processing and knowledge management.

This paper presents a structural proposal based on MDA for the storage of data related to PM2.5 particulate matter and health events, which are being generated in the city of Medellin. as part of two research processes that are being led by several higher education institutions in the city together with the environmental authority and health sector entities. From the design of a metamodel and using the CWM standard, which allows its use in data warehouses and the design of an architecture that supports knowledge management.

Methodology

For the development of this research, the use of MDA for the storage of environmental health data is proposed. The focus is the participation of experts and the construction of artifacts (conceptual diagrams) that allowed the identification of the Platform (*Platform Independent Model -PIM-*).

From this, a metamodel was proposed, which in turn allowed the definition of instances of a Platform Definition Model (PDM) and thus finally achieve an approximation to the Platform-Specific Model (PSM). Figure 1 summarizes the work methodology, detailing the activities carried out and the artifacts generated in the application of the MDA.

doi: 10.25100/iyc.v26i3.13644

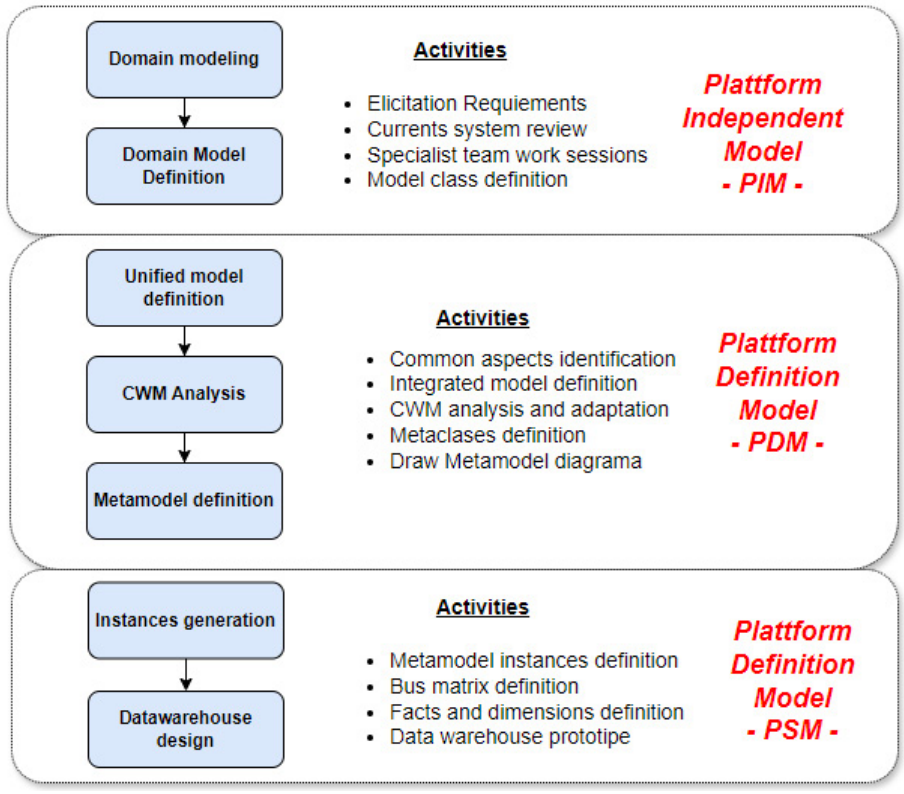


Figure 1. Work Methodology for MDA application. Source: author

As a starting point, a conceptualization of the domains was carried out, in which information was gathered and the data storage platforms and their representation schemes were recognized. Then, together with the experts of each domain, the conceptual models were built in UML class diagrams.

Once the conceptual models were built, the common aspects were identified in order to obtain a unified conceptual model. In addition, a conceptual integration was achieved based on a data abstraction approach, i.e. the definition of metadata. With this, an extension to the CWM standard was defined, based on the definition of meta classes and a diagram of the metamodel.

Once the metamodel was obtained, we proceeded to the generation of instances related to the air quality and public health domains. This was achieved with the logical design of the data warehouse, which includes the definition of the dimensions and their respective facts. With this, it was possible to identify key elements for the unification of the models, such as the date.

As a working mechanism for validation, a prototype-based approach was used. The conceptualizations and decisions about the project were worked on conceptual diagrams, which were refined - as prototypes evolved - in periodic meetings. It should also be noted that validations were carried out based on the opinion of experts and on the contrast with data sources.

Ambient domain model - PM2.5 particulate matter characterization

Figure 2 presents the conceptual model of the chemical characterization of PM2.5. PM

doi: 10.25100/iyc.v26i3.13644

is defined as a complex mixture of organic and inorganic, solid and liquid substances suspended in the air with aerodynamic diameters between 50 µm and less than 2.5 µm (PM2.5) (16), the latter having undesirable effects on the health of its inhabitants, inhalable due to its capacity to enter the pulmonary alveoli, accumulate, pass through the pulmonary mucosa, enter the lung and in some cases be transported through the bloodstream and reach other types of organs. For this reason, it is considered one of the most harmful atmospheric pollutants due to its effect on local and regional deterioration of air quality and severe effects on respiratory and cardiovascular diseases. This affectation is directly related to its chemical composition including diverse elements and compounds, mostly organic matter including polycyclic aromatic hydrocarbons, sulfates, nitrates, ammonium, sodium chloride, minerals and water, whose association with an estimated 7 million deaths per year (17), is today considered of high importance for the consolidation of robust and valid information on this subject in relation to health events. The impact on the cellular system is biological, genotoxic or cytotoxic.

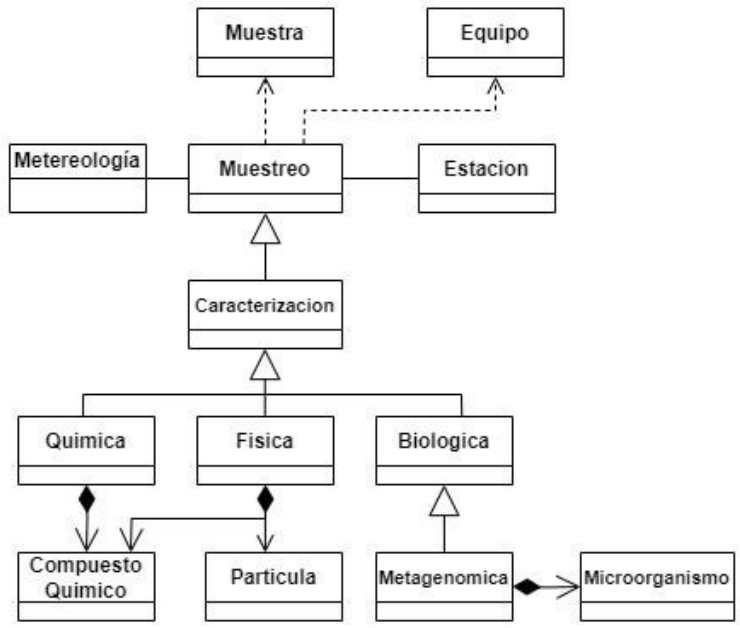


Figure 2. Conceptual model of chemical and microbiological characterization from PM2.5. Source: author

This diagram specifies that special equipment is used to collect PM2.5 samples, which are located at monitoring stations. Once the samples are collected, they are taken to the laboratory for characterization. The types of characterization are: chemical, physical and biological. Then the characteristics of the sample are associated, such as its label and concentration, and by means of the sampling date, meteorological characteristics that interfere in the analysis are related. Once the samples have been recovered and made available to the laboratory, they are classified according to their label and lot, which makes it possible to identify the type of characterization that will be carried out for each sample.

For the chemical characterization, information is available for all the components analyzed, determining in each of the samples their concentration (%) and in micrograms per cubic meter (µg/m3). For the physical characterization, the particles to be studied are linked; therefore, all particles reported in the physical characterization must be registered as particles.

doi: 10.25100/iyc.v26i3.13644

Microbiological characterization can be of the metagenomic type, and has associated genotoxic or cytotoxic impacts. With respect to genomic analysis, the situation is very similar to that mentioned in the chemical and

physical characterizations, where a class of microorganisms is available from DNA. In this case, the main variables of interest will be to determine the number of microorganisms and genes in each sample. Finally, the genotoxic analysis seeks to determine the mutagenic potential present in each batch of samples.

It is now possible to obtain genomic information directly from microbial communities in their natural habitats in order to infer taxonomic and functional profiles of a microbial community. Thus, metagenomics provides the ability to study microorganisms at the genomic level in order to understand the relationships between microorganisms, communities, and the habitats in which they live (18).

Metagenomic studies associated with air quality are of great importance for understanding the potential impact of microorganisms on human health through periodic monitoring of air-associated microbiota. These studies may also lead to the discovery of new genes or metabolic pathways relevant to industrial, meteorological, environmental bioremediation, and biogeochemical cycling applications (19).

Some studies have associated microorganisms present in the atmosphere to different human diseases, for example, fungi identified as causing respiratory problems (20), lymph node syndrome, Kawasaki disease (21), as well as endotoxins from airborne bacteria have also been associated with health problems (22).

Until recently, many of the microbial diversity studies in air relied on culture media-based methods, however, culture-independent methods using DNA sequencing technologies are widely employed lately. This technology allows a broader picture of the diversity of microorganisms and thus compares their temporal variation under different meteorological conditions.

In this way, metagenomics provides the ability to study microorganisms from the genomic level in order to understand the relationships between microorganisms, communities, and the habitats in which they live.

Health event domain model

For MinSalud (23), a health event is related to “circumstances that may affect the health situation of an individual or community. Health events are classified into physiological conditions, diseases, disabilities and deaths; protective factors and risk factors related to environmental conditions, consumption and behavior; specific protective actions, early detection and care of diseases and other associated determinants”.

Figure 3 presents the conceptual model of health events, represented in a UML class diagram, which shows the different stages of collection and integration of the information of this dimension, as a result of the implementation of a cohort study in environmental health (24), whose objective is to investigate the relationship between exposure to particulate matter and the occurrence of health events in a population over time (25).

doi: 10.25100/iyc.v26i3.13644

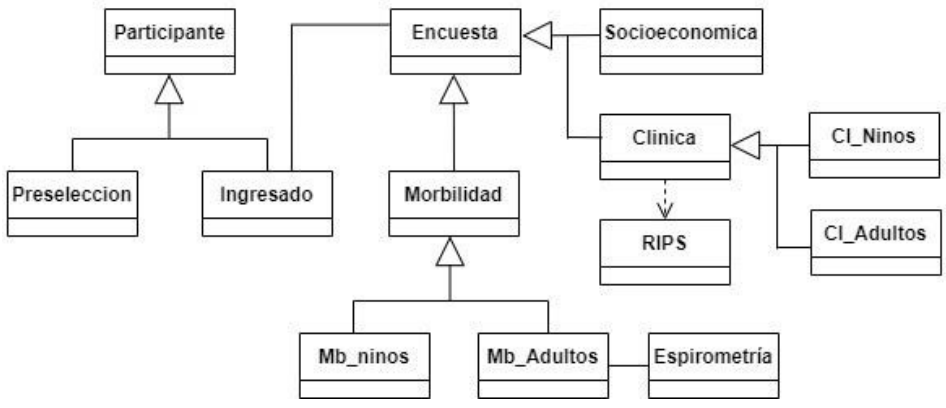


Figure 3. Conceptual Model of Health Events. Source: author

This study has formed a fixed cohort of men and women in the age groups under 15 years and over 44 years, who were previously selected, considering aspects such as their clinical history, time of residence in the study area and their acceptance to participate and to provide information on their health status and living conditions and authorization to access their health service care records. The selection of participants is part of a cohort study, and their characteristics are not part of the study design.

Participants entered the cohort based on a pre-selection process of families from the study areas. The participants were selected from the study areas. The persons identified as having entered the cohort were subjected to different data collection procedures: information-gathering procedures: physical examination, clinical survey according to their age group. socioeconomic surveys, and felt morbidity surveys.

We also accessed the Individual Health Services Provision Records (RIPS), which compiles all the care provided to an individual. Additionally, in a subgroup of participants, the health information was complemented with pulmonary function information by means of Spirometric tests (26). Finally, to broaden the spectrum of analysis of the health impacts of PM2.5 to a molecular scale, a genotoxicity analysis was performed using the Ames test, which seeks to determine the mutagenic potential present in each batch of PM 2.5 samples at times of environmental contingency.

Application of MDA

MDA allows the creation of highly abstract, machine-readable models that are developed independently of the implementation technology and stored in standardized repositories. Tools can access them repeatedly and automatically transform them into schemas, code skeletons, test cases, integration code and deployment scripts for various platforms (27).

Figure 4 presents the traditional OMG modeling architecture, which consists of a hierarchy of model levels, each (except the top one) characterized as “an instance” of the top level. The lower level, also referred to as M0, contains real-world elements corresponding to the “user data”, i.e., real data objects for which the software has been designed to manipulate, while level M1 contains models of the real-world objects. Level M2 contains models of the models in M1, and level M3 contains models of the models in M2 (28).

doi: 10.25100/iyc.v26i3.13644

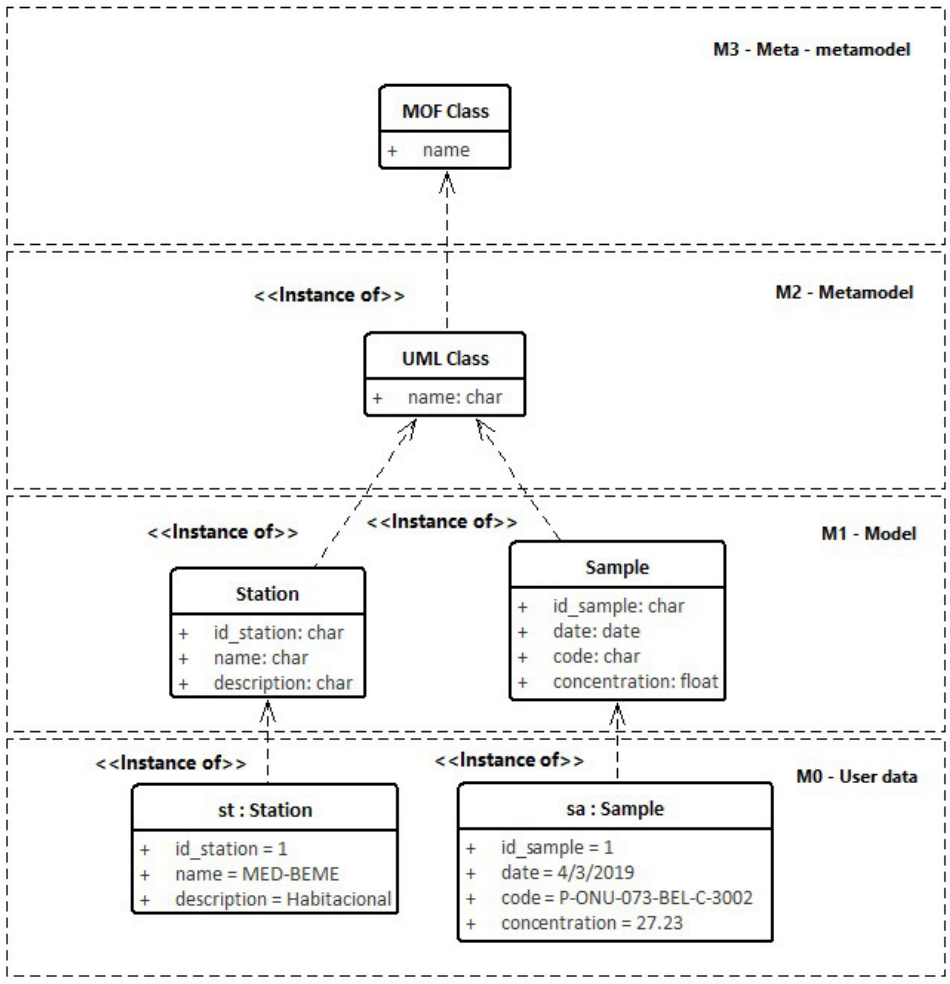


Figure 4. Arquitechture of 4 levels proposed by OMG. Source: author

For the definition of the metadata-oriented metamodel, use is made of the Common Warehouse Metamodel (CWM) standard, aligned with that proposed by the Meta-Object Facility (MOF). The latter provides the modeling standard resides in the M3 layer. Specifically for the problem of PM2.5 characterization data warehousing and health events, the layers of the MDA architecture are as follows:

Layer M0: This layer has as its initial input the structured data from each of the domains, represented in turn as databases, spreadsheets, flat files, and physical files. This layer is the lowest hierarchy of the models and represents the most concrete and detailed level of the metamodel.

For the PM2.5 domain there is a representation of data by means of flat files in .cvs and .txt format. The PM2.5 data and the characterization of its samples, includes the processing of the source files to be interpreted by the domain models. In the case of data related to the health events domain, the surveys are stored in relational databases linked to running applications. Likewise, spreadsheets are available for automation in macros.

Layer M1: E This layer comprises the instance of the metamodel, i.e. the defined domain models (characterization and health events), together with their respective business rules

doi: 10.25100/iyc.v26i3.13644

with which it is possible to reason in an integrated manner. Mainly, a conceptual model was obtained from the data present in the M0 ayer. The logic and basic functionality around the data were reflected in these models.

In this M1 layer, there is a platform-independent view, since it describes the logic and behavior of the system, without going into specific details. An advantage of this is the ability to generate schemes that can be exploited by other domains, which have not been taken into account so far, such as meteorological or satellite information.

To model the M1 layer, a specific notation was taken into account, this was the UML class diagrams, with which a formality accepted by the community is achieved. In addition, UML includes extensions for the representation of meta-classes and meta-meta-classes, which facilitates the definition of schemes for adaptation to specific models of the platform.

Layer M2: The proposed metamodel is present in this layer. Mainly, from this layer it is possible to generate instances where common aspects of the models of layer M1 are visualized in an integrated way. That is to say, from layer M2, a specific representation of the platform can be made, in this case for data warehouses.

The models generated in the M2 layer are an abstraction of the M1 layer, where the specific data execution platform, either a data schema or a specific technology, is reflected. Specifically, for the research, the metamodel is defined based on what is proposed by the CWM standard. Thus, the M2 layer instances are structured in such a way that they can be easily carried in data warehouses.

From this layer, it is possible to obtain a detailed implementation of its instances, since it is expected that the models generated in the M2 layer serve as a basis for the generation of source code that can be implemented in software tools. This means that the capacity to rewrite the execution logic according to the defined platform is strengthened.

Layer M3: Not applicable to the scope of this research.

Results and discusion

Proposed Metamodel (Layer M2)

The proposed metamodel is based on the metaclasses schemas, tables and columns. With these metaclasses, instances of the metamodel can be generated that are appropriate for the integration of data from the characterization and health event domain models. The metamodel that complements the CWM standard was then defined. Figure 5 presents the proposed metamodel in UML notation. The CWM standard is defined as the central package, which is complemented by an integration schema management model and a model for managing the data generated in the domains.

Regarding the domain models, the introduction of the Factory Method design pattern in the metamodel efficiently presents how to manage the integration of the different domain information, specifically in the areas of environment and health. This pattern was applied based on the experience of the work team. The use of the Factory Method pattern This approach effectively allows the creation of objects, facilitating the extensibility of the metamodel by incorporating new domain schemas in a simple and organized manner. The adoption of the Factory Method pattern provides flexibility and scalability to the model, where the creation of meta class instances, related to the environment and health domains, is centrally managed,providing a consistent standard for the incorporation of future domain schemas.

doi: 10.25100/iyc.v26i3.13644

The adoption of the CWM standard in the metamodel design facilitates the coherent and structured representation of metadata describing both relational databases and the complexity of a data warehouse. Thus, a "star" schema can be represented as a relational database. Typically, data warehouses are relational databases, which are processed in a relational database management system (29).

The proposed design provides a unified framework for describing diverse domain schemas, allowing the seamless integration of data corresponding to the environment and health domains, as well as the incorporation of other possible domain schemas in the future. The standardization offered by CWM simplifies the relationship of metadata, allowing for efficient management of information in the data warehouse.

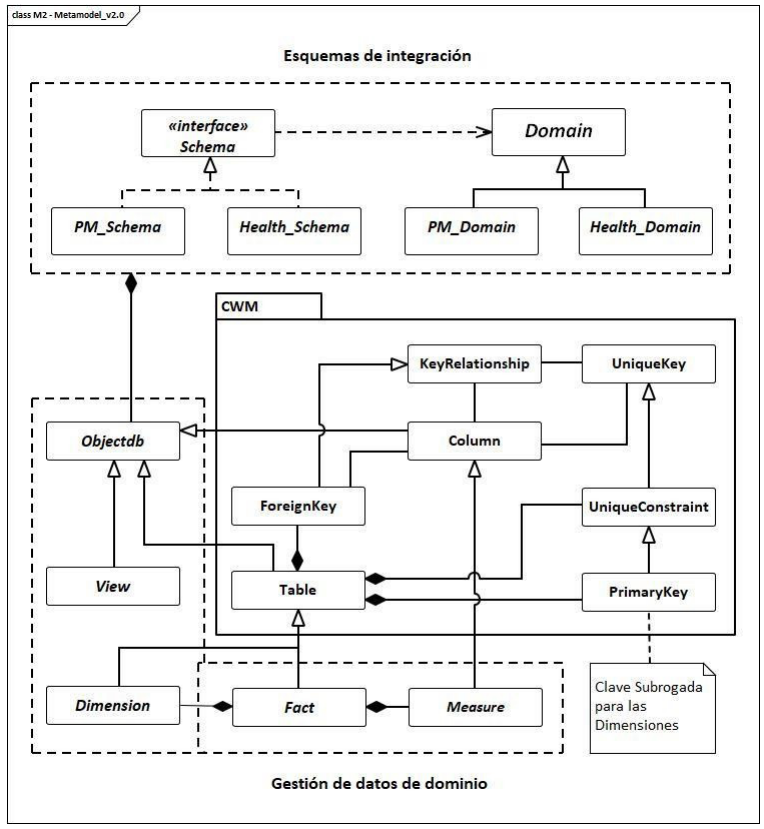


Figure 5. Proposed Metamodel. Source: author

This metadata-based approach has the advantage of ensuring consistency in data relationships, regardless of the complexity and diversity of domain schemas, by uniformly defining key database objects such as tables, views or columns, as well as the constraints between them, which facilitates the understanding and maintenance of the metamodel. Likewise, the flexibility of the CWM standard provides a solid and scalable foundation, allowing the metamodel to adapt to new requirements and domain schemas without compromising the integrity and consistency of the data warehouse. In this sense, the columns that act as primary keys in the dimension tables are surrogate keys, artificially generated to identify each record uniquely in the dimension table (30).

Finally, the implementation of a multidimensional approach as part of the metamodel design of the data warehouse, represents in a comprehensive and structured way, the

doi: 10.25100/iyc.v26i3.13644

management of the data generated in the domains by using dimensions, facts and measures, allowing a clearer understanding of the relationships between the stored data. In this way, it is possible to classify and organize the data into relevant categories that describe the domains, as well as the events or phenomena that occur in them.
The design also introduces a metaclass, whose instances refer to the variables that allow the quantification and analysis of data, facilitating the storage of heterogeneous information in a coherent manner, as well as enriching the metamodel by allowing data related to the environment and health domains to be represented in a more complete and understandable way for decision making.

Implementation and validation

Based on the proposed metamodel, the preliminary design of a prototype of a data warehouse is presented to validate the instances generated from the metamodel. Initially, the design of the bus matrix is presented, to later define the star model that represents it.

Proposed bus matrix

In the field of research on the impact of air pollution on human health, the construction of a bus matrix is a fundamental resource designed with the primary purpose of facilitating the integration and analysis of multidimensional data, thus allowing a deeper understanding of the relationship between exposure to air pollutants and health effects. In a bus matrix, granularity refers to the level of detail at which facts or measurements are recorded, while hierarchy relates to the structured organization of dimensions at different levels of aggregation or detail (31).

Table 1 presents data from various sources, including measurements of contaminant concentrations (expressed in micrograms per cubic meter), meteorological characteristics (temperature, wind speed and direction, rainfall and radiation), and contextual attributes (time, geographic location of monitoring stations, details on study participants, information on monitoring equipment, and specific chemical elements present in the particulate matter samples).

Table 1. Bus matrix for the design of the data warehouse

Dimenssions		**Facts**	
		Deaths	
Time	●	●	●
Geography	●	●	●
Station	●		
Equipment	●		
Element	●		
Participant		●	●
Diagnosis		●	●
Service			●
Medication			●
Procedure			●

doi: 10.25100/iyc.v26i3.13644

The cells of this matrix store specific data representing the concentration of PM2.5 particles in a given geographic location and at a given time, as well as the number of medical care and deaths related to exposure to that concentration. The dimensions (Time and Geography) allow these data to be organized and contextualized, which facilitates their multidimensional analysis.

For the validation of the metamodel, P.M.2.5 pollutant concentration data were collected with a sampling frequency of every 3 days. The dimensions of time, geography and monitoring station are included. Each of these dimensions has hierarchies that allow data analysis at different levels of detail. As an example, the “Time” dimension has a hierarchy that ranges from the highest level (year) to more detailed levels (month, day) to allow temporal analysis at different scales. The “Geography” dimension includes hierarchies ranging from the highest level (Municipality) to more detailed levels (Commune, Neighborhood, Block, Coordinates of the monitoring point).

The columns of this matrix represent the facts to be analyzed according to the dimensions that describe them. For the three facts an analysis will be required from the dimensions of Time and Geography. To describe the fact of the Concentration level of the pollutant P.M.2.5 it will also be necessary to include aspects such as the Monitoring Station and Equipment where the sample was collected, and additionally, the Elements that compose each sample.

On the other hand, the facts related to morbimortality (both care and deaths) require the addition of dimensions that present contextual data of the Participant of the cohort and its determined Diagnosis, while those recorded in the medical records obtained during the process of care in the different health Services and the Medications and Procedures provided will be essential only in the fact of Care.

The construction of the data warehouse for the integration of health and environmental information, as an instance of the data metamodel, poses significant challenges in terms of heterogeneity of data sources, quality, privacy and scalability. For this purpose, a star schema was defined based on the proposed bus matrix. Figure 6 shows the star schema defined, whose main fact is Morbidity, on which the Geography, Participant and Time dimensions depend.

doi: 10.25100/iyc.v26i3.13644

Prototype of a star schema for the data warehouse

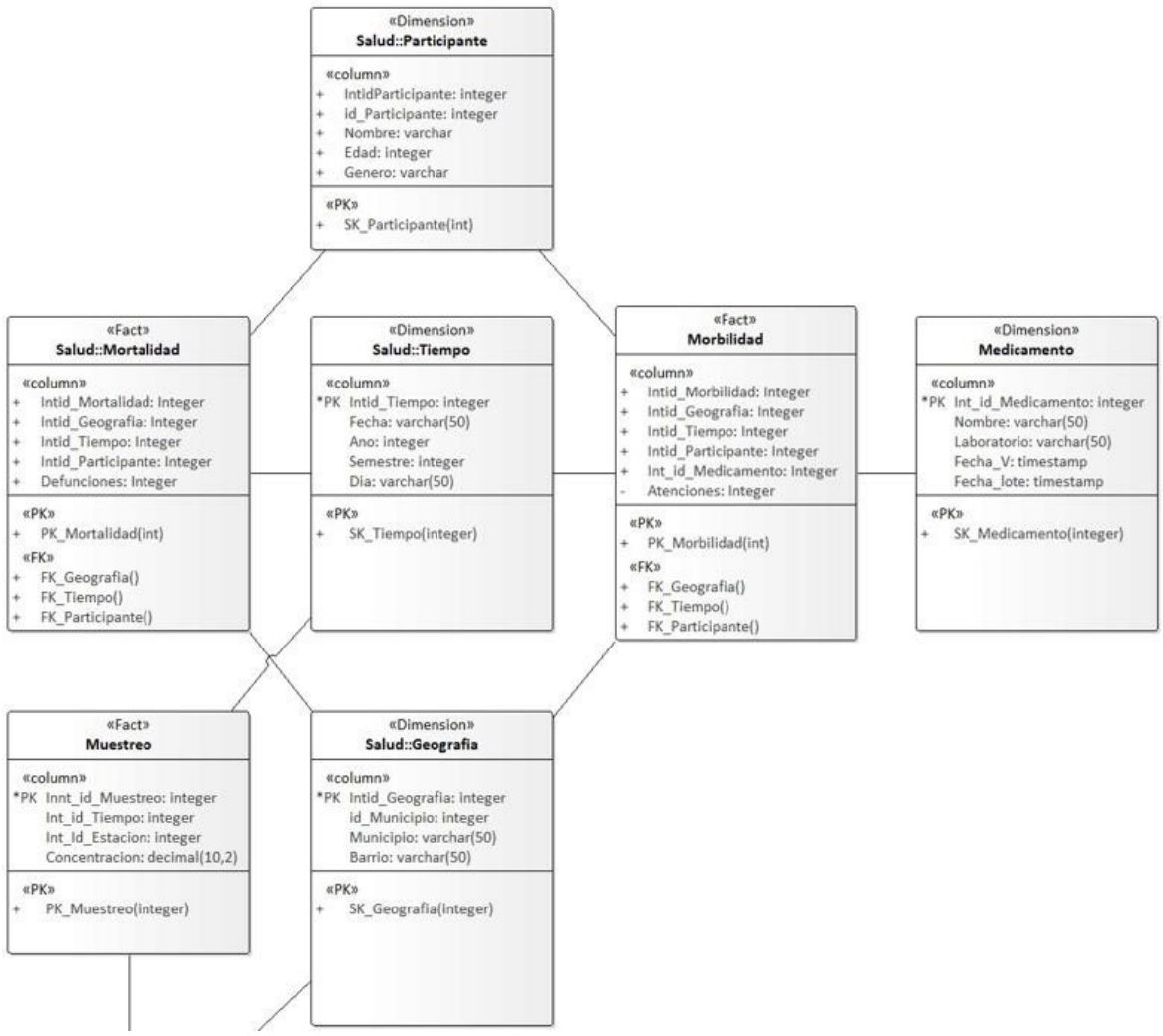


Figure 6. Star Schema. Source: author

Managing missing data and correcting recording errors are essential procedures to avoid biases in the analysis. In addition, data availability presents challenges in terms of geographic coverage and temporal continuity, requiring effective data collection and processing strategies. The complexity of the relationships between contaminant exposure and health effects demands advanced analytical approaches to capture nonlinear relationships and control for confounding factors. Similarly, privacy and security of participants' personal data are key considerations.

In addition, scalability of the data infrastructure and ongoing record keeping are critical as more data accumulates over time. Finally, effective communication of results to technical and non-technical audiences is essential to translate findings and informed decisions into public health and air quality management. The main difficulty of data management at the

doi: 10.25100/iyc.v26i3.13644

M0 level of the metamodel is the constant modification and risk of data corruption. This is reflected in changes of sources and file extensions. Likewise, there is a dependence on maintenance and privacy processes. These difficulties can affect the other layers, hence the importance of parameterization. In the case of environmental information, different types of data are also available, according to the chemical characterization, the levels of each specific element and compound and other potential analyses in relation to the grouping of these elements and compounds into more robust profiles according to specific sources of contamination.

The temporal component must be kept in mind in this integration process; a longitudinal design implies the integration of environmental and health data in different time frames, i.e., data taken at various points in time. Health data is available from 2008 to date, from different sources, which have been taken at various times or at a single moment, depending on the objectives of the study. Likewise, the proposed metamodel must be able to integrate nested designs to the analyses, for example, cross-sectional analyses to the cohort, nested cases and controls, panel studies or cohort-cohort design. This will allow the metamodel to expand the possibilities for data integration.

On the other hand, the advantages inherent in this construct are substantial. The synergy between health and environmental data can lead to a more holistic understanding of the causal relationships between exposure and human health effects. This, in turn, can more accurately inform the formulation of public policy and health strategies, and provide the opportunity for more sophisticated scientific research.

Early detection of adverse public health trends related to air quality can provide a preventive advantage and promote more informed decision making, while transparency in the availability of data and models can foster public and scientific confidence in environmental and health mitigation efforts.

Conclusions

The work presented highlights the possibility of consolidating, through a metamodel, data related to PM2.5 particulate matter and public health events. From this, it is possible to store data in a conceptually coherent manner in order to facilitate decision making and promote health protection programs.

The results derived from this analytical approach can be used as a solid basis for decision making in public health policies and air quality management strategies. The information obtained can contribute to the design of more effective pollution control measures and to the formulation of informed recommendations for the preservation of the health of the exposed population.

The adoption of MDA and especially CWM, has proven to be an effective architectural design approach for the representation of a metamodel where metadata describing a data warehouse is presented, however, the integration of DDD (Domain Driven Design) principles could be considered to improve the identification and delimitation of specific domains within the data warehouse. DDD provides a robust design approach to understanding the business, allowing domain boundaries to be clearly established for a consistent representation of the data.

doi: 10.25100/iyc.v26i3.13644

Likewise, frameworks such as DIF (Data Integration Framework) or MIS (Metadata Integration Strategy), represent areas of interest for future research, with the objective of approaching efficient data integration, considering the variety of data sources that can be stored in a data warehouse, as well as incorporating effective strategies to address metadata in an efficient manner.

CRediT authorship contribution statement

James A. Vergara-Correa: Investigation, Methodology, Software, Writing – original draft, Writing – review & editing. Jorge E. Giraldo-Plaza: Investigation, Project administration, Software, Writing – original draft, Writing – review & editing. Miriam Gómez-Marín: Conceptualization, Funding acquisition, Resources, Supervision. Juan P. Holguín-Marín: Data curation, Methodology, Validation, Visualization. Nora A. Montealegre-Hernández: Formal analysis. Juan G. Piñeros-Jiménez: Conceptualization, Resources.

Financing

This article was financed by the General Royalties System of the Republic of Colombia, through the institution strengthening project entitled "Development of a knowledge management program on air pollution and its effects on health in the Aburrá Valley Antioquia" with BPIN code 2020000100410.

Conflict of interest

The authors declare that they did not receive resources for the writing or publication of this article.

Ethical implications

The authors do not have any type of ethical involvement that should be declared in the writing and publication of this article.

References

[1] Elmasri, R, Navathe S, Castillo V, Pérez G, Espiga, B. Fundamentos de sistemas de bases de datos. Earson educación; 2007.

[2] Jarke, M, Lenzerini, M, Vassiliou, Y, Vassiliadis, P. (2002). Fundamentals of data warehouses. Springer Science & Business Media.

[3] Olivé A. A universal ontology-based approach to data integration. Enterprise Modelling and Information Systems Architectures (EMISAJ), 13, 110-119; 2018

[4] Dumas M, La Rosa M, Mendling J, Reijers, H. Fundamentals of business process management. Springer; 2013..

[5] Poole, J., Chang, D., Tolbert, D., & Mellor, D. (2002). Common warehouse metamodel. John Wiley & Sons; 2002.

[6] Object Management Group Model Driven Architecture (MDA). OMG MDA Guide rev. 2.0; 2014.

[7] Sajji A, Rhazali Y, Hadi Y. A methodology for transforming BPMN to IFML into MDA; Bulletin of Electrical Engineering and Informatics, 2022; 11(5), 2773-2782.

doi: 10.25100/iyc.v26i3.13644

[8] Sun S, Meng F, Chu D. A model driven approach to constructing knowledge graph from relational database. In Journal of Physics: Conference Series (Vol. 1584, No. 1, p. 012073). IOP Publishing; 2020.

[9] Azzaoui A, Rabhi O, Mani A. A model driven architecture approach to generate multidimensional schemas of data warehouses; 2019.

[10] Belkadi F, Esbai R. A Model-Driven Engineering: From Relational Database to Document-oriented Database in Big Data Context. In ICSOFT (pp. 653-659); 2021.

[11 Xie J, Xu F, Li Z, Li X. Data Mining Method under Model-Driven Architecture (MDA). Security and Communication Networks; 2022.

[12] Hanine M, Lachgar M, Elmahfoudi S, Boutkhoum O. MDA Approach for Designing and Developing Data Warehouses: A Systematic Review & Proposal. International Journal of Online & Biomedical Engineering; 2021;

[13] Esbai R, Hakkou R, Habri A. Modeling and automatic generation of data warehouse using model-driven transformation in business intelligence process. Indonesian Journal of Electrical Engineering and Computer Science Vol. 30, No. 3, June 2023, pp. 1866~1874 ISSN: 2502-4752, DOI: 10.11591/ijeecs.v30.i3.pp1866-1874

[14] Peláez O, y otros.as Bermejo, P. Brotes, epidemias, eventos y otros términos epidemiológicos de uso cotidiano. Revista Cubana de Salud Pública, 46, e2358; 2020.

[15] Mercuriali, L, Oliveras L, Gómez A, Marí, M, Montalvo T, Villalbí J. Un sistema de vigilancia de salud pública Para el cambio climático en las ciudades. Gaceta Sanitaria, 36, 283-286; 2022.

[16] United States Environmental Protection Agency. Particulate matter (PM) basics; 2017.

[17] Novillo-Ortiz D, D'Agostino M, Becerra-Posada F. El rol de la OPS/OMS en el desarrollo de capacidad en eSalud en las Américas: análisis del período 2011-2015. Revista Panamericana de Salud Pública; 2016; 40, 85-89

[18] Wooley J, Godzik A, Friedberg I. A Primer on Metagenomics. PLoS Comput Biol 6(2): e1000667.https://doi.org/10.1371/journal.pcbi.1000667;

[19] Behzad H, Gojobori T, Mineta K. Challenges and Opportunities of Airborne Metagenomics. Genome Biol Evol;7:1216–doi: 10.1093/gbe/evv064; 2015

[20] Grinn-Gofroń A, Strzelczak A. Changes in concentration of Alternaria and Cladosporium spores during summer storms. Int J Biometeorol. Sep; 57(5):759-68; 2013

[21] Rodó X, Curcoll R, Robinson M, Ballester, J, Burns, J, Cayan R., ... Morguí, J. Tropospheric winds from

northeastern China carry the etiologic agent of Kawasaki disease from its source to Japan. Proceedings of the National Academy of Sciences, 111(22), 7952-7957; 2014.

[22] Mueller-Anneling L, Avol E, Peters JM, Thorne PS. Ambient endotoxin concentrations in PM10 from Southern California. Environ Health Perspect. Apr; 112(5):583-8; 2004.

[23] Ministerio de Salud, S. D. S., & Inspección, S. (2006). DECRETO 3518 DE 2006 (OCTUBRE 09).

[24] Lazcano-Ponce E, Fernández E, Salazar-Martínez E, Hernández-Avila, M. Estudios de

doi: 10.25100/iyc.v26i3.13644

cohorte. Metodología, Sesgos y aplicación. Salud pública de México, 42, 230-241; 2000.

[25] World Health Organization. WHO global air quality guidelines: particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbonmonoxide: executive summary; 2021.

[26] Taylor J, Shrubsole C, Symonds P, Mackenzie I, Davies, M. Application of an indoor air pollution metamodel to A spatially-distributed housing stock. Science of the Total Environment, 667, 390-399; 2019.

[27] Kleppe A, Warmer J, Bast W. MDA explained: the model driven architecture: practice and promise. Addison-Wesley Professional; 2003.

[28] Atkinson C, Kühne T. Model-driven development: A metamodeling foundation. IEEE Software, 20(5), 36–41. https://doi.org/10.1109/MS.2003.1231149;

[29] Imran, S., Mahmood, T., Qamar, A. M., Siddiqui, A. J., Ahmed, I., & Shariq, N. (2024). NODW Framework for Data Warehousing-A NoSQL Big Data Perspective. Authorea Preprints-

[30] Wijaya, W., & Wiratama, J. (2024). The Implementation of Data Warehouse and Star Schema for Optimizing Property Business Decision Making. G-Tech: Journal Teknologi Terapan, 8(2), 1242-1250.

[31] Kimball R, Ross M. The data warehouse toolkit: the definitive guide to dimensional modeling. John Wiley &Sons; 2013

doi: 10.25100/iyc.v26i3.13644