

# CoMoDID: Combining explainable artificial intelligence and conceptual modeling for data intensive-domains management

Oscar Pastor<sup>1,\*</sup>, Diana Martínez Minguet<sup>1,†</sup>, Jose Fabián Reyes Román<sup>1,†</sup>,  
Alberto García S.<sup>1,†</sup>, Ana Leon<sup>1,†</sup>, Mireia Costa<sup>1,†</sup> and Ferran Pla<sup>1,†</sup>

<sup>1</sup>Valencian Research Institute for Artificial Intelligence (VRAIN). Universitat Politècnica de València, Camí de Vera S/N, Valencia, 46022, Spain

## Abstract

The large and heterogeneous data sets that characterize Data-Intensive Domains (DID) pose a challenge to developing data analysis and management approaches. A successful and efficient data knowledge extraction from DID-based systems is determined by assembling and analyzing such data sets, but integrating their different sources is arduous work. Finding sound solutions for this problem has become a relevant research goal, that existing DID-based systems are not solving in a final, convincing way. To solve this problem, a conceptual characterization of the data sets that constitute DID-based systems is essential. The use of foundational ontologies and conceptual modeling provides an adequate strategy to face the complexity of this problem by clarifying the data structure that is to be analyzed and managed. In this project we tackle this principle, by defining a method grounded on a conceptual model to develop efficient DID-based systems, and by making use of a well-grounded combination of Explainable Artificial Intelligence (XAI) and Machine Learning (ML) techniques to perform data analytics. In addition, the characterization of a platform for the implementation of the method is going to be designed and developed. The project's chosen domain of application is genomics, specifically in predicting critical diseases before symptoms manifest. Leveraging XAI and ML with genomic information can contribute to the advancement of precision medicine, allowing for the prediction of future diseases based on the available genomic data. The ML dimension will cover the predictive knowledge (is a disease present in a patient?), while the XAI dimension will deal with the explainable part (why the patient has the disease).

## Keywords

Data-Intensive Domains, Conceptual Modeling, Explainable Artificial Intelligence, Precision Medicine

---

ER2023: Companion Proceedings of the 42nd International Conference on Conceptual Modeling: ER Forum, 7th SCME, Project Exhibitions, Posters and Demos, and Doctoral Consortium, November 06-09, 2023, Lisbon, Portugal

\*Corresponding author.


†These authors contributed equally.

✉ opastor@dsic.upv.es (O. Pastor); dmarmin@pros.upv.es (D. Martínez Minguet); jreyes@pros.upv.es (J.F. Reyes Román); algarsi3@pros.upv.es (A. García S.); aleon@vrain.upv.es (A. Leon); miscossan@pros.upv.es (M. Costa); fpla@dsic.upv.es (F. Pla)

ORCID 0000-0002-1320-8471 (O. Pastor); 0009-0002-3191-1969 (D. Martínez Minguet); 0000-0002-9598-1301 (J.F. Reyes Román); 0000-0001-5910-4363 (A. García S.); 0000-0003-3516-8893 (A. Leon); 0000-0002-8614-0914 (M. Costa); 0000-0003-4822-8808 (F. Pla)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

# 1. Introduction

Data has become an invaluable asset in today's society, and its production is unparalleled, continually increasing. This presents significant challenges for modern software platforms, which must store, analyze, and quickly provide access to data for numerous users. Consequently, various research fields related to data management and processing have undergone profound transformations [1]. One of the most current, relevant challenges in the software development context is dealing with DID-based systems, which require extensive and heterogeneous datasets [2] to create knowledge from data. To develop effective and efficient methods and facilities for data analysis and management, software developers must integrate complex, distributed, and heterogeneous datasets from increasingly diverse data-generating technologies (e.g., sensors, the internet, genome sequencing machines, and other sophisticated devices). Therefore, managing this massive amount of data to find the most critical and actionable pieces of knowledge has become a significant challenge.

A fascinating example of DID-based systems are those that analyze the human genome [3]. Understanding the human genome is a significant scientific challenge, requiring the application of sound conceptual modeling techniques to manage such complex systems adequately. The continuous generation of genomic data from improved sequencing technologies [4, 5, 6] necessitates selecting the right data management strategy for software platforms. Developing software systems to deal with these DID are key for a proper genome analysis that would lead to anticipating future illness in the human population [7].

To address these issues, this proposal will be grounded on an interdisciplinary scientific policy especially interested in combining two strong lines of research: conceptual modeling (CM) [8] and explainable artificial intelligence (XAI) [9]. To this aim, two main components need to be explored, designed, and developed: i) A method to deal with DIDs problem's management (the methodological perspective) correctly and efficiently, and ii) a "materialization" of the method in the form of a platform intended to assess the solution's value in a challenging and specially selected DID as the one related to the understanding of the human genome (the practical perspective).

In this scenario, applying a methodological framework based on XAI and CM to address DIDs concerns effectively becomes a relevant, promising strategy that forms the basis of the scientific approach used to achieve the project's major goal. On the one hand, CM is recognized as crucial for developing data-oriented computer systems, ensuring an accurate representation of the application domain independently of the system that will be developed to address a real-world problem. This is especially relevant when we want to "understand data" in a DID context, which in our case applies to genomics. On the other hand, there is the application of XAI principles [10, 11], which describe a system in which humans can easily understand the results that an AI system provides, focusing primarily on understanding exactly "how" and "why" decisions are taken to reach results [9, 12]. For DID-based systems, where the right representation of concepts becomes a crucial step, CM becomes the perfect partner for a useful XAI application [10] since by visualizing the relevant concepts, the structure of meaning people

use to understand the domain is clearly represented.

Our approach -both methodological (a method) and practical (a platform for the genomics domain)- is based on the group's expertise [13, 14], focusing on understanding data's true nature, employing CM techniques, and addressing challenges such as data volume and processing.

### 1.1. Details of the project

The project combines XAI and CM for Data Intensive-Domains Management (CoMoDID). It is a four-year project (Sept. 2022 – Dec. 2025). Currently, the Research team is constituted by Óscar Pastor López, Juan Carlos Casamayor Ródenas, Tanja E. Vos, Lluís-F. Hurtado, Encarna Segarra, Ferran Pla, Fernando García Granada, José F. Reyes Román (Postdoctoral Researcher), Alberto García Simón (Postdoctoral Researcher) and Diana Martínez Minguet (Predoctoral Researcher), in collaboration with the Genomics Team of the PROS research group. The project is supported by the Generalitat Valenciana through the CIPROM/2021/023 project.

## 2. Project goals, tangible outputs & expected outcomes

The research proposed in this project focuses on the design of solutions for DID's problems since existing frameworks to build DID-based systems lack a sound conceptual modeling grounding, and too frequently, ad-hoc implementations are built. Both the method and the platform to materialize the solution in order to show how it works for a selected DID conform to the two major objectives of this project:

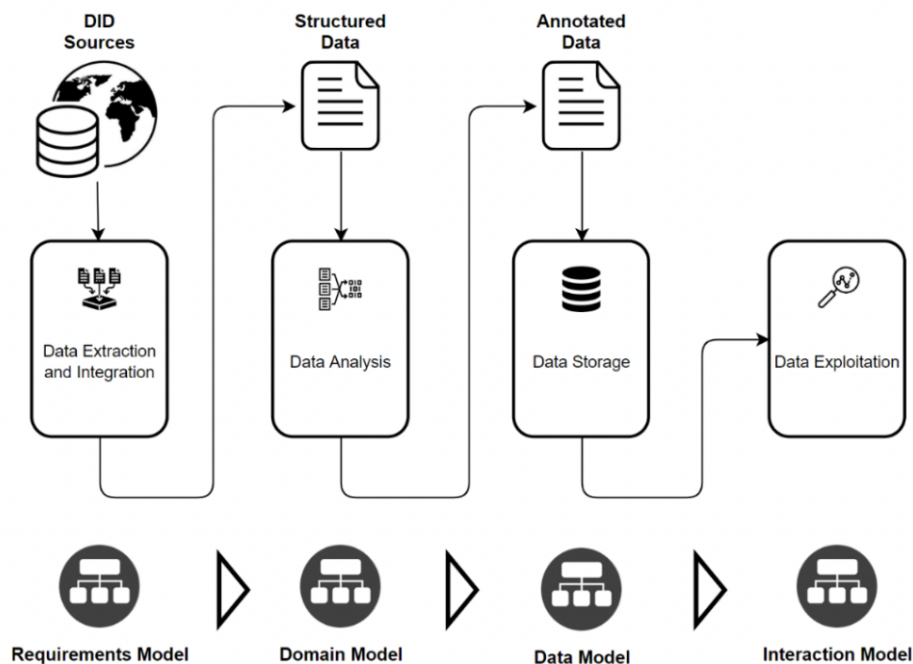
- Definition of a general method (the so-called DELFOS method (Figure 1), as the method to be used for the CoMoDID project) applicable to any DID for facing its analysis and design, which is based on a sound combination of conceptual modeling techniques and XAI technologies.
- Development of a technological platform, that will instantiate and support the method in a particularly challenging and complex DID context: the genomic domain.

To address the methodological and practical components of the approach, we break these objectives into specific goals (G) with associated work packages (WPs) which the tangible outputs obtained so far result from.

### 2.1. Specific goals for a general method

Figure 1 represents the different components that constitute the approach proposed by the DELFOS method, which is the adopted name for the method we aim to develop in this project proposal. The two specific goals to achieve our objective are:

- G1. **Ontological characterization of DIDs.** (WP1: *Ontological characterization of DIDs*): The study and analysis of existing foundational ontologies related to DID characterization in conjunction with the state of the art about existing solutions for the development of DID platforms are reflected in the doctoral thesis:



**Figure 1:** Components of the proposed DELFOS method. The first component oversees the data extraction and integration, based on the requirements represented in a conceptual model. With data structured and ready to be processed, the second component applies the most suitable XAI and/or ML techniques to satisfy knowledge requirements. The third component stores classified data using the adequate technology and data schema according to the analytical requisites. Stored data is used to build specific tools (fourth component), based on interaction models, that allow the extraction of knowledge.

- *García Simón, A. (2022). Understanding the Code of Life: Holistic Conceptual Modeling of the Genome. Universitat Politècnica de València. <https://doi.org/10.4995/Thesis/10251/191432>*

A preliminary definition of a foundational ontology for the development of DID platforms is published in:

- *Bernasconi, A., Guizzardi, G., Pastor, O., & Storey, V. C. (2022). Semantic interoperability: ontological unpacking of a viral conceptual model. BMC Bioinformatics, 23(11), 1-23. DOI: <https://doi.org/10.1186/s12859-022-05022-0>*

**G2. Integration of XAI techniques for data management and exploitation.** (WP2: *Integration of XAI techniques for data management*): Focusing on the case study of the genomic domain as one of the major DIDs that currently exist, a study and statistical comparison of different data sources with information associated with two groups of diseases: cancer and heart disease has been carried out. The results of these studies have been reported in:

- *Costa, M., García S, A., & Pastor, O. (2022). A Comparative Analysis of the Completeness and Concordance of Data Sources with Cancer-Associated*

**Information.** *In International Conference on Conceptual Modeling (pp. 35-44). Springer, Cham. DOI: [https://doi.org/10.1007/978-3-031-22036-4\\_4](https://doi.org/10.1007/978-3-031-22036-4_4)*

- Costa, M., García S, A., & Pastor, O. (2022). **Conceptual Modeling-Based Cardiopathies Data Management.** *In International Conference on Conceptual Modeling (pp. 15-24). Springer, Cham. DOI: [https://doi.org/10.1007/978-3-031-22036-4\\_2](https://doi.org/10.1007/978-3-031-22036-4_2)*

In order to replicate some of the criteria used by clinicians for genomic data to streamline the process of selecting relevant data by reducing the effort of manual activities performed by experts, different XAI techniques have been established according to the data collections to be analyzed, and different data storage and integration techniques have been evaluated.

- García S, A., Costa, M., Leon, A., & Pastor, O. (2022). **The challenge of managing the evolution of genomics data over time: a conceptual model-based approach.** *BMC Bioinformatics, 23(11), 1-33. DOI: <https://doi.org/10.1186/s12859-022-04944-z>*

(WP4: *Analysis and design of a tool to help with the writing of medical case reports in a genomic domain*): As a first draft, a Deep Learning (Transformers) model has been trained for a multi-class, multi-label classification task of radiology medical reports using Transfer Learning techniques:

- Contreras-Ochando, L., León, A., Hurtado, L.F., Pla, F., Segarra, E. (2023) **Enhancing Precision Medicine: An Automatic Pipeline Approach for Exploring Genetic Variant-Disease Literature** *The 4th International Workshop on Conceptual Modeling for Life Sciences (CMLS) @ER202 (Under revision)*

## 2.2. Specific goals for a technological platform

The instantiation of the method in the Genomics domain aims to validate that the method is a very complex DID, to provide a technological platform to collect, manage and analyze the generated data in practical settings in order to improve the understanding of the human genome challenge, and ultimately to obtain relevant value by the extraction of knowledge from the data. To achieve these objectives, the following goals are defined:

G3. **Definition of the interaction mechanisms for DID-based systems.** (WP3: *Interaction Definition for DID-based Systems*): The analysis of the interaction requirements for DID-based systems and the elicitation of such requirements to design sustainable interfaces are developed in:

- Bernasconi, A., García S, A., Ceri, S., & Pastor, O. (2022). **A Comprehensive Approach for the Conceptual Modeling of Genomic Data.** *In International Conference on Conceptual Modeling (pp. 194-208). Springer, Cham. DOI: [https://doi.org/10.1007/978-3-031-17995-2\\_14](https://doi.org/10.1007/978-3-031-17995-2_14)*
- García, A., Costa, M., León, A., Reyes, J. F., & Pastor, Ó. (2023). **Human-Centered Design for the Efficient Management of Smart Genomic Information.** *In Proceedings of the 18th International Conference on Evaluation of Novel Approaches to Software Engineering-ENASE (pp. 1-12). DOI: <https://www.doi.org/10.5220/0011635800003464>*

- Pastor, Ó., León Palacio, A., Panach, J. I., García, A., Costa, M. & Reyes Román, J. F. (2023). **Usability Evaluation of a Method to Analyze Data Intensive Domains**. *Multimedia Tools and Applications (To be published)*.

Remaining goals to be tackled are G4. **Development of a platform to support the DELFOS method** and G5. **DELFOF method and platform validation**, with the associated homonymous word packages (WP5 and WP6). These WPs leverage on all the previous ones which are in the process of improvement and development.

Finally, WP7: *Communication, dissemination and exploitation of the results*, is ubiquitous and ensures consistent dissemination, visibility and outreach to all relevant stakeholders.

The expected outcomes of the project are the development of a solid method that can be applied to any complex DID, and the implementation of a platform that enables the instantiation of the method for a particular DID-based system (genomics), providing the technological support.

### 3. Relevance for ER

The proposed project is aligned with several research topics relevant to the conceptual modeling community. It is highly relevant to the topics of Ontological and cognitive foundations and Semantics in conceptual modeling since the incorporation of foundational ontologies and conceptual modeling in the project contributes to a solid theoretical foundation concerning DID-based systems, in combination with the project's focus on developing standardized approaches for data integration and analysis which involves addressing semantic aspects. In the same line, the project is relevant to the topic of Complex management of large conceptual models, given that the project addresses the challenge of managing large and heterogeneous data sets in complex DID-based systems.

In another direction, the project aims to develop a method and platform to automate the development of DID-based systems, including data modeling. In this context, using Artificial Intelligence is useful for optimizing and automatizing data analysis. However, in the Precision Medicine field, where the practical instantiation of the project is embedded, the necessity of transparency and minimization of uncertainties is essential for the resulting decisions to be explainable. XAI satisfies these requirements, thus being suitable for data analysis counseling. The use of XAI and ML techniques involves knowledge representation and reasoning for accurate data analysis, being directly related to the topic of Logic-based knowledge representation and reasoning.

Overall, the proposed project's alignment with various research topics highlights its relevance and potential contributions to conceptual modeling, as well as knowledge representation and reasoning in the context of DIDs. It aims to address existing challenges and improve the efficiency and accuracy of DID-based systems, offering valuable insights for data analysts in diverse research fields based on conceptual modeling techniques and foundational ontologies.

## 4. Current Project Status

The project is in the first quarter of its development and is well on schedule. So far, the project tasks have involved the exhaustive characterization of the framework elements, as well as the development of precursory solutions. The ongoing tasks concern further investigation and the extension and improvement of the proposed solutions, for instance, the generation of a preliminary platform prototype and precursory validation tests for both the method and the platform. On the other hand, fruitful discussions with external companies have revealed future lines of research which are being addressed in a new Ph.D. thesis conducted within the scope of this project, regarding domain-specific concerns related to the genomic field.

## Acknowledgments

This work was supported by the Generalitat Valenciana through the CoMoDiD project (CIPROM/2021/023), through a GVA-Predoctoral Research Grant (ACIF/2021/117), a Margarita Salas Grant, and the Spanish State Research Agency through the DELFOS (PDC2021-121243-I00, MICIN/AEI/10.13039/501100011033) and SREC (PID2021-123824OB-I00) projects, and co-financed with ERDF and the European Union Next Generation EU/PRTR.

## References

- [1] A. Margara, G. Cugola, N. Felicioni, S. Cilloni, A model and survey of distributed data-intensive systems, 2022.
- [2] A. Elizarov, B. Novikov, S. Stupnikov, Data Analytics and Management in Data Intensive Domains, Springer International Publishing, 2020.
- [3] M. Felderer, B. Russo, F. Auer, On Testing Data-Intensive Software Systems, 2019, pp. 129–148. doi:10.1007/978-3-030-25312-7\_6.
- [4] M. Cowley, R. Davis, Next-generation sequencing and emerging technologies, Seminars in Thrombosis and Hemostasis 45 (2019). doi:10.1055/s-0039-1688446.
- [5] D. Rigden, X. Fernandez, The 27th annual nucleic acids research database issue and molecular biology database collection, Nucleic Acids Research 48 (2020) D1–D8. doi:10.1093/nar/gkz1161.
- [6] W. McCombie, J. McPherson, E. Mardis, Next-generation sequencing technologies, Cold Spring Harbor Perspectives in Medicine 9 (2018) a036798. doi:10.1101/cshperspect.a036798.
- [7] B. Louie, P. Mork, F. Martin-Sanchez, A. Halevy, P. Tarczy-Hornoch, Data integration and genomic medicine, Journal of biomedical informatics 40 (2007) 5–16. doi:10.1016/j.jbi.2006.02.007.
- [8] A. Olivé, Conceptual Modeling of Information Systems, 2007. doi:10.1007/978-3-540-39390-0.
- [9] S. Spreeuwenberg, AIX: Artificial Intelligence needs explanation: Why and how transparency increases the success of AI solutions., Amsterdam: LibRT: the Lab for Intelligent Business Rules Technology, 2019.

- [10] O. Pastor, A. Palacio, J. Reyes Román, J. Casamayor, Modeling Life: A Conceptual Schema-centric Approach to Understand the Genome, 2017, pp. 25–40. doi:10.1007/978-3-319-67271-7\_3.
- [11] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbadó González, S. Garcia, S. Gil-Lopez, D. Molina, V. R. Benjamins, R. Chatila, F. Herrera, Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai, *Information Fusion* 58 (2019). doi:10.1016/j.inffus.2019.12.012.
- [12] J. Qin, M. Žumer, X. Wang, W. Fan, Conceptual models and ontological schemas for semantically sustainable digital libraries, 2020, pp. 441–442. doi:10.1145/3383583.3398545.
- [13] L. Kalinichenko, A. Volnova, E. Gordov, K. Nadezhda, D. Kovaleva, O. Malkov, I. Okladnikov, N. Podkolodnyy, A. Pozanenko, N. Ponomareva, S. Stupnikov, A. Fazliev, Data access challenges for data intensive research in russia 10 (2016) 2–22. doi:10.14357/19922264160101.
- [14] S. Spreeuwenberg, Choose for AI and for Explainability, 2020, pp. 3–8. doi:10.1007/978-3-030-40907-4\_1.