IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 29, 2022, accepted April 11, 2022, date of publication April 21, 2022, date of current version April 28, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3169268

An Advanced Search System to Manage SARS-CoV-2 and COVID-19 Data Using a Model-Driven Development Approach

A. LEÓN, A. GARCÍA SIMON, AND O. PASTOR, (Member, IEEE)
Valencian Research Institute for Artificial Intelligence (VRAIN), Universitat Politécnica de Valéncia, 46022 Valéncia, Spain

Corresponding author: A. León (aleon@vrain.upv.es)

This work was supported in part by the Spanish State Research Agency and the Generalitat Valenciana under the Project PROMETEO/2018/176 and Project INNEST/2021/57, in part by the Spanish Ministry of Universities and the Universitat Politècnica de València under the Margarita Salas Next Generation EU Grant, and in part by European Regional Development Fund (ERDF) and the European Union NextGenerationEU/Plan de Recuperación, Transformación y Resiliencia (PRTR).

ABSTRACT The pandemic outbreak of COVID-19 has allowed the proliferation of an unprecedented amount of data that must be organized and connected in a way that allows its efficient management. Nevertheless, the speed at which all of this knowledge is being generated has highlighted the shortcomings of the research community in creating well-organized, standardized, and structured databases. Despite the efforts of the community to develop advanced integrative platforms such as CovidGraph, we have identified some limitations when using these solutions that we think are derived from the lack of a sound ontological schema to guide the collection, standardization, and integration of data. This work explores the advantages and disadvantages for the final user of building advanced information systems using a Model Driven Development approach to integrate heterogeneous and complex data using an ontological background as a basis. As a proof of concept, we built a database (CovProt) to integrate data about different aspects of SARS-CoV-2 using this approach, we analyzed the advantages and disadvantages of using this approach compared to CovidGraph by performing a set of queries in CovProt and CovidGraph, and finally, we compared the structure and redundancy of the retrieved data.

INDEX TERMS Conceptual model, graph data model, MDD, COVID-19, design methods.

I. INTRODUCTION

The pandemic outbreak of COVID-19, caused by the virus SARS-CoV-2, has allowed the proliferation of an unprecedented number of scientific results about the genetics of the virus and the clinical manifestations of the disease. Several projects, data sources, and consortia have been created to understand the causes of the infection from multiple perspectives (e.g., genetic, clinical, and environmental) and to reduce the devastating consequences that the pandemic is bringing to our population. To be successful, all of the pandemic-related information must be organized and connected so that it can be analyzed both correctly and efficiently. Nevertheless, the speed at which all of this knowledge is being generated has highlighted the shortcomings of the research community in creating well-organized, standardized, and structured databases.

One of the many relevant approaches intended to help researchers explore all of this knowledge from an interconnected perspective is the CovidGraph project (https://covidgraph.org/). CovidGraph offers a set of advanced tools to explore papers, patents, existing treatments, and medications related to the family of the coronaviruses, using a knowledge graph as a basis to represent the fundamental entities of biology (e.g., genes, proteins, and pathways). The information stored in CovidGraph is extracted and integrated from multiple sources such as the COVID-19 Open Research Dataset (CORD-19), Ensembl, Reactome, UniProt, RefSeq, and medRxiv.

CovidGraph uses a knowledge graph that provides a basic structure to the collected data (https://covidgraph.org/). However, we have identified some limitations when exploring the database. The information is stored as it comes from the databases, using the original format and structure. This leads to having redundancies due to the heterogeneity of data formats that are used by the different sources to

The associate editor coordinating the review of this manuscript and approving it for publication was Gianmaria Silvello.

represent the same information (e.g., the same gene or protein is represented by multiple nodes, with different properties and different connections with other nodes). Another consequence is that relevant data or data connections may be missing because of the existing ontologically imprecise characterization (e.g., the genes that are associated with all of the small pathways that constitute a complex biological process instead of being associated only to the pathway where they specifically act).

Despite the limitations mentioned above, storing the data using its original format and structure offers advantages. The main advantage is that integrating the information from different sources is faster since it does not require a data processing task to transform the different source fields into a common data structure. Nevertheless, the process of search, retrieval, and analysis of the collected information becomes more complex because a deep knowledge of how each element is represented in the integrated sources is required.

Our work aims to present how a Model-Driven Development (MDD) can help to mitigate the impact of these problems. To do so, we have developed a conceptually sound and structured database called CovProt. By conceptually sound, we mean that a precise conceptualization process is supporting the conceptual model on which the database is based. Then, we performed a set of queries, in both CovProt and CovidGraph, in order to compare the structure of the retrieved data and the complexity of the queries required to obtain the results.

This work is not intended to present the technological details of a new search engine, but to explore the advantages and disadvantages for the final user of building advanced information systems using a MDD approach to integrate heterogeneous and complex data using an ontological background as a basis. Following this reasoning, the work intends to go beyond mappings between databases to make queries easier. We try to explain how the use of models improves the development process and has a real impact on building information systems that can be more intuitive for the user to explore and free of redundancies.

The remainder of our work is structured as follows. After the introduction, Section 2 describes what MDD is, and Section 3 describes the materials and methods used to achieve the objective of this work. Section 4 describes the Conceptual Model (CM), and Section 5 shows the Graph Data Model (GDM) used to implement the CovProt database. In Section 6, we describe the population of the database. The CovidGraph and CovProt databases are compared in Section 7. Finally, our results are presented in Section 8, and the conclusions are presented in Section 9.

II. MODEL-DRIVEN DEVELOPMENT (MDD)

Model-Driven Development (MDD) consists of building complex systems from models that are smaller and more abstract representations of the different parts of the system [1]. One of the main advantages of developing complex systems using an MDD approach is that they are built using concepts that are independent of the implementation technology and are closer to the problem. Having technological-independent concepts that focus on the problem rather than on the solution makes models easier to specify, understand, and maintain [2]. Furthermore, they are adaptive and easily expandible, which is a key characteristic for domains where the knowledge evolves quickly. Examples of use in the medical field are [3] and [4].

Different types of models can be used in the different stages of the software lifecycle. In this work, we focus on the models that are used to develop a database to store and query data about SARS-CoV-2: the Conceptual Model (CM) and the Graph Data Model (GDM) or Property Graph (PG).

CMs are Platform-Independent Models (PIMs) that are commonly supported by well-founded, precise, and accurate ontologies. A CM can be translated into a Platform Specific Model (PSM) to design graph databases or into a PSM to design relational databases. An example of a translation process that is commonly used to design relational databases is the Entity-Relationship model [5]. Like the Entity-Relationship model in relational databases, a GDM is a PSM that is used to design graph databases, where data structures are represented as graphs or generalizations of them. Despite the ‘‘schema-less’’ philosophy behind graph databases, the definition of a conceptually well-grounded data model is highly encouraged in order to ensure that the data correctly represents the domain [6], [7].

III. MATERIALS AND METHODS

Due to the complexity of the data to be represented and to facilitate the understanding of the process, we have focused on the fundamental elements of the genetic perspective of COVID-19 that corresponds to the BioMedical view of the CovidGraph knowledge graph shown in Fig. 1. Thus, the information to be integrated is related to the host genes and proteins that interact with the virus facilitating its entrance into the cells, the pathways where they are involved, and the possible variants and diseases caused by their alteration.

First, we have defined a PIM model. This model describes the information to be integrated, making the relationships among the concepts explicit, and facilitating the understanding of the domain. Second, the model is transformed to a PSM. The PSM is a GDM since the technology used to store and query the data is a Neo4j database. Third, the data is collected and stored in a database instance following the structure defined by the GDM. Fig. 1 shows the structure of the MDD components followed in this work.

The data is retrieved from eight data sources: NCBI Gene [8], UniProtKB [9], NCBI Taxonomy [10], ClinVar [11], Reactome [12], PubMed [10], Human Phenotype Ontology [13], and GeneHancer [14]. This data is then stored in a Neo4j database.

IV. CONCEPTUAL MODEL OF THE CovProt DATABASE

The first stage of the MDD approach used in this work consists of defining the conceptual model representing

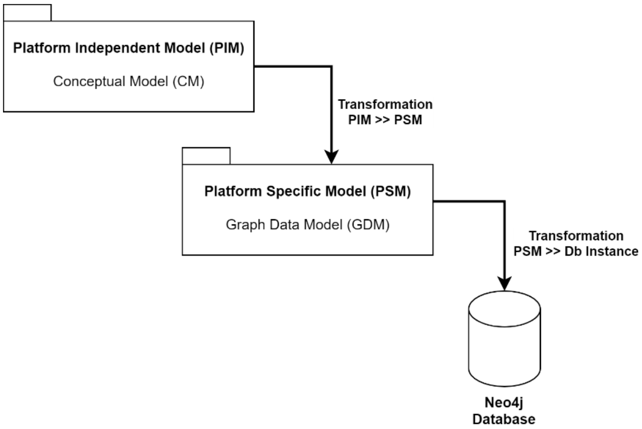


FIGURE 1. Components of the MDD approach used to develop the CovProt database.

the fundamental concepts that are associated with the information to be stored. Therefore, the following knowledge needs to be represented:

- Details regarding the host (Homo Sapiens) and the virus (SARS-COV-2).
- The viral proteins that interact with the host proteins or complexes and vice versa.
- The genes that codify the proteins.
- The regulatory elements that control the transcription and expression of the genes.
- The pathways where the host proteins are involved.
- The diseases where the host proteins play a significant role.
- The variants that may cause a disease.
- The evidence that supports the role of the variant in the development of disease.
- Cross-references to the data sources.

The resulting CM, which is described using the UML Class Diagram modeling language (https://www.uml.org/index.htm), is shown in Fig. 2.

This model represents the basic concepts as classes. The relationships among classes are represented as associations. The minimum properties required to characterize the concepts are represented as class attributes. The model can be extended with additional concepts and properties to represent as much knowledge as required.

V. GRAPH DATA MODEL OF THE CovProt DATABASE

Since the analysis done in this work focuses on identifying relationships among the data, the technology selected to support CovProt is a graph database. Graph databases excel at easing the analysis of highly connected data. Thereby, reducing the cost of doing multiple joins when the depth of the connections is high. We have used Neo4j (https://neo4j.com/), which is a widely used graph database that provides mechanisms to ensure data integrity, scalability, and ACID compliance (Atomicity, Consistency, Isolation, and Durability). In this section, the GDM is derived from the previously defined CM and is adapted to the requirements of a graph database, resulting in the model shown in Fig. 3.

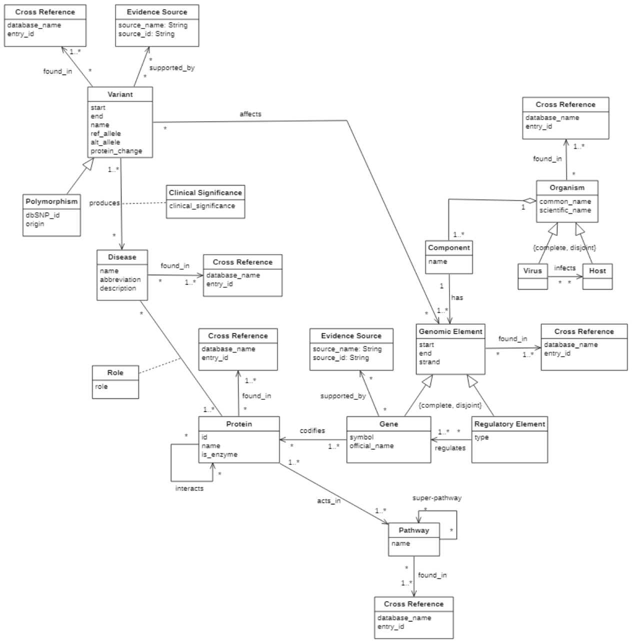


FIGURE 2. CovProt conceptual model.

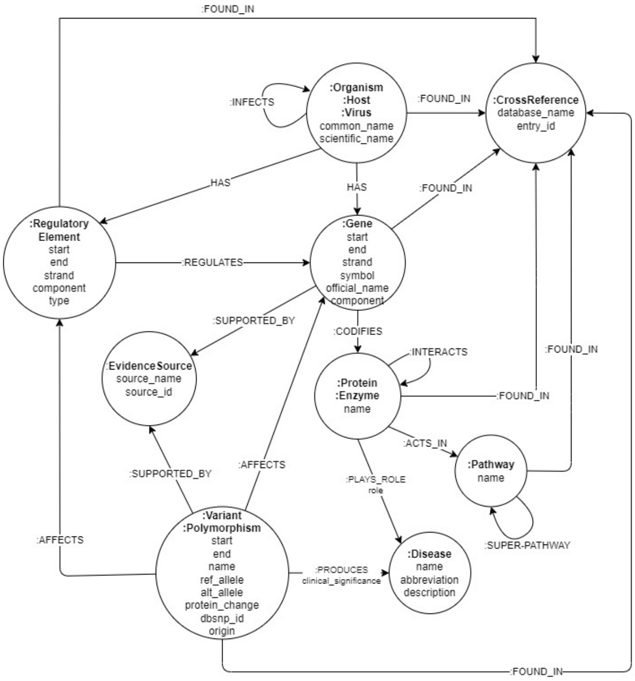


FIGURE 3. CovProt graph data model.

Since the CM used in this work is not a complex model in terms of number of classes, attributes, and relationships, the CM-to-GDM transformation has been done manually. Examples of the different approaches for mapping CMs to GMDs can be found in [15], [16], and [17]. This GDM represents how the data is structured in the instantiation of

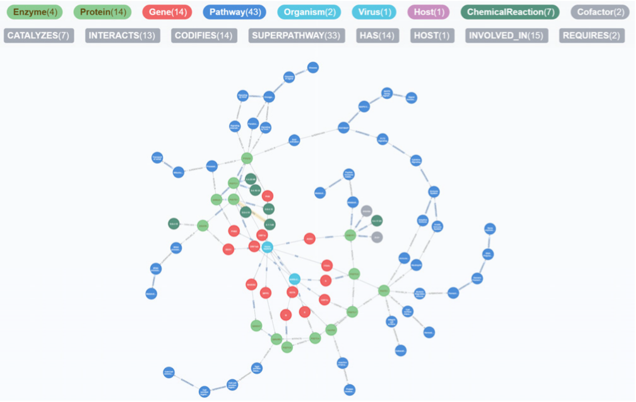


FIGURE 4. Results of the interactions between the virus and the host proteins, along with the pathways where they are involved.

the database. Any change or improvement in the CM can be translated into the GDM easily, adapting the structure of the underlying database as the model evolves.

VI. POPULATING THE CovProt DATABASE

Once the GDM has been defined, the next step is to populate the database. To do this, we used the following data sources: NCBI Gene, UniProtKB, NCBI Taxonomy, ClinVar, Reactome, PubMed, Human Phenotype Ontology (HPO), GeneHancer. The information extracted from the data sources is collected, integrated, and stored in the CovProt database according to the structure defined in the GDM.

To do so, one of the steps to be performed in any MDD approach is the definition of the mapping rules that allow the representation of the raw data into the new data schema. The mapping rules are defined using as a strict basis the conceptual model of the domain and having in mind the specific structure of each data source. These rules determine how the source fields are mapped to the corresponding concept of the conceptual schema and translated into the corresponding field in the destination database. For example, considering the GDM represented in Fig. 4, the mapping rules required to populate a Gene node must define the correspondence between each of its attributes (start, end, strand, symbol, official_name, and component) and the data source attribute that provides the required data. A mapping rule also specifies if the original data requires transformation to adapt to the destination format (e.g., adding a certain prefix or parsing a string). In our development cycle, the application of the mapping rules has been automated to be included during the extraction and transformation process.

The use of a conceptual schema of the domain along with the definition of the mapping rules are key to solve well-known integration problems when the data sources have different schemas. Using a conceptual modeling-based approach can provide semantic interoperability through a conceptual modeling characterization that delimit concepts even when their database representation is different in different data sources. Nevertheless, defining the mapping rules is not an easy task and requires a deep knowledge of the underlying structure of each data source. Furthermore, if the data source or the GDM structure changes, the mapping rules must change too. In order to keep the focus of this document in the MDD approach, the technological details about how the mapping rules are defined and implemented has been omitted.

Once the database is populated, it can be queried to determine that six of the 16 proteins that conform the virus (ORF1ab, ORF1a, ORF7a, E, N, and S) interact with eight host proteins (PHB, PHB2, DDX1, ITGGAL, SGTA, MPP5, SH2D3C, and ACE2). Furthermore, the host proteins are involved in 15 pathways that can be grouped into the following categories: Transport of small molecules, Metabolism of RNA, Immune System, Gene expression (Transcription), Hemostasis, Extracellular matrix organization, Protein localization, Cell-cell communication, Disease (Oncogenic MAPK signaling), and Metabolism of proteins. Fig. 4 shows how this data is interconnected in the CovProt database.

A total of 258 host proteins interacts with the eight used by the virus to infect the cells. The host proteins used by the virus to infect the cells interact with the other 258 proteins.

Following the GDM defined in Section 5, the user can also explore the variants that affect the genes and the regulatory elements that codify the proteins as well as the diseases related to these variants.

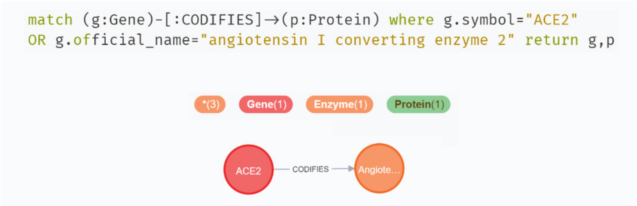
VII. COMPARING CovProt AND CovidGraph

To compare the two approaches (CovProt and CovidGraph), we have executed a set of queries that are of interest to the experts that work with COVID data, and we have analyzed the structure of the results, the level of redundancy, and the complexity of the queries required to obtain the desired results.

When exploring the genetic implications of the response to SARS-COV2 infection, the most important queries to be answered are related to:

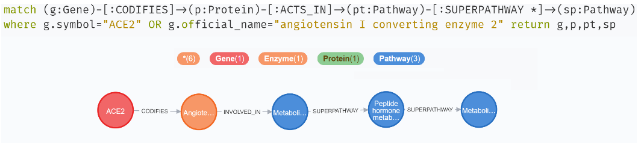
1. Retrieving information about the host genes that interact with the virus to understand how the virus enters the cells. In this example, we have simplified the query to retrieve information about one of the most well-known genes (ACE2).
2. Retrieving information about the biological pathway where the genes are involved to identify the altered mechanisms that can lead to cellular malfunction.
3. Identifying all the biological pathways that could be altered based on all the host proteins that interact with the virus to have an idea of the clinical manifestations of the disease.

These three queries help to understand key points of the infection such as how the virus can attack the cells, which biological mechanisms are altered, and how these alterations produce the characteristic symptoms of the disease. This is the reason why we have selected these three queries for the example.



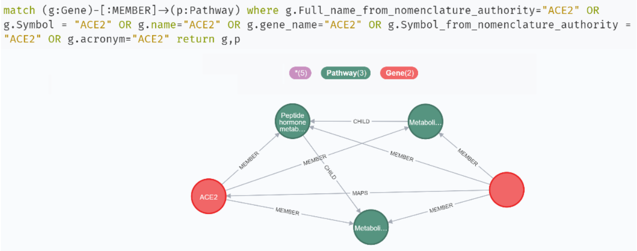
```
match (g:Gene)-[:CODIFIES]→(p:Protein) where g.symbol="ACE2"
OR g.official_name="angiotensin I converting enzyme 2" return g,p
```

FIGURE 5. Cypher query and results for the ACE2 gene and the codified protein in the CovProt database.



```
match (g:Gene)-[:CODIFIES]→(p:Protein)-[:ACTS_IN]→(pt:Pathway)-[:SUPERPATHWAY *]→(sp:Pathway)
where g.symbol="ACE2" OR g.official_name="angiotensin I converting enzyme 2" return g,p,pt,sp
```

FIGURE 6. Cypher query and results for the pathways where the ACE2 gene is involved in the CovProt database.



```
match (g:Gene)-[:MEMBER]→(p:Pathway) where g.Full_name_from_nomenclature_authority="ACE2" OR
g.Symbol = "ACE2" OR g.name="ACE2" OR g.gene_name="ACE2" OR g.Symbol_from_nomenclature_authority =
"ACE2" OR g.acronym="ACE2" return g,p
```

FIGURE 7. Cypher query and results for the pathways where the ACE2 gene is involved in the CovidGraph database.

A. OBTAINING DATA ABOUT THE ACE2 GENE

The first query is intended to obtain the data associated with the *angiotensin I converting enzyme 2* gene (also known as ACE2) and the protein it codifies. In the CovProt database, the query should match the path gene – codifies – protein, filtering by the gene symbol (ACE2) or by the gene name (*angiotensin I converting enzyme 2*). As a result of the query, two nodes are returned (one for the gene and one for the codified protein). All of the associated information is stored in the corresponding node as properties. The result of the query is shown in Fig.5.

To retrieve the same data in CovidGraph, the structure of the origin databases from where the information has been collected must be considered. In this case, the query should follow the path *gene – codes – transcript – codes – protein*. In CovidGraph, the genes have multiple properties to represent symbols and names such as:

- *Full_name_from_nomenclature_authority*
- *Other_designations*
- *Symbol*
- *Symbol_from_nomenclature_authority*
- *Synonyms*
- *Acronym*
- *Gene_name*
- *Synonyms*
- *Name.*

As can be observed, there are different attributes that seem to represent the same concept (e.g., Synonyms and synonyms). Furthermore, since the information is not structured and the concepts are not ontologically well-grounded, the different databases do not assign the same meaning to symbol and name, which means that these terms are usually mixed, leading to confusion. Therefore, the query to get the desired results must consider all these options:

MATCH (g:Gene)-[:CODES]-(t:Transcript)-[:CODES]-(p:Protein)
WHERE
g.Full_name_from_nomenclature_authority = "ACE2" OR
g.symbol = "ACE2" OR g.name = "ACE2" OR
g.gene_name = "ACE2" OR
g.Symbol_from_nomenclature_authority = "ACE2" OR
g.acronym = "ACE2"
RETURN g, t, p

As a result, the database returns two gene nodes, four transcript nodes, and 10 protein nodes. Both gene nodes represent the same gene, with different attributes and with attributes representing inconsistent content. For example, ACE2 is represented as a symbol in one node and as a name in another node.

Similarly, each gene codifies two transcripts that are duplicated in the database. The 10 protein nodes correspond to redundant data about the two isoforms of the same protein. Transcripts and proteins also have a different number of attributes. This means that in order to obtain global knowledge about the protein, the data from all of the nodes must be retrieved and combined, which involves removing duplicates and fixing inconsistencies.

B. OBTAINING DATA ABOUT PATHWAYS

The second query is intended to obtain information about the pathways where the ACE2 gene is involved. In the CovProt database, this information can be obtained following the path gene – codifies – protein – acts_in – pathway. The result of the query is shown in Fig. 6.

The ACE2 gene is involved in one pathway (*the Metabolism of Angiotensinogen to Angiotensins*), which is part of the *Peptide hormone metabolism pathway*, which is part of the *Metabolism of proteins*.

In the CovidGraph database, the same information must be retrieved following the path gene – member – pathway. Using the first query to find the ACE2 gene, the results returned are shown in Fig. 7.

According to the data stored in the CovidGraph database, the results returned are the same since the gene is duplicated. Nevertheless, the difference with the CovProt results is that the gene acts as a member in all of the pathways (related to each other by the CHILD association). This means that it is not possible to know exactly in which specific part of a complex pathway the gene takes part.

C. OBTAINING THE PATHWAYS WHERE THE HOST PROTEINS ARE INVOLVED

The previous queries were intended to solve very specific questions about a gene or a protein; however, a real research context requires answering more general questions. To represent such contexts, we are going to query all of the pathways where host proteins that are used by the virus to infect the cell are involved. The aim of this query is to get a more complete view of the effect of the infection. As a starting point, we are going to consider the eight proteins mentioned in Section 6 (PHB, PHB2, DDX1, ITGGAL, SGTA, MPP5, SH2D3C, and ACE2). In the CovProt database, the path to follow is the same used for Query 2, but filtering by the names of the eight genes.

Seven of the eight proteins are associated with pathways in the database, and the PHB and PHB2 proteins are involved in the same pathway (Processing of SMDT1). The same query in the CovidGraph database requires extra effort to filter the genes in order to ensure that no relevant information is missing. Even though the results are the same, the query is much more complex, and the number of returned nodes and connections is higher.

VIII. RESULTS

The structure of the data returned by CovidGraph and CovProt is easy to understand because both approaches are based on a graph data model that helps connect the different concepts in a way similar to how experts understand the connections between the main concepts of the domain. However, the lack of an ontological foundation to integrate the results produces that the datasets obtained from CovidGraph have redundancies that increase the number of nodes returned and the complexity of the connections among these nodes. For example, while CovProt returns one node representing the protein codified by the ACE2 gene, CovidGraph returns 10, and they have a different number of attributes. In addition, the high degree of heterogeneity of the nodes of CovidGraph increases the complexity of some queries (see Subsection 7.C), and it requires a thorough knowledge of the underlying schema of each integrated data source in order to obtain the full benefit of the stored data.

The strength of developing an information system with an ontological background supporting it, is that the data are stored and structured in a way that is conceptually consistent. Since there is no need to deeply understand the internal structure of each data source to query the data, the user only requires his knowledge of the domain to navigate through the data structure, and building queries are more natural for the user. Furthermore, the frequent inconsistencies that can appear in the vocabulary that is used to define the concepts, are solved by the underlying ontology. This means that the information is easier to retrieve.

IX. CONCLUSION

Our concrete intention in this work is to point out the advantages and disadvantages of an MDD approach to manage such complex data as genomics is, serving as a guide for improving data scientists and developers' work when designing and developing information systems. To such aim, we have used an MDD approach to build the CovProt database, and we have performed a set of queries to determine its advantages and disadvantages compared to CovidGraph, which integrates the data without the support of any conceptual schema. CovProt integrates information about SARS-COV-2 and the context of the host proteins that interact with the virus in an ontologically well-grounded and structured way. The CovProt database presented in this document is a proof of concept that is currently under development to be publicly available as future work, along with a performance benchmark to describe its advantages and disadvantages in terms of speed, performance, and size.

The use of a CM that is independent of the technological implementation of the database allows the correct conceptualization and representation of the information to be stored. CovidGraph also integrates information based on a knowledge graph that connects the fundamental entities of the biological domain. However, without a sound ontology as a basis, redundancies and inconsistencies in its data arise, which hinders the data analysis process.

One of the advantages of using an MDD approach such as CovProt is that it results in a database without redundancies, in which all of the information is well-organized with a structure that is intuitive to navigate and query. In contrast, CovidGraph provides a navigational structure that is apparently easier to query but building accurate queries and retrieving the desired information is more complex. The reason for this is that the data has been stored as obtained from the original sources, and its lack of standardization and complexity requires a deep knowledge of the structure of each source. This lack of standardization increases the probability of missing relevant data due to executing a wrong or incomplete query. In addition, the higher number of connections and nodes makes it difficult to understand and analyze the results efficiently.

One of the disadvantages of using an MDD approach is that it requires an extra effort to define the CM that represents the domain in order to transform the model to a GDM. Another difficulty to be considered is derived from the already known complexity of integrating data coming from heterogeneous sources that commonly have different structure (e.g., format disparities, variable level of quality, and duplicates). These challenges can be addressed by defining the mapping and transformation rules that are required to represent the data into a common format. Nevertheless, as has been mentioned in section VI, the definition of these rules is not a trivial task.

The storage of the data as it comes from the sources is faster and easier thanks to the existence of ‘‘schema-less'' databases like Neo4j. Taking advantage of this more flexible approach reduces the time and effort required to integrate new data sources into the system. Another disadvantage is that if any external source changes its data structure, the transformation rules must be reevaluated.

Both approaches have advantages and disadvantages that must be carefully considered when building advanced search systems that require the integration of multiple sources. Important considerations that must be taken into account include the following: the heterogeneity of the original data source schemes from where the information is collected; how it can complicate the creation of nodes, attributes, and associations; and how it can affect the building of queries to obtain the desired information. It is also necessary to consider the impact of redundant data on both the performance when executing queries and on the complexity of the analysis of the results obtained. Finally, if many heterogeneous sources are going to be integrated, the effort required to define and implement the mapping and transformation rules of an MDD approach must be carefully analyzed.

REFERENCES

[1] O. Pastor, S. España, J. I. Panach, and N. Aquino, "Model-driven development," *Informatik-Spektrum*, vol. 31, no. 5, pp. 394–407, Oct. 2008, doi: 10.1007/s00287-008-0275-8.

[2] B. Selic, "The pragmatics of model-driven development," *IEEE Softw.*, vol. 20, no. 5, pp. 19–25, Sep. 2003, doi: 10.1109/MS.2003.1231146.

[3] J. A. Maldonado, M. Marcos, J. T. Fernández-Breis, V. M. Giménez-Solano, M. D. C. Legaz-García, and B. Martínez-Salvador, "CLIN-IK-LINKS: A platform for the design and execution of clinical data transformation and reasoning workflows," *Comput. Methods Programs Biomed.*, vol. 197, Dec. 2020, Art. no. 105616.

[4] S. K. Shahzad, D. Ahmed, M. R. Naqvi, M. T. Mushtaq, M. W. Iqbal, and F. Munir, "Ontology driven smart health service integration," *Comput. Methods Programs Biomed.*, vol. 207, Aug. 2021, Art. no. 106146.

[5] V. C. Storey, "Relational database design based on the entity-relationship model," *Data Knowl. Eng.*, vol. 7, no. 1, pp. 47–83, Nov. 1991, doi: 10.1016/0169-023X(91)90033-T.

[6] R. Angles, "A comparison of current graph database models," in *Proc. IEEE 28th Int. Conf. Data Eng. Workshops*, Apr. 2012, pp. 171–177, doi: 10.1109/ICDEW.2012.31.

[7] R. Angles, "The property graph database model," in *Proc. CEUR Workshop*, vol. 2100, 2018, pp. 1–10.

[8] *Gene*, National Center for Biotechnology Information, National Library of Medicine, Bethesda, MD, USA, 2004.

[9] The UniProt Consortium, "UniProt: The universal protein knowledgebase in 2021," *Nucleic Acids Res.*, vol. 49, no. D1, pp. D480–D489, Jan. 2021, doi: 10.1093/nar/gkaa1100.

[10] E. W. Sayers, T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, R. Edgar, S. Federhen, and M. Feolo, "Database resources of the national center for biotechnology information," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D23–D28, Jan. 2019, doi: 10.1093/nar/gky1069.

[11] M. J. Landrum, J. M. Lee, M. Benson, G. R. Brown, C. Chao, S. Chitipiralla, B. Gu, J. Hart, D. Hoffman, W. Jang, and K. Karapetyan, "ClinVar: Improving access to variant interpretations and supporting evidence," *Nucleic Acids Res.*, vol. 46, no. D1, pp. D1062–D1067, Jan. 2018, doi: 10.1093/nar/gkx1153.

[12] B. Jassal, S. Jupe, L. Matthews, K. Sidiropoulos, M. Gillespie, P. Garapati, R. Haw, B. Jassal, F. Korninger, B. May, and M. Milacic, "The reactome pathway knowledgebase," *Nucleic Acids Res.*, vol. 46, pp. D649–D655, Nov. 2019, doi: 10.1093/nar/gkz1031.

[13] S. Köhler, M. Gargano, N. Matentzoglu, L. C. Carmody, D. Lewis-Smith, N. A. Vasilevsky, D. Danis, G. Balagura, G. Baynam, A. M. Brower, and T. J. Callahan, "The human phenotype ontology in 2021," *Nucleic Acids Res.*, vol. 49, no. D1, pp. D1207–D1217, Jan. 2021, doi: 10.1093/nar/gkaa1043.

[14] S. Fishilevich, R. Nudel, N. Rappaport, R. Hadar, I. Plaschkes, T. I. Stein, N. Rosen, A. Kohn, M. Twik, M. Safran, D. Lancet, and D. Cohen, "GeneHancer: Genome-wide integration of enhancers and target genes in GeneCards," *Database*, vol. 2017, p. 17, Jan. 2017, doi: 10.1093/database/bax028.

[15] D. Gweldand, S. Gerso, and J. Cabot, "UMLtoGraphDB: Mapping conceptual schemas to graph databases," in *Proc. Int. Conf. Conceptual Modeling (ER)*, 2016, pp. 430–444, doi: 10.1007/978-3-319-46397-1_33.

[16] K. Shin, C. Hwang, and H. Jung, "NoSQL database design using UML conceptual data model based on Peter Chen's framework," *Int. J. Appl. Eng. Res.*, vol. 12, no. 5, pp. 632–636, 2017.

[17] P. Atzeni, F. Bugiotti, L. Cabibbo, and R. Torlone, "Data modeling in the NoSQL world," *Comput. Standards Interface*, vol. 67, Jan. 2020, Art. no. 103149, doi: 10.1016/j.csi.2016.10.003.

A. LEÓN received the Ph.D. degree in computer science from the Universitat Politècnica de València, in 2019. Currently, she is a Researcher with the Research Center on Software Production Methods (PROS-UPV), Universitat Politècnica de València, where her research activity is focused on the use of conceptual models for the development of genomic information systems and the definition of a systematic process for the search, identification, load, and exploitation of DNA variants in the context of precision medicine. She is also an University Expert in medical genetics and genomics with the Universidad Católica de Murcia, Spain. Her research interests include conceptual modeling, genomic data science, data quality, and information systems.

A. GARCÍA SIMON is currently pursuing the Ph.D. degree with the PROS Research Center, Universitat Politècnica de València, Spain. He is also an in charge of the development of a technological platform to provide support for the identification of relevant variants and their application to precision medicine. His research is focused on the application of conceptual modeling to describe genetic data independently of the species. His research interests include conceptual modeling, genomic data science, and information systems.

O. PASTOR (Member, IEEE) is currently a Full Professor and the Director of the PROS Research Center, Universitat Politècnica de València, Spain. He is currently leading a multidisciplinary project linking information systems and bioinformatics to designing and implementing tools for conceptual modeling-based interpretation of the human genome information. He has published more than 300 research papers in conference proceedings, journals, and books, received numerous research grants from public institutions and private industry, and a keynote speaker at several conferences and workshops. His research interests include conceptual modeling, web engineering, requirements engineering, information systems, and model-based software production.