DATOS-CAT: Methodologies for the standardization, integration and analysis of population-based biomedical data using semantic technologies

Judith Martinez-Gonzalez[1,2,*,†], Guillem Bracons Cucó[1,†], Aikaterini Lymperidou[1,3], Xavier Escribà-Montagut[4], Marta Huertas[1,5], Ramon Mateo-Navarro[1,4], David Sarrat-González[4], Santiago Frid[6], Rafael de Cid[3], Juan R González[4] and Alberto Labarga[2]

[1]*Institute for Bioengineering of Catalonia, Barcelona, Spain*
[2]*Barcelona Supercomputing Center, Barcelona, Spain*
[3]*Genomes for Life- GCAT lab- Germans Trias i Pujol Research Institute, Badalona, Spain*
[4]*Barcelona Institute for Global Health, Barcelona, Spain*
[5]*Centre for Genomic Regulation, Barcelona, Spain*
[6]*Hospital Clínic de Barcelona, Barcelona, Spain*

1. Introduction

In the context of personalized medicine, long-term data collection allows researchers to track how diseases progress over time, identify patterns of environmental and genetic risk, and assess the impact of different treatment strategies. The project DATOS-CAT aims to increase the visibility and scientific impact of the population-based cohorts developed in Catalonia, GCAT|Genomes for life and the COVICAT-CONTENT sub-cohort, and to contribute to the development of procedures applicable to other cohorts, improving the level of interoperability of their data in the context of the FAIR data ecosystem principles (Findable, Accessible, Interoperable, Reusables) to facilitate their exploitation and scientific use. As a basis for development, the European Genome-Phenome Archive (EGA) infrastructure will be used for the genomic data, and the Observational Medical Outcomes Partnership Common Data Model (OMOP-CDM) compatible standards will be used for working with structured clinical data.

SWAT4HCLS '24: The 15th International SWAT4HCLS conference, February 26-29, 2024, Leiden, The Netherlands.
*Corresponding author.
†These authors contributed equally.
judith.martinez@bsc.es (J. Martinez-Gonzalez); bracons@recerca.clinic.cat (G. Bracons Cucó); alymperidou@igtp.cat (A. Lymperidou); xavier.escriba@isglobal.org (X. Escribà-Montagut); marta.huertas@crg.eu (M. Huertas); ramon.mateo@isglobal.org (R. Mateo-Navarro); david.sarrat@isglobal.org (D. Sarrat-González); frid@clinic.cat (S. Frid); rdecid@igtp.cat (R. d. Cid); juanr.gonzalez@isglobal.org (J. R. González); alabarga@bsc.es (A. Labarga)
0009-0009-9441-5971 (J. Martinez-Gonzalez); 0000-0003-1274-7403 (G. Bracons Cucó); 0009-0002-3425-2188 (A. Lymperidou); 0000-0003-2888-8948 (X. Escribà-Montagut); 0009-0000-9334-108X (R. Mateo-Navarro); 0000-0002-9064-3303 (D. Sarrat-González); 0000-0001-8400-5770 (S. Frid); 0000-0003-3579-6777 (R. d. Cid); 0000-0003-3267-2146 (J. R. González); 0000-0001-6781-893X (A. Labarga)

2. Cohorts

The GCAT project has been following 20.000 individuals living in Catalunya since 2014. Extensive clinical, genomic and lifestyle data has been collected, set up to integrate and assess the role of the environmental and genetic factors (i.e, genomic, metabolomic, proteomic, epigenomic) in the development of chronic diseases. The project aims to assess the prevalence of risk factors and their association with disease incidence over time. The COVICAT-CONTENT is a sub-cohort involving 11.833 participants, mostly derived from the matrix GCAT cohort. Created in 2020 with the specific aim of monitoring COVID-19, the final objective was to evaluate determinants, ultimately contributing to the definition of predictive models and control policies for similar situations in the future.

3. Methods

The OMOP-CDM is intended to be used as a standardized framework for data representation and analysis. The OMOP-CDM is a person-centric model, which establishes a common vocabulary and data format, allowing for seamless integration across disparate heterogeneous sources.

Building an OMOP-CDM database requires Extract-Transform-Load (ETL) processes that facilitate the data conversion to a standard model and terminology. ETL plays an important role in ensuring data consistency and integrity through the integration process, enabling a smooth transition from heterogeneous datasets to a coherent and standardized structure. DATOS-CAT project evaluates two different ETL methodologies: a traditional ETL and a semantic ETL.

On one hand, the traditional ETL framework involves sequential steps. First, the Extract phase allows us to gather raw data from multiple sources, such as EHR. General information from the database can be extracted using the White Rabbit software. Once extracted, the Transform phase allows the standardization and harmonization of diverse data formats, terminologies and structures into the OMOP-CDM, where the data is restructured in tables and mapped to Standardized Vocabularies. In this second phase, Rabbit-in-a-Hat software can be useful for mapping variables to the corresponding tables and columns of the OMOP-CDM. Based on the resulting mapping, a specific SQL transformation process can be implemented to transform raw data into the desired format. This last point usually involves mapping source-specific codes (e.g., ICD-10, RxNorm, etc.) to OMOP standard terminologies. After the transformation, the Load phase allows to populate the unified database to a OMOP-CDM compliant format. Each of the steps is crucial for effectively converting heterogeneous raw data into a standardized OMOP format.

On the other hand, having standardized data with their semantic meaning and syntactic structure is key for conducting scalable and interoperable secondary data studies. To achieve such data normalization, the Ontobridge tool has been developed. Ontobridge is an ontology-based tool, created by the Hospital Clínic de Barcelona, that transforms local databases to CDMs. It uses semantic mappings of data between an ontology representing the local data model, and one or more ontologies, representing data standards to transform local databases into standards (OMOP-CDM on our use case), in a secure, fast, scalable manner, that maintains the semantics and syntactic structure of the data. Specifically, Ontobridge uses Ontop to replicate relational

data into RDFs through R2RML mappings, and then inserts the triples into an ontology that represent the local data model. Local concepts are mapped to standard ones by means of the owl:sameAs property to allow for semantic interoperability, while the syntactic equivalence is performed by defining that local properties to be instances of metaclasses that model attributes of the common data models. These ontologies are loaded into a Jena Fuseki server, and the corresponding SPARQL queries are executed to generate the OMOP-CDM tables.

On this project, we examine the adaptability, efficiency and scalability of each ETL approach in achieving data consistency and integrity.

4. Results

The standardization of both cohorts is indispensable, as it offers more robust pathways for collaborative researchers and facilitates their integration to global genomic and clinical data repositories. During this process, ETL plays a crucial role since it enables the transformation of raw heterogeneous data from GCAT and COVICAT-CONTENT into standardized and consistent formats. At the end, the integration empowers researchers to access direct sources of information, enabling a deeper understanding of how genetic profiles relate to clinical outcomes, unlocking potential breakthrough discoveries and providing new knowledge that can significantly influence health care practices around the world. DATOS-CAT seeks to contribute to making these cohorts more competitive at a global level, by connecting them with direct sources of genomic and clinical information, ultimately resulting in a greater benefit for society. The proposal also contains the elements to align and collaborate with the efforts carried out by the IMPaCT programme and also with other European parallel developments.