Judith Martinez Gonzalez

The DATOS-CAT project: Leveraging OMOP CDM for the standardization, integration and analysis of population-based biomedical data in Catalonia

MASTER'S THESIS

Supervised by Alberto Labarga Gutiérrez

Master's Degree in Biomedical Data Science

Tarragona, 2024

Mr. ALBERTO LABARGA GUTIERREZ, certifies that the student JUDITH MARTINEZ GONZALEZ has elaborated the work under his direction, and he authorizes the presentation of this Master's Thesis for its evaluation.

Advisor signature:

Acknowledgments

I would like to extend my extreme gratitude and heartfelt appreciation to everyone who has supported, contributed and guided me through the completion of this Master's thesis.

First and foremost, I would like to thank my thesis advisor, Alberto Labarga, for his invaluable patience, insightful advice, and constant support. His experience and dedication have been fundamental to the development of this work.

I also extend my gratitude to all my colleagues at the INB group of the BSC and the collaborators of DATOS-CAT, for their knowledge, teaching and for the journeys we have shared. Their passion for the subject has significantly contributed to my professional development and has been a constant source of inspiration.

To my fellow master's colleagues, Eric Torralba and Carlos Colmenero, for the shared moments and mutual support throughout this journey. Your friendship has made this experience truly special.

I wish to express my extreme gratitude to Arnau, Maria Rosa, Antonio, Andrea, my family and friends, for their unconditional love, understanding and support. Their constant encouragement has been a fundamental pillar throughout this process.

Finally, I would like to extend a special thanks to *Plan Complementario de Biotecnología aplicada a la Salud*, whose funding has been essential to the completion of this work. Their financial support has enabled this project to take place and has been crucial to its success.

Thank you very much,

Este proyecto ha sido financiado por el "Plan Complementario de Biotecnología aplicada a la Salud", coordinado por el Institut de Bioenginyeria de Catalunya (IBEC) en el marco del Plan de Recuperación, Transformación y Resiliencia (C17.I1) - Financiado por la Unión Europea - NextGenerationEU.

Authorship Attribution

I, Judith Martinez Gonzalez, solemnly declare that all the content present in this Master's Thesis is the result of my own research, analysis and effort, except for the parts indicated. I have dedicated time and resources to data cataloging, data standardization, data analysis, as well as to the literature review and individual writing of this document.

However, I must highlight that this project is part of a collaboration among seven institutions, as mentioned throughout this work. In sections where collaborative work has been included, I have properly attributed the contribution of each institution involved.

Furthermore, I have referenced all external sources used to support my arguments or ideas, adhering to principles of academic integrity and avoiding plagiarism in all its forms.

Therefore, I confirm that this Master's Thesis represents my individual and original work, except for the parts where collaboration with other entities has been acknowledged.

Abstract

Personalized medicine is based on the use of individual data to develop specific and effective treatments. This study focuses on contributing to the DATOS-CAT project through the development of guidelines, tools, and protocols to improve the interoperability and visibility of the Catalan population cohort GCAT and the COVICAT-CONTENT sub-cohort, ensuring their alignment with the FAIR principles and the policies of the European Commission.

The main objective of DATOS-CAT is to develop the methodology and infrastructure necessary for the publication of the data generated in the context of the GCAT/COVICAT projects for use by the scientific community, following FAIR standards and ensuring the privacy and security of patients' personal information.

In this master's thesis project, the data flow is defined, and different technologies to be used are evaluated. Key aspects, such as creating a comprehensive data catalog and standardizing data to a common data model, are addressed. For this standardization, we evaluate two approaches for the Extract-Load-Transform (ELT) processes, one based on data engineering tools commonly used in the tech industry (Meltano, DBT, etc.) and another based on a tool developed within the project framework using semantic technologies (OntoBridge).

We emphasize the importance of developing a platform for the discovery and federated analysis of data, maintaining their privacy and security, demonstrating the usefulness of open-source tools such as OBiBa, DataSHIELD, and Beacon v2.

This study provides a roadmap for future research and improvements within health projects, thus contributing to the continuous improvement of data management and analysis.

Keywords: DATOS-CAT, FAIR principles, OMOP CDM, interoperability, data standardization, data discovery, federated data analysis, DataSHIELD, Beacon.

Resum

La medicina personalitzada es basa en la utilització de dades individuals per desenvolupar tractaments específics i efectius. Aquest estudi se centra a contribuir al projecte DATOS-CAT mitjançant el desenvolupament de directrius, eines i protocols per tal de millorar la interoperabilitat i la visibilitat de la cohort poblacional catalana GCAT i de la sub-cohort COVICAT-CONTENT, assegurant la seva alineació amb els principis FAIR i les polítiques de la Comissió Europea.

L'objectiu principal de DATOS-CAT és el desenvolupament de la metodologia i la infraestructura necessàries per a la publicació de les dades generades en el context del projecte GCAT/COVICAT per al seu ús per part de la comunitat científica, seguint estàndards FAIR i assegurant la privacitat i seguretat de la informació personal dels pacients.

En aquest projecte de final de màster, es defineix el flux de dades i s'avaluen diferents tecnologies a utilitzar. S'aborden aspectes clau com la creació d'un catàleg de dades exhaustiu i l'estandardització de dades a un model de dades comú. Per a aquesta estandardització avaluem dos enfocaments per als processos *Extract-Load-Transform* (ELT), un basat en eines d'enginyeria de dades comunament utilitzades a la indústria tecnològica (Meltano, DBT, etc.) i un altre basat en una eina desenvolupada en el marc del projecte basada en tecnologies semàntiques (OntoBridge).

Destaquem la importància de desenvolupar una plataforma per al descobriment i anàlisi federada de dades, mantenint la privacitat i seguretat de les mateixes, demostrant la utilitat d'eines lliures com OBiBa, DataSHIELD i Beacon v2.

Aquest estudi proporciona una fulla de ruta per a futures investigacions i millores dins dels projectes de salut, contribuint així a la millora contínua de la gestió i l'anàlisi de dades.

Paraules clau: DATOS-CAT, principis FAIR, OMOP CDM, interoperabilitat, estandardització de dades, descobriment de dades, anàlisi federada de dades, DataSHIELD, Beacon.

Resumen

La medicina personalizada se basa en la utilización de datos individuales para desarrollar tratamientos específicos y efectivos. Este estudio se enfoca en contribuir al proyecto DATOS-CAT mediante el desarrollo de directrices, herramientas y protocolos para mejorar la interoperabilidad y visibilidad de la cohorte poblacional catalana GCAT y de la sub-cohorte COVICAT-CONTENT, asegurando su alineación con los principios FAIR y las políticas de la Comisión Europea.

El objetivo principal de DATOS-CAT es el desarrollo de la metodología y la infraestructura necesaria para la publicación de los datos generados en el contexto del proyecto GCAT/COVICAT para su uso por la comunidad científica, siguiendo estándares FAIR y asegurando la privacidad y seguridad de la información personal de los pacientes.

En este proyecto de final de máster, se define el flujo de datos y se evalúan diferentes tecnologías a utilizar. Se abordan aspectos clave como la creación de un catálogo de datos exhaustivo y la estandarización de datos a un modelo de datos común. Para esta estandarización evaluamos dos enfoques para los procesos *Extract-Load-Transform* (ELT), uno basado en herramientas de ingeniería de datos comúnmente usadas en la industria tecnológica (Meltano, DBT, etc.) y otro basado en una herramienta desarrollada en el marco del proyecto basada en tecnologías semánticas (OntoBridge).

Destacamos la importancia de desarrollar una plataforma para el descubrimiento y análisis federado de datos, manteniendo la privacidad y seguridad de los mismos, demostrando la utilidad de herramientas libres como OBiBa, DataSHIELD y Beacon v2.

Este estudio proporciona una hoja de ruta para futuras investigaciones y mejoras dentro de los proyectos de salud, contribuyendo así a la mejora continua de la gestión y análisis de datos.

Palabras clave: DATOS-CAT, principios FAIR, OMOP CDM, interoperabilidad, estandarización de datos, descubrimiento de datos, análisis federado de datos, DataSHIELD, Beacon.

Index

List of abbreviation

Acronym	Full name
1+MG	1+ Million Genomes
API	Application Programming Interface
B1MG	Beyond 1 Million Genomes
BSC	Barcelona Supercomputing Center
CDM	Common Data Model
CNAG	*Centre Nacional d'Anàlisi Genòmic*
CRG	*Centre de Regulació Genòmica*
CTE	Common Table Expression
dbt	Data Build Tool
DCM	Detailed Clinical Model
DDL	Data Definition Language
DMP	Data Management Plan
DQ	Data Quality
DQD	Data Quality Dashboard
ECHRS	European Community Respiratory Health Survey
EDA	Exploratory Data Analysis
EGA	European Genome-Phenome Archive
EHDEN	European Health Data & Evidence Network
EHDS	European Health Data Space
EHR	Electronic Health Records
ELT	Extract-Load-Transform
EMBL	European Molecular Biology Laboratory
EMBL-EBI	EMBL-European Bioinformatics Institute
EOSC	European Open Science Cloud
ETL	Extract-Transform-Load
EU	European Union
EuCanSHare	EU-Canada joint infrastructure for next-generation multi-Study Heart research
FAIR	Findable-Accessible-Interoperable-Reusable
FHIR	Fast Healthcare Interoperability Resources
GA4GH	Global Alliance for Genomics & Health
GCAT	Genomes for Life
GDI	Genome Data Infrastructure
GDPR	General Data Protection Regulation
H2020	Horizon 2020
HIPAA	Health Insurance Portability and Accountability Act
HL7	Health Level 7

HPI	Protected Health Information
IBEC	Institut de Bioenginyeria de Catalunya
ICD	International Classification of Diseases
ICU	Intensive Care Unit
IGTP	Germans Trias i Pujol Research Institute
IMI 2	Innovative Medicines Initiative 2
IMPaCT	*Infraestructura de Medicina de Precisión asociada a la Ciencia y a la Tecnología*
INB	Spanish National Bioinformatics Institute
INMA	*Infancia y Medio Ambiente*
ISCIII	Carlos III Health Institute
ISGlobal	Barcelona Institute for Global Health
LA2	Action Line 2
LeRAgs	*Lesión renal aguda en trabajadores agrícolas en España*
LOINC	Logical Observation Identifiers Names and Codes
MCC-Spain	*Estudio multicast-control poblacional*
NHGRI	National Human Genome Research Institute
NLP	Natural Language Processing
OBiBA	Open source software for BioBank
OHDSI	Observational Health Data Science and Informatics
OMOP	Observational Medical Outcomes Partnership
PADRIS	*Programa d'analítica de dades per a la recerca i la innovació en salut*
R2RML	RDB to RDF Mapping Language
RBD	Relational Database
RDF	Resource Description Framework
RxNorm	Standardized Nomenclature of Medicine Clinical Terms
SISCAT	Integrated Public Health System of Catalonia
SNOMED-CT	Systematized Nomenclature of Medicine – Clinical Terms
SNS	Spanish National Health System
SQL	Structured Query Language
UK	United Kingdom
URI	Uniform Resource Identifier

List of figures

Figure 23. *Screenshot showing the GCAT Network view.*

Figure 24. *Screenshot showing the search functionality by Collected Datasets in the Mica web data portal.*

Figure 25. *Screenshot showing the AH_Episodis collected dataset.*

Figure 26. *Screenshot showing all the information related to the variable circ_alta_desc.*

Figure 27. *Meltano's tap-csv extractor configuration.*

Figure 28. *Meltano's target-postgres loader configuration.*

Figure 29. *Diagram of the raw data tables (patient, diagnoses, procedures, episodes and external_cause) illustrating the initial structure and variables, prior to the transformation to the OMOP CDM.*

Figure 30. *Athena screenshot showing the list of downloaded vocabularies.*

Figure 31. *Diagram illustrating the tables within the OMOP CDM vocabulary schema, detailing the structure and the relationship between tables.*

Figure 32. *White Rabbit configuration screen for connecting to the datoscat_tfm PostgreSQL database.*

Figure 33. *White Rabbit scan screen.*

Figure 34. *Screenshot of the Rabbit-in-a-Hat diagram illustrating the relationships between the datoscat_tfm PostgreSQL database and the OMOP CDM v5.4.*

Figure 35. *Screenshot of the Rabbit-in-a-Hat with a detailed mapping of the "patient" table variables to the "person" table in the OMOP CDM v5.4.*

Figure 36. *Importing codes for the variable circ_alta into USAGI and the subsequent configuration.*

Figure 37. *USAGI mapping results showing the highest scoring match for the circ_alta variable.*

Figure 38. *SQL code to transfer raw data from the patient table to the OMOP CDM person table. The model is called person.sql.*

Figure 39. *Configuration settings of the person model within the SQL code used to map the patient table to the OMOP CDM person table.*

Figure 40. *Segment of the person model within the SQL code, showing the patients CTE.*

Figure 41. *Segment of the person model within the SQL code, showing the person CTE.*

Figure 42. *SQL code responsible for executing the macros used in the transformation process.*

Figure 43. *Screenshot showing the result of the execution of the transformation process from the raw data to the OMOP CDM.*

Figure 44. *Diagram of the OMOP CMD created. It illustrates the relationships and structure of the key tables: person, visit_occurrence, visit_detail, death, condition_occurrence, procedure_occurrence, observation, condition_era and observation_period.*

Figure 45. *Results of the Data Quality Dashboard Assessment.*

Figure 46. *Diagram illustrating the structure of the local ontology developed, showing the relationships and clinical concepts modeled from the Mica data dictionary.*

Figure 47. *Fragment of the automatically generated local ontology OWL file, highlighting the semantic representation of the variable sex.*

Figure 48. *Fragment of the R2RML code in Turtle format, used to map relational data to RDF, facilitating integration with the OWL local ontology.*

Figure 49. *Fragment of the OWL file of the standard dictionary ontology, related to gender_concept_id*

Figure 50. *Screenshot of the DataSHIELD administration in the Opal server.*

Figure 51. *Screenshot of the DATOS-CAT resources section in the Opal server.*

Figure 52. *DataSHIELD command to access "gender_concept_id".*

Figure 53. *Output of the DataSHIELD command showing the "gender_concept_id" results.*

Figure 54. *Histogram of the year_of_birth distribution.*

Figure 55. *Descriptive Statistics of the year_of_birth.*

Figure 56. *DataSHIELD command to retrieve the concept_id 320128 from the condition_occurrence table.*

Figure 57. *Output of the DataSHIELD command to retrieve the concept_id 320128 from the condition_occurrence table.*

Figure 58. *Implementation of Beacon v2 in our database.*

Figure 59. *Output after applying a filter to count the number of male records in our database.*

Figure 60. *Opal homepage where you are asked for credentials.*

Figure 61. *Opal administration main page.*

Figure 62. *Pop-up window for adding a new project in Opal.*

Figure 63. *DATOS-CAT project home screen in Opal.*

Figure 64. *DATOS-CAT project home screen in Opal II.*

Figure 65. *Pop-up window for adding/updating tables from a dictionary in Opal.*

List of tables

Chapter 1. Background and motivation

In recent decades, medicine has undergone a significant transformation from a traditional approach to a more personalized, patient-centered paradigm, known as **personalized medicine**. Personalized medicine[1], as defined by the National Human Genome Research Institute (NHGRI), *is an emerging practice of medicine that uses an individual's genetic profile to guide decisions made in regard to the prevention, diagnosis, and treatment of disease* (National Human Genome Research Institute, 2024 June). This concept has arisen as a transformative approach to healthcare delivery, promising personalized interventions and treatments, uniquely designed for each patient. This approach involves the comprehensive collection and analysis of large amounts of medical data, including genomic information, clinical data, imaging and laboratory results, among others.

Looking for significant advances in personalized medicine, prospective cohort studies emerge as fundamental building blocks for understanding the interaction between genetics, environmental and behavioral factors in human health. In contrast to case-control studies, prospective cohort studies provide a longitudinal perspective on health and disease. Data is gathered prior to the onset and treatment of disease, enabling the examination of interactions among multiple factors over time [2,3].

1.1. GCAT cohort

The *Germans Trias i Pujol* Research Institute (IGTP) has a strategic project to develop a population-based cohort called **Genomes for Life** (GCAT) [4,5]. GCAT is a prospective cohort study that recruited volunteers living in Catalonia with the aim of integrating and evaluating the role of epidemiological, environmental and omics factors in the development of chronic diseases, as well as defining the risk factors.

Prior a pilot study was conducted from April 2014 to June 2014, involving 191 participants across two centers, to assess the feasibility and refine procedures. Once the pilot study was successfully finished, the main cohort study was initiated, with a recruitment spanning from April 2014 to April 2018.

The cohort was open to any volunteer, with recruitment primarily targeting individuals enrolled in the Blood and Tissue Bank, a public agency of the Catalan Department of Health, known as *Banc de Sang i Teixits*. In line with one of the objectives to identify chronic disease events, adults aged 40 to 65 years were included in the cohort. In total, 19.329 participants were enrolled in the cohort. In addition to age criteria, participants must (i) possess an Individual Health System Identification Card, (ii) be current residents of Catalonia, and (iii) understand one of the two official languages in Catalonia, which are Catalan or Spanish. Exclusion criteria included (i) mental or health disorders that prevented communication, and (ii) whether the volunteer had plans to leave Catalonia within the next 5 years.

All participants who agree to take part in the study signed an informed consent that allows access to the electronic health records (EHR) of the Catalan Public Healthcare System for passive follow-up. More concretely, the EHR are extracted from the *Programa d'analítica de dades per a la recerca i la innovació en salut* (PADRIS) [6]. PADRIS aims to make health-related data available to the scientific community to promote research, innovation and evaluation in health. In this way, access to reuse health data generated by the Integrated Public Health System of Catalonia (SISCAT) is facilitated,

following the legal and regulatory framework, as well as the principles of ethics and transparency towards citizens. Additionally, other data was collected, including:

- An epidemiological questionnaire, which included 142 questions for men and 149 questions for women. Detailed information related to sociodemographic, socioeconomic status, lifestyle, personal and familiar medical history, etc. was asked.
- Biological samples, such as blood plasma, blood serum and white blood cells.
- Anthropometry measurements, such as cardiac frequency.

Omics techniques - genome, metabolome, epigenome and proteome - were also used to determine the molecular profiles of 6550 participants, of whom 6400 were unrelated and 50 were family trios. Among the unrelated participants, around 5459 genomics profiles were characterized by comprehensive genotyping. Moreover, the complete genome of 808 participants was sequenced.

The idea was to follow-up participants biannually for at least 20 years after recruitment to ensure long-term data collection and analysis. During this period, the advent of the COVID-19 pandemic created several significant challenges and opportunities for epidemiological research. In response to the health crisis, Barcelona Institute for Global Health (ISGlobal) and IGTP decided to start an epidemiological study called COVICAT project [7,8].

COVICAT project, also known as **COVICAT-CONTENT**, driven by the need to understand the impacts of the coronavirus SARS-CoV-2 in the Catalan population, is based on a part of the GCAT-Genomes for Life cohort, along with various segments of adult population and specific target groups, including: (i) *Estudio multicast-control poblacional* (MCC-Spain) [9], (ii) I*nfancia y Medio Ambiente* (INMA) - Sabadell Cohort [10], (iii) European Community Respiratory Health Survey (ECHRS) [11], (iv) Urban Training [12] and (v) *Lesión renal aguda en trabajadores agrícolas en España* (LeRAgs) [13].

The COVICAT-CONTENT sub-cohort was created in 2020 with the specific aim of monitoring COVID-19, through a prospective epidemiological study. The cohort involved around 10,246 individuals, most of them derived from the GCAT cohort, as mentioned before. The final objective was to evaluate determinants, ultimately contributing to the definition of predictive models and control policies from similar situations in the future.

Over the years, from the baseline to the present date, multiple follow-ups and data collections have been carried out, providing a detailed picture of the health and well-being of the Catalan population. These follow-ups, meticulously organized, are presented clearly and concisely in Table 1.

Events	Time period	Final number of participants
Baseline	2014-04-07 to 2018-04-11	19,329
First follow-up, known as "Follow-up 2018"	2018-05-30 to 2019-11-11	9,252 (8,913 complete questionnaire)
Second follow-up, known as "Covicat 2020"	2020-05-28 to 2020-11-27	10,246 (9,480 complete questionnaire)
Third follow-up, known as "Covicat-Content 2021"	2021-06-15 to 2022-02-06	8,021 (7,065 complete questionnaire)
Fourth follow-up, known as "Covicat-Content 2023"	2023-02-07 to 2023-09-06	5,215 (5,068 complete questionnaire)

Table 1. *List of data collection events along with their corresponding time period, as well as the final number of participants included in each follow-up. In some follow-ups, the global number of participants is indicated first and then in brackets, the final number of participants who have completed the questionnaire, considering the exclusion of duplicates. The final number of participants may vary because participants who request to be eliminated are not included.*

During the different follow-ups, extensive clinical, genomic and lifestyle data has been collected, as can be seen in Figure 1.

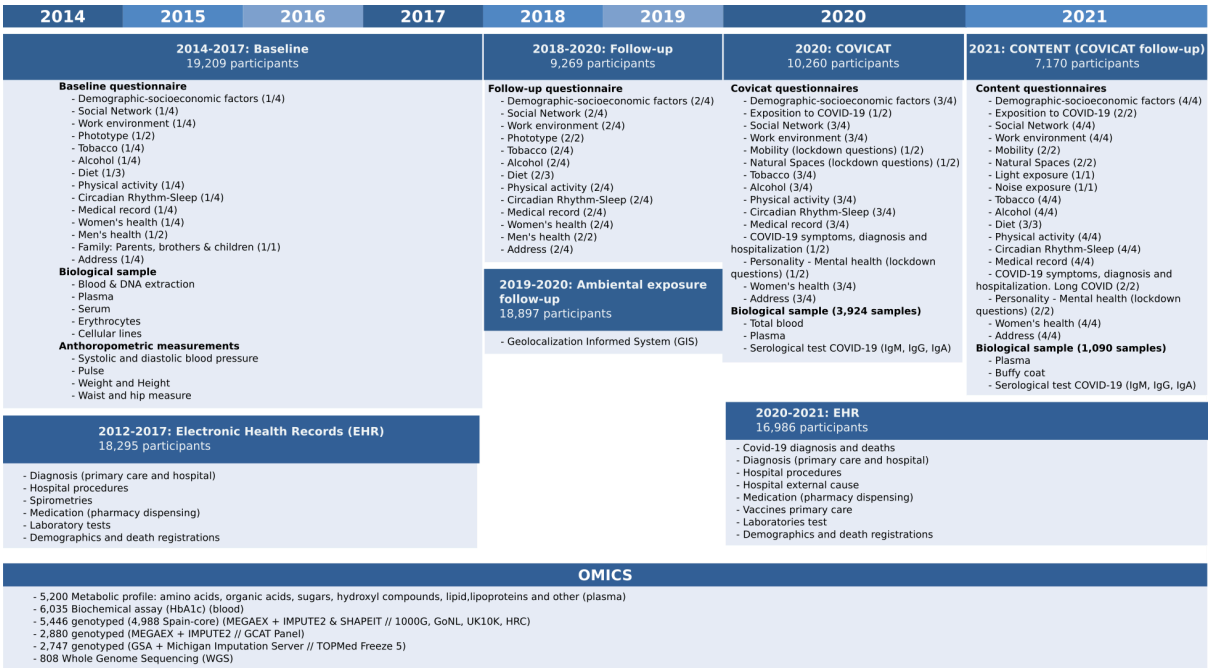


Figure 1. *Summary of GCAT data. Source: http://www.gcatbiobank.org/investigadors/en_gcat-summary-aggregate-data/*

However, it is important to note that some of the information that can be found about the cohort may be outdated. For example, the website mentions that sharing GCAT data is one of the aims of the project, and that data can be found in a Maelstrom Catalogue [14]. Reference is also made to resources

such as a catalog of GCAT variables in Mica. However, it is imperative to note that some of these resources may be decommissioned, disabled or out of date. This raises significant challenges, such as the loss of data or the replication crisis, both related.

1.2. Open science challenges related to data

For the past few years, the amount and diversity of generated data have seen an exponential growth. It is known that data is a valuable resource capable of driving research, enhancing knowledge, and aiding in decision-making across diverse domains. Despite their potential, effective data management presents several significant challenges that can hinder their usefulness and restrict their overall impact.

Replicability in science ensures that other scientists can repeat an experiment or a study to see if the results are the same. This is important because it helps scientists confirm that their findings are accurate and reliable. It also allows other scientists to build on the work of others, which is how science progresses. To make replicability possible, the data used in the experiment or study must be available.

An article published by Gibney, E., & Van Noorden, R. titled "*Scientists losing data at a rapid rate*" (2013) [15], a relevant issue in the field of scientific research is addressed. It notes that, when contacting the authors of various studies, it was observed that the availability of data decreased over time. Specifically, it was found that while the vast majority of data from studies published two years ago were still accessible, this proportion decreased by 17% per year. Furthermore, it was estimated that after 20 years, up to 80% of the data may no longer be available. These findings highlight the importance of addressing and resolving challenges related to long-term data availability and accessibility in scientific research.

Even when data is available, according to the Open Science glossary "*a large proportion of scientific studies published across disciplines do not replicate. This is considered to be due to a lack of quality and integrity of research and publication practices, such as publication bias, questionable research practices and a lack of transparency, leading to an inflated rate of false positive results.*" (Korbmacher, M., 2023) [16].

1.3. Open science solutions related to data

In 2016, Wilkinson et al. published "*The FAIR Guiding Principles for scientific data management and stewardship*" [17] manuscript, in which they designed and endorsed a set of principles called FAIR Data Principles, with the ultimate goal of improving data management and promoting its usefulness and reuse in research and science. This initiative is based on four foundational principles, which are detailed below according to the original article and the GO FAIR Initiative [18, 19]:

- **Findable -** *"Metadata and data should be easy to find for both humans and computers."*.
  To be Findable:
    F1. (meta)data are assigned a globally unique and persistent identifier.
    F2. data are described with rich metadata (defined by R1 below).
    F3. metadata clearly and explicitly include the identifier of the data it describes.
    F4. (meta)data are registered or indexed in a searchable resource.

- **Accessible -** *"Once the user finds the required data, she/he/they need to know how they can be accessed, possibly including authentication and authorization.".*
  To be Accessible:
  A1. (meta)data are retrievable by their identifier using a standardized communications protocol
  A1.1 the protocol is open, free, and universally implementable
  A1.2 the protocol allows for an authentication and authorization procedure, where necessary
  A2. metadata are accessible, even when the data are no longer available

- **Interoperable -** *"The data usually needs to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing.".*
  To be Interoperable:
  I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
  I2. (meta)data use vocabularies that follow FAIR principles
  I3. (meta)data include qualified references to other (meta)data

- **Reusable -** *"Metadata and data should be well-described so that they can be replicated and/or combined in different settings.".*
  To be Reusable:
  R1. meta(data) are richly described with a plurality of accurate and relevant attributes
  R1.1. (meta)data are released with a clear and accessible data usage license
  R1.2. (meta)data are associated with detailed provenance
  R1.3. (meta)data meet domain-relevant community standards

These four principles refer to three different types of entities: (i) data, (ii) metadata and (iii) infrastructure.

Furthermore, the European Commission has established **open science** as a policy priority and as a main focus in research and innovation funding programs [20]. This new approach stands out for its ability to improve the quality, efficiency, and responsiveness of research projects. This requires beneficiaries to make publications open access and their data *as open as possible and as closed as necessary*.

As mentioned, open science has become a priority for the European Commission, and this is reflected in its 8 ambitions of the EU's open science policy. These ambitions include the promotion of open data, assuming the FAIR principles. However, in the field of medicine, most health data is considered personal data due to its sensitive and confidential nature, this adds an additional layer of complexity to the management and sharing of information in the framework of open data initiative.

Increasingly stringent regulations around data protection, such as the **General Data Protection Regulation** (GDPR) [21] in the European Union (EU) or the **Health Insurance Portability and Accountability Act** (HIPAA) [22] in the United States, impose rigorous requirements to ensure the privacy and security of personal information. GDPR covers all types of personal data, while HIPAA

specially focuses on the protection of medical and health information, known as Protected Health Information (HPI). Working with health data is therefore a challenge that needs to be addressed, as it is valuable to the scientific community and can lead to improvements in healthcare.

One promising solution is **federated data analysis** [23, 24, 25]. Federated data analysis is a method that allows multiple databases to be analyzed together without the need to move or copy the data. This approach maintains data within its original jurisdiction, ensuring compliance with data protection laws like GDPR and HIPAA. By bringing the analysis to the data, federated data analysis minimizes security risks and maintains data privacy. Researchers can perform analyses across various datasets securely, facilitating global collaboration and accelerating scientific discovery. This method also reduces costs related to data transfer and storage, increases compliance with data protection regulations, and promotes sustainability by minimizing data duplication. Ultimately, federated data analysis democratizes access to valuable health data, enhancing research capabilities while safeguarding privacy.

However, federated data analysis requires a common data model, enabling queries to retrieve data from all harmonized databases, ensuring accurate data processing. **Data harmonization** *refers to all efforts to combine data from different sources and provide users with a comparable view of data from different studies* (Inter-university Consortium for Political and Social Research, n.d.) [26], i.e. the process of transforming and standardizing the data to make them comparable. Harmonization of data is also considered a challenge, as combining data from different sources involves addressing variations in data collection methods, structures, vocabularies, and interpretations. To address this issue, various standards and protocols have been established.

Another key concept is **data standardization**, *a data processing workflow that converts the structure of different data sets into a common data format* (Advances in Computers, 2021) [27]. In healthcare, it plays a crucial role due to the inherent sensitivity and complexity of medical information. Standardization ensures uniformity and consistency between different sources and types of data, known as interoperability.

There are some widely extended healthcare standards used for data standardization [28, 29, 30]:

- **Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM).** The OMOP CDM is widely used in Europe and is designed to facilitate the analysis of observational data. It supports the conduct of research to identify and evaluate associations between interventions and outcomes. OMOP organizes data into a person-centric model with standardized vocabularies, ensuring interoperability across disparate data sources. Its extensive adoption in Europe is due to its robustness in handling diverse healthcare datasets and supporting large-scale, collaborative research initiatives.

- **Fast Healthcare Interoperability Resources (FHIR).** FHIR is developed by **Health Level 7 (HL7)**, a global "*standard for medical informatics exchange between healthcare providers"* (Kimura M, 1999) [31]. Therefore, FHIR focuses on facilitating the exchange of healthcare information electronically. It enables quick data exchange by using discrete and independent data elements called resources, which can be combined into complex documents. FHIR is widely used for real-time data exchange in clinical settings, making it suitable for applications requiring rapid data integration and interoperability.

- **OpenEHR.** OpenEHR is a community working on a comprehensive open standard specification for EHRs. It emphasizes semantic interoperability and long-term data preservation. OpenEHR uses archetypes and templates to define clinical concepts, ensuring that data is consistently interpreted across different systems and contexts.

Data standardization using these models enhances the ability to integrate and compare healthcare data, thereby improving research quality and healthcare outcomes. The choice of OMOP CDM as the primary model in Europe is driven by its comprehensive structure, extensive support for various types of health data, and its capability to facilitate large-scale, federated research projects.

1.4. DATOS-CAT

Revisiting the previously mentioned scenario, it is evident that the Catalan cohort does not conform to the FAIR principles, nor does it adhere to the European Commission's policies. Hence arises the **DATOS-CAT** project [32], a dynamizing action aimed at addressing the identified limitations and enhancing adherence to the FAIR principles.

The main objective of DATOS-CAT is to increase the visibility and scientific impact of the Catalan cohort, ensuring alignment with current standards and best practices, as well as, safeguarding the privacy and security of patients' personal information .

This ambitious project of the *Plan Complementario de Biotecnología Aplicada a la Salut* is developed through the collaboration of seven different Catalan institutions: Barcelona Supercomputing Center (BSC) as scientific coordinator, *Institut de Bioenginyeria de Catalunya* (IBEC), *Institut de Recerca Germans Trias i Pujol* (IGTP), *Centre de Regulació Genòmica* (CRG), *Centre Nacional d'Anàlisi Genòmic* (CNAG), *Institut de Salut Global de Barcelona* (ISGlobal), a centre promoted by “Fundació la Caixa”, and *Hospital Clínic de Barcelona*.

As an active participant of DATOS-CAT, my involvement is focused on contributing to the design, development, and evaluation of the project. Specifically, I am committed to leading the implementation of tools and best practices for data cataloging and data standardization, as well as defining the steps for data analysis and data discovery on the data collected from the Catalan cohorts.

This participation not only provides an opportunity to apply the knowledge acquired during my master's program but also represents a commitment to improve data management and data re-use in a context of scientific and social relevance.

Chapter 2. Objectives

This Master's Thesis is part of the DATOS-CAT project, a national initiative that aims to increase the visibility and scientific impact of the population-based cohorts developed in Catalonia, the GCAT cohort and the COVICAT-CONTENT sub-cohort. At the same time, the project seeks contribute to the development of procedures applicable to other cohorts, by improving the level of interoperability of their data in the context of the FAIR (Findable, Accessible, Interoperable, Reusable) data ecosystem principles to facilitate their exploitation and scientific use.

A key aspect of the DATOS-CAT project is the use of existing technological infrastructure to manage and analyze the data. The European Genome-Phenome Archive (EGA) infrastructure will serve as the backbone for storing genomic data, and the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) will provide a standardized framework for structured clinical data.

On the other hand, to address the challenge of sharing the cohort data with the scientific community, in relation to privacy concerns and data governance issues, DataSHIELD will be used to perform non-disclosive federated analysis. Another key component intended to be used is the GA4GH Beacon v2, a tool that allows federated data discovery (i.e. discover data sources that meet specific genomic or phenotypic criteria).

All these technologies not only ensure the robustness and reliability of the data, but also enhance its interoperability, enabling seamless integration with other cohorts (such as UK Biobank) and datasets, facilitating a wide range of research applications.

2.1. Objectives of the Master’s Thesis

The main objective of this Master's Thesis is to contribute to the DATOS-CAT project by establishing guidelines, tools and protocols to facilitate the implementation and use of these different technologies, as well as contributing to the improvement of interoperability and visibility of the Catalan population-based cohort.

The specific objectives of the project are detailed below:

1. Define a **data flow** within the DATOS-CAT project, for the whole data life cycle of the different datasets generated from the GCAT cohort, analyzing, and defining an efficient path of data from its origin to its final use with the project.

2. Develop a comprehensive **data catalog** that encompasses the various data sets available, ensuring comprehensive documentation and organization of the metadata to facilitate their accessibility and use.

3. **Standardize to a common data model** that harmonizes the structure and format of data across cohorts, facilitating interoperability and consistency in data management and analysis.

4. Compare and evaluate **two different Extract-Load-Transform (ELT) processes** within the DATOS-CAT framework, assessing their effectiveness, efficiency, and compatibility with the established data model.

5. Deploy a platform that allows **federated data discovery** and **analysis** across multiple datasets within the Catalan cohorts, while ensuring data privacy and security.

6. Develop **synthetic data** to test and validate the different objectives mentioned above. The synthetic data will attempt to adequately reflect the characteristics and variability of the real data, while preserving privacy and confidentiality.

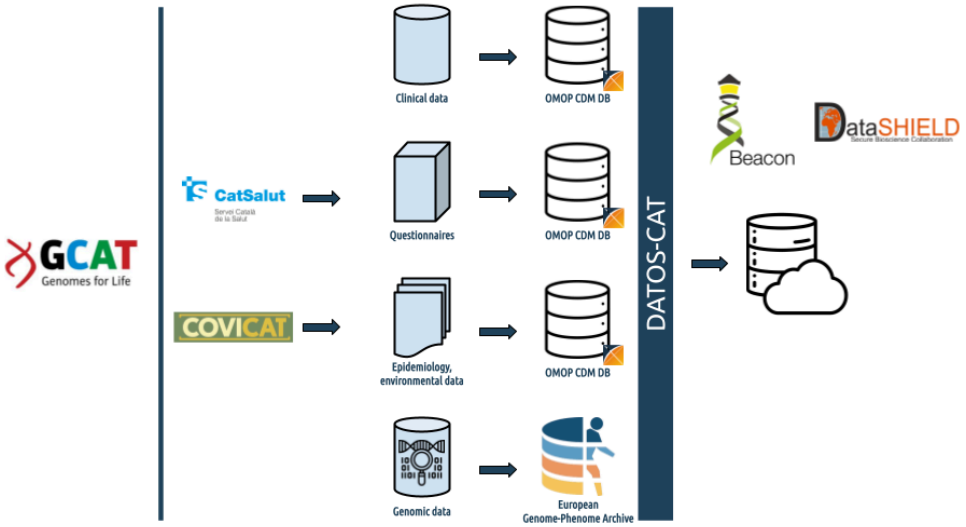


Figure 2. *Overview of the DATOS-CAT project, where the main components of the project can be seen.*

All the methods and results of the final master's thesis have served and will serve to lead and guide the DATOS-CAT project team. With the final objective of achieving a satisfactory execution of the project within the specified deadlines.

Chapter 3. State-of-the-art

3.1. International Context

The **Global Alliance for Genomics & Health** (GA4GH) [33] is an international non-profit organization that builds the foundation for a broad and responsible use of genomic data, by setting standards, policy frameworks and tools for their use within a human rights framework.
GA4GH was founded in 2013, and currently includes more than 500 organizations connected to genomics - across healthcare, research, patient advocacy, ethics, government, etc. Members of different organizations work together in initiatives known as *Driver Projects*, which play a key role in shaping GA4GH products and their application in real genomic data. There are also GA4GH *products*, which are tools and resources developed by the alliance to address specific genomics-related challenges. Below two key products are presented:

- **Phenopackets** - *offers a human and machine-readable way to structure clinical and phenotypic data about a patient or individual* (GA4GH, n.d.) [34].
- **Beacon** - *enables researchers to discover new, relevant data sets while honoring patient consents and legal requirements* (GA4GH, n.d.) [35].

Another initiative, more focused on the use of observational data in the field of health, is the **Observational Health Data Science and Informatics** (OHDSI) [36]. It is an international network of researchers and observational health databases, with a central coordination center based at Columbia University.

OHDSI was founded in 2014 because of the success of the **Observational Medical Outcomes Partnership** (OMOP) [37]. The OMOP Common Data Model (CDM) is a standardized framework for data representation and analysis, i.e. it was designed to standardize both the structure and content of the observational data and to allow analysis that produces reliable evidence.

With more than 2,000 collaborators, OHDSI aims to enhance healthcare by empowering a community-driven approach to generate evidence that facilitates improved health decisions and care. Leveraging its OMOP CDM, OHDSI's data network enables collaborative, federated analytics among its partners.

3.2. European Context

To contextualize the project within the European framework, it is essential to understand the current scenario and the proposed challenges.

Since open science has become a priority for the European Commission, the implementation of a **European Open Science Cloud** (EOSC) has become one of the ambitions of the EU's open science policy [38]. EOSC aims to provide to the European researches, companies and citizens a federated and open multi-disciplinary environment, to make it possible to publish, find and reuse data, tools and services for research, innovation and education purposes.

A very related initiative is the **European Health Data Space** (EHDS) [39], which is one of the 15 European Common Data Spaces. EHDS focuses on creating a common framework for the secure exchange and use of health-related information across Europe. It includes several initiatives, such as the implementation of a Federated European infrastructure for genomics data, related to a previous initiative called 1+ Million Genomes (1+MG). EHDS leverages the federated and open environment provided by EOSC to facilitate better access, sharing, and reuse of health data for research and innovation. This synergy ensures that health data is accessible and interoperable, driving improvements in healthcare delivery and research outcomes.

Following Digital Day 2018, a coalition of 24 European (EU) countries, plus the United Kingdom and Norway, ratified the Member State Declaration, which aimed for access to a minimum of 1 million sequenced genomes within the EU by 2022. This strategic plan involves intensified efforts to build a European data infrastructure tailored to genomic data and the implementation of standardized national protocols to facilitate federated data accessibility. The 1+MG initiative [40, 41, 42] aimed to enable secure access to genomics, and its associated clinical data, across Europe to improve research, allow personalized healthcare and support health policy making. This initiative is implemented through the joint efforts of the **Beyond 1 Million Genomes** (B1MG) [43] project and the **Genome Data Infrastructure** (GDI) [44], both European-funded projects.

The B1MG project started in 2020 with the objective of providing guidelines for the implementation of the 1+MG initiative. On the other hand, the GDI project, launched in 2022, builds on the preliminary work done by B1MG, focusing on developing the technical infrastructure needed to access genomic data.

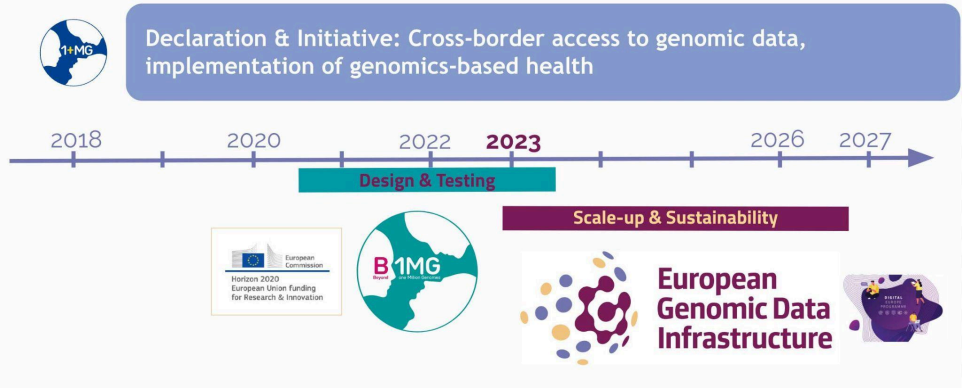


Figure 3. *Timeline illustration of the 1+MG framework, featuring Horizon 2020, B1MG and GDI initiatives. Source: https://framework.onemilliongenomes.eu/about-the-framework* [42]

Focusing on European organizations, there is **ELIXIR** [45], an intergovernmental organization that brings together life science resources, - including databases, software tools, training material, cloud storage, etc. - across Europe. The main goal is to coordinate these resources to form a single structure. Founded in December 2013, the ELIXIR organization includes 22 members and 3 Observers (i.e. countries working towards full membership) that work together using a "*Hub and Nodes*" model.

- **ELIXIR Nodes**. Each member country has a "Node", i.e. a collection of research institutes in the country. However, the European Molecular Biology Laboratory (EMBL) is the only Node that is not associated with a country, due to it is an intergovernmental organization.
  Each node has also a lead institute that supervises the work of that Node. For example, ELIXIR Spain's lead institute is the Barcelona Supercomputing Center (BSC).

- **ELIXIR Hub**. The ELIXIR Hub is basically like a headquarter that coordinates the work done by the Nodes. It is located at the Wellcome Genome Campus in Hinxton, Cambridge.

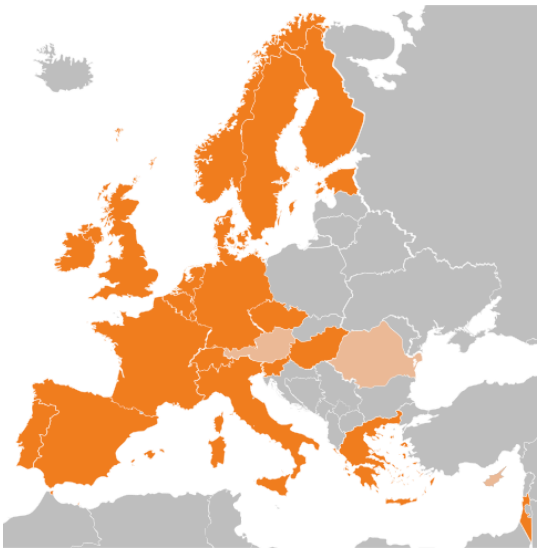


Figure 4. *Map of ELIXIR members and observers, highlighting member countries in orange and observers in light orange. Source: https://elixir-europe.org/about-us/who-we-are* [45]

Within the ELIXIR infrastructure part, the **European Genome-Phenome Archive** (EGA)[46, 47] serves as a Core Data resource. This collaborative partnership between the **EMBL-European Bioinformatics Institute** (EMBL-EBI) and the CRG underscores its pivotal role in securely storing and facilitating access to genetic and phenotypic data for research purposes. More specifically, EGA serves as a global network service used to store and share personally identifiable genetic, phenotypic and clinical data. It hosts data from individuals whose consent agreements authorize the sharing of these data for specific research.

The EGA project [48] is governed by strict protocols for data management, storage and distribution to ensure data security and privacy. Data submitted to the EGA is accessible only to authorized researchers who have obtained the necessary permissions, ensuring compliance with ethical and legal standards. The EGA's infrastructure supports various data types, including raw sequencing data, genotype data, and phenotypic data, making it a comprehensive resource for biomedical research.

Another organization is the **European Health Data & Evidence Network** (EHDEN) [49, 50], a consortium of 25 European participants, which is part of the Innovative Medicines Initiative 2 (IMI 2) program of Horizon 2020 (H2020). The aim of the EHDEN project was to exploit patient-related data contained in EHRs and other types of databases, whether structured or unstructured, to improve clinical practices by improving understanding of diseases and treatment methods.

The idea is to create a platform that complies with FAIR principles and that is in alignment with OHDSI, which is its main partner. The core of this consortium is the use of OMOP CDM, standardized outcomes evaluation and open-source analysis.

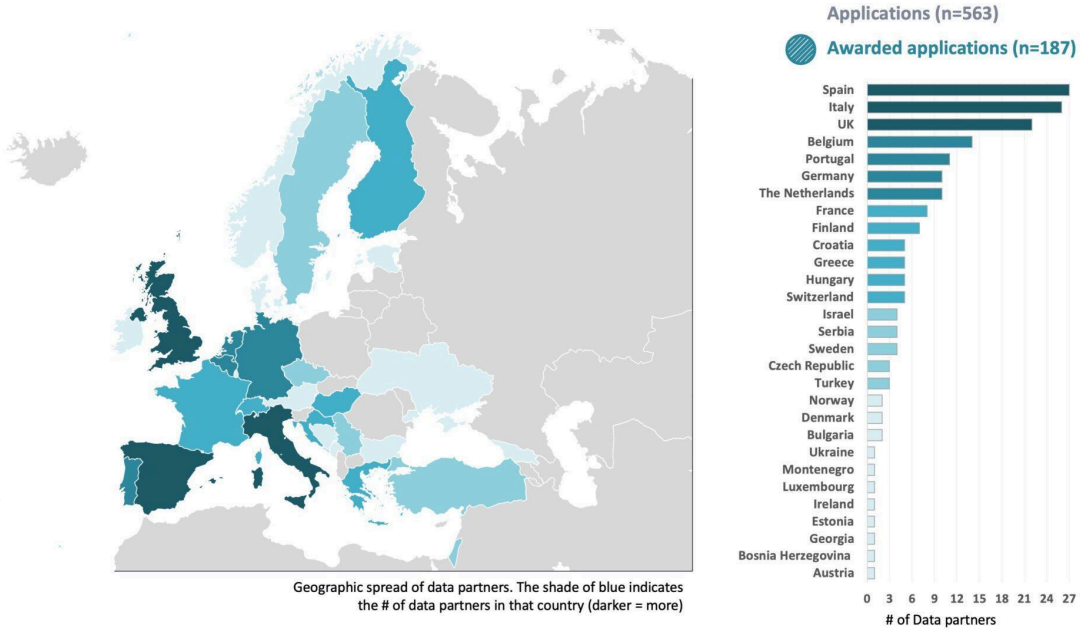


Figure 5. *Map of EHDEN data partners which are mapping their data to OMOP CDM. There are 187 data partners from 29 different countries. Source: https://www.ehden.eu/datapartners/* [49]

3.3. National Context

Within the national landscape, the **IMPaCT** initiative [51], which stands for *Infraestructura de Medicina de Precisión asociada a la Ciencia y a la Tecnología*, has emerged as an important effort to advance personalized medicine practices within Spain.

IMPaCT's Strategic Plan is configured around three strategic axis listed below, and two transversal lines: (i) *Ética e Integridad Científica*, and (ii) *Internacionalización*.

- **IMPaCT-cohort**. This program aims to include 200,000 participants in a prospective study, to establish a population-based cohort representative of the population of Spain. The ultimate goal is to design precision strategies and predictive models for the prevention, early diagnosis and treatment of the main diseases.

- **IMPaCT-data** [52]. This program aims to develop recommendations and best practices for the entire health data environment, from data collection to interoperability and analysis.

- **IMPaCT-genómica**. This program aims to develop standard operating procedures, as well as the infrastructure, for the proper management, development and analysis of genomic data.

The DATOS-CAT project is framed within the Action Line 2 (LA2) of the *Plan Complementario Biotecnología Aplicada a la Salud* [53]. The action proposed under LA2 is directly related to both the national IMPaCT-data initiatives (in terms of technology) and IMPaCT-cohorte (as specific application).

Figure 6. *IMPaCT Strategic Plan, where the 3 strategic axis are depicted as squares and the 2 transversal lines as arrows. Source:*https://www.isciii.es/QueHacemos/Financiacion/IMPaCT/Paginas/Plan.aspx [54]

The **Spanish National Bioinformatics Institute** (INB) [55], known as *Instituto Nacional de Bioinformática*, is a network of 26 nodes located at 14 different research institutions. INB was founded in 2003, and since January 2018 it has been the bioinformatics technology platform of the Carlos III Health Institute (ISCIII).

The INB serves in the coordination, integration and development of Spanish bioinformatics resources, in different areas such as genomics, proteomics and translational medicine. The institute has two main objectives: (i) increase and maintain the alignment with the ELIXIR organization, and (ii) increase the translational capacity toward the Spanish National Health System (SNS). It is important to mention that Spain has been a member of ELIXIR since 2015, with the INB as the Spanish national node.

3.4. Similar Projects

The presented project, DATOS-CAT, shares many synergies with other national and European projects, as many of them focus on longitudinal medical data.

If we focus on cohort projects, one standout example is the **United Kingdom** (UK) **Biobank** [56], a large population-based prospective study with over 500,000 participants. As mentioned by Sudlow C. et al. in the article "*UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age*" [57], this large-scale biomedical database is available online for open access to researchers, without the need for collaboration, enabling new scientific discoveries to improve public health.

During COVID-10 pandemic, EDHEN launched a series of calls to convert different data sets to OMOP, with the goal of providing new COVID-19 insights on a large scale across countries. Performing federated analysis across different data sets can be challenging, as the data may have different terminologies, use different schemas, etc. However, this is solved by converting the data to a common data model such as OMOP. Papez et al. published "*Transforming and evaluating the UK Biobank to the OMOP Common Data Model for COVID-19 research and beyond*" [58], which explains how they converted the UK Biobank data to OMOP CDM v. 5.3.

On the other hand, another project in line with the UK Biobank's ambition, called IMPaCT-Cohorte [59], aims to establish a Spanish cohort of approximately 200,000 people. Although still at an early stage, IMPaCT-Cohorte seeks to collect and integrate lifestyle data, clinical and genetic information, etc. with the aim of generating predictive models that allow the implementation of Precision Medicine.

In this section, we also decided to explore relevant projects that share infrastructure similarities. In this context, the **EU-Canada joint infrastructure for next-generation multi-Study Heart research** (EuCanSHare) [60] project stands out. EuCanShare aims to establish a cross-border data sharing and multi-cohort cardiovascular research platform, and it has an infrastructure analogous to the one proposed in this work.

Chapter 4. Design and development

4.1. Data Flow

The primary aim of the DATOS-CAT project is to enhance the visibility and scientific impact of the Catalan population-based cohorts. Since most of the information and resources available on the GCAT cohort are outdated, decommissioned, or disabled; we should implement the **FAIR principles** throughout almost the entire data life cycle.

We did a first review and we saw that the lack of updated information may affect the "F" for Findable, as users may have difficulty locating and accessing relevant GCAT project data and resources. In addition, the possible unavailability of resources could compromise the "A" of Accessible, limiting the ability of researchers to access and use the data in their research projects. Also, the lack of updating and possible obsolescence of the resources could hinder the “I” for Interoperability of the data by not having access to up-to-date standards and protocols needed to effectively integrate and combine the data with other data sets. Finally, the lack of availability of up-to-date resources and data could undermine the “R” for Reusability of GCAT project data in future research, as researchers may find it difficult to validate, replicate or build on existing findings.

In this context, it is crucial to address and resolve these issues to ensure that GCAT data and resources comply with the FAIR principles. Therefore, we set up a data flow that starts with data cataloging. Here, the metadata related to each database is described. Once the different data sources have been described, the standardization of the raw data to the common data model is undertaken. Finally, a last step that enables federated data analysis, allowing local data from different institutions to be analyzed without compromising patient privacy and security. In general, the whole process can be grouped into three main phases: (i) data cataloging, (ii) data standardization, and (iii) data analysis. Figure 7 shows an overview of the data flow.

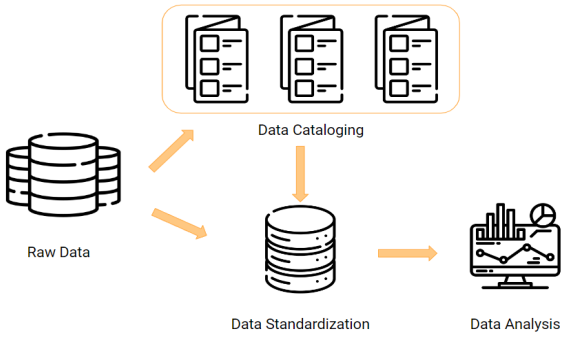


Figure 7. *Data flow diagram illustrating the three main phases of DATOS-CAT project: Data Cataloging, Data Standardization and Data Analysis.*

Several tools have been evaluated to address the three phases of the project. After a detailed analysis, the use of the OBiBa environment has been identified as the most suitable option for both data cataloging and data analysis in a federated environment. It has been also concluded that the adoption of OMOP, as a common data model standard, will ensure data consistency and interoperability at all stages of the work.

Regarding **OBiBa** [61], which stands for Open-source software for BioBank, it is an international project dedicated to build open-source software for epidemiological studies. Aligned with the Maelstrom Research program objectives, the OBiBa software package covers the entire data management lifecycle: collection, integration, harmonization, sharing and analysis. During the project, several tools (products) were developed to address the different stages of the data lifecycle. A brief description of the products we have used for the DATOS-CAT project follows:

- **Opal** - is the core data warehouse application, serving as a central data repository for data curation, analysis and harmonization. This server application provides tools to import, transform and describe data.

- **Mica** - is a software application that allows the creation of web-based data portals for large-scale studies and research consortia. It is a metadata catalog that facilitates a structured description of cohort studies, as it allows the definition and description of networks, studies, data sets and variables. In addition, it allows the search of all of them.

- **Agate** - is a web application that provides user authentication, user profile management and user notification to the OBiBa related services.

- **Rock** - is the OBiBa's R analysis server, i.e. Opal uses Rock to form clusters of R servers and handle analysis requests on multiple hosts.

- **DataSHIELD** [62] - is an open-source software that provides an infrastructure for secure and distributed statistical analysis, i.e. federated analysis. It is a platform designed to protect data privacy while enabling collaboration and statistical analysis on data distributed across different locations, without the need to share the actual data between the parties involved. Instead of transferring data, DataSHIELD uses advanced encryption techniques and secure communication protocols to perform complex statistical queries on local data and then, securely combine the results.

  There are different types of infrastructures: (i) Multi-site DataSHIELD, (ii) Multi-site DataSHIELD with reference to a resource , and (iii)Single-sit DataSHIELD. In the DATOS-CAT project, it is intended to use the second type of infrastructure, the main DataSHIELD infrastructure schema can be seen in Figure 8.

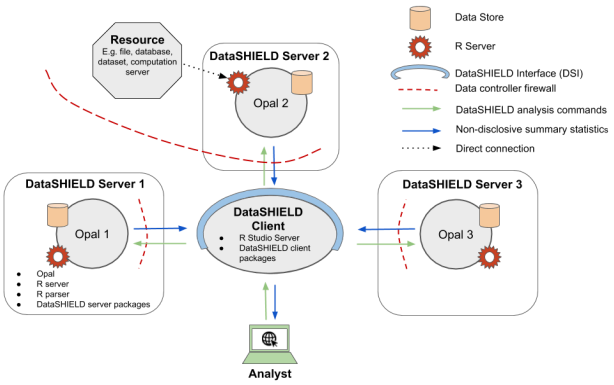


Figure 8. *Multi-site DataSHIELD with reference to a resource infrastructure schema. Source:* https://www.datashield.org/about/about-datashield-collated [62]

To streamline the configuration and deployment process of all the essential tools, such as Opal and Mica, a collaborative project called EUCAN-Connect [63] decided to provide the required applications for each data server through a Docker stack, referred to as the **Coral Stack**. The Coral stack consists of a series of Docker containers, built from images originally created by OBiBa, adapted and customized for easy configuration and deployment. This approach ensures consistency, scalability and efficient setup across environments, due to it simplifies the management of dependencies and version control.

4.2. Data Catalog

In line with the FAIR principles, it is essential to have a tool such as a data catalog to facilitate the location of information, either within an organization or in the context of a large data cohort, as is the present case [64].

Within the framework of this project, effective management of the data collected is also important to ensure the quality and integrity. In this regard, a detailed and comprehensive Data Catalog has been developed, which plays a central role in the organization and documentation of the data collected throughout the study.

More concretely, the DATOS-CAT project has deployed the Mica infrastructure to centralize the previously collected data from the GCAT cohort and COVICAT sub-cohort. This catalog not only facilitates detailed documentation of the collected data but also ensures its findability and accessibility for further analysis and use in future research. In general terms, it promotes the collection and incorporation of new data, as well as it improves their dissemination and exploitation.

4.2.1. Data Catalog Template

The Opal platform offers a template to fill in new data set information, and we decided to use it as a template to create the data catalog. Moreover, almost all the fields are customizable, i.e. the information can be added, modified or deleted according to the needs of the study.

To ensure that the cataloging of all the datasets was the same, we internally decided the fields and the variable description rules. The Excel template consists of two main spreadsheets that serve as a structural framework for the organization and understanding of the data: (i) variables tab and (ii) categories tab.

The variables tab is intended for the detailed explanation of the variable attributes contained in the data set. Meanwhile, the categories tab is reserved for the explanation of the categorical variables present in the data set, and at the same time, to describe specific cases of variables with missing or outliers.

4.2.2. Data Catalog Automation

In order to optimize the process of filling in the Excel template mentioned above, a Python script was developed as part of the methodology. This script was designed with the specific purpose of

facilitating and speeding up the task of completing the required fields necessary to build up the data catalog.

To start creating the data catalog, the user first need to use the OHDSI White Rabbit tool [65, 66], an open-source software tool that scans the source data - in our case, data sets in Excel or csv formats - and provides a summary of the tables, fields and values that appear in them, generating a report in Excel format called *Scan Report*. Once generated, the user only needs to specify the path to this file. Then, the script reads the report and transforms the information into the format required by the data catalog. As output, the user receives an Excel file which contains almost all the information needed to catalog the data set. Figure 9 shows a schematic view of the data set cataloging process. The data catalog automation script is available in the Supplementary Material.

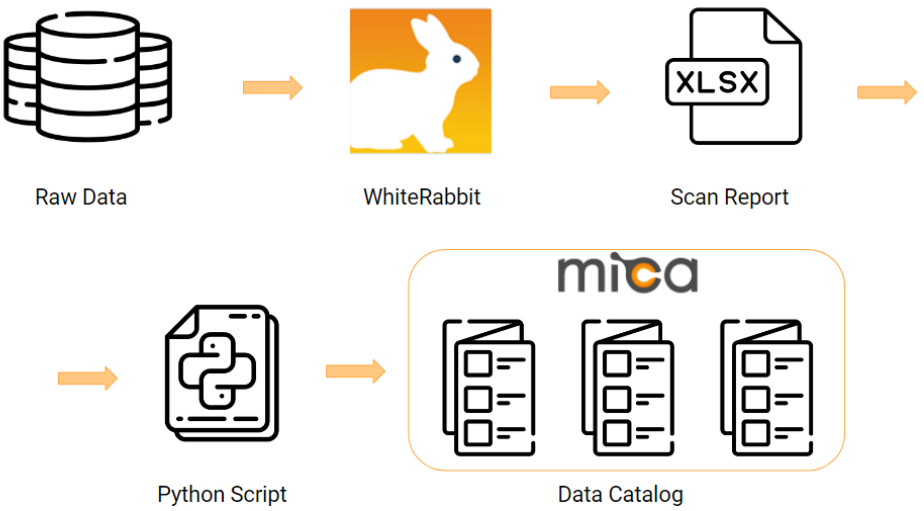


Figure 9. *Schematic view of the automation process of data set cataloging. Source: https://github.com/DATOS-CAT/DataCatalog* [67]

This automated approach has several advantages, including significantly reducing the time spent on manual tasks and minimizing human error in data entry. In addition, the inherent flexibility of Python allows the script to be adapted to the specific requirements of any other project.

4.2.3. Data Catalog OBiBa Deployment

Once the metadata of a data set is documented, the information must be integrated into the OBiBa software, through two main tools: Opal and Mica. The workflow between these two platforms is detailed below.

The first step is to upload and organize the data in Opal. Opal acts as a database and data management platform, where the organization follows several hierarchical levels.

- **Projects**. The information is organized into projects, which represent specific studies. Within a project, different tables, variables and resources can be found.

- **Tables**. Most data are stored in tables, which represent specific sets of information. Tables should be in Excel or CSV format. The metadata (i.e. data dictionary) related to a dataset is imported at this hierarchical level.

- **Variables**. Variables are the columns within the tables that contain the individual data. Each variable has properties defined in the data dictionary, including its name, type, description, etc.

- **Resources**. In some cases, data is not stored in formats such as Excel or CSV, and therefore cannot be treated as tables in Opal. For these cases, Opal allows the management of additional elements known as resources. A resource is described as *a dataset or computation unit whose location is described by a URL and access is protected by credentials* (Opal, n.d.) [68].

Within each Opal project, there is the option to add tables (those must be in excel or csv format) and/or resources. In our case, we have created a project called DATOS-CAT and we have imported the data dictionaries (i.e. the Excels with the description of variables) as tables.

After organizing the data in Opal, the next step is to create the catalog in Mica. Mica [69] is the OBiBa server for creating data web portals, and it supports several types of documents, some of them are described below:

- **Network** - "*a group of epidemiological studies that has specific research interests*". It includes studies.

- **Study** - "*any epidemiological study (e.g. cohort, case control, cross sectional, etc.) conducted to better understand the distribution and determinants of health and disease*". It can include one or more populations.

- **Population** - "*a set of individuals sharing the same selection criteria for enrollment in a study*". Depending on the number of follow-ups, a population can be linked to one or more data collection events.

- **Data Collection Event** - "*a collection of information on one or more population(s) over a specific period of time*". It includes data sets.

- **Harmonization Initiative** - "*a research project harmonizing data across individual studies to answer specific research questions*".

- **Collected data set** - "*metadata about the variables collected within a data collection event. The metadata is described using a standardized format of data dictionary which provides information on collected variables' definitions and characteristics*", i.e. what we have called Mica data catalog. It is associated with a study, specifying a data collection event.

Each type of document has an internal structure to allow associations between them. It also plays an important role in the organization of the metadata. Therefore, within the framework of our project, we have identified and defined several types of documents that are essential. The following table presents each document type and how it is specifically applied in the context of the DATOS-CAT project.

Document type	DATOS-CAT metadata structure
Network	GCAT project
Study	- GCAT study - COVICAT study
Population	- GCAT cohort - COVICAT cohort: MCC-Spain, INMA, ECHRS, Urban Training, and LeRAgs.
Data collection event	- Baseline - Follow-up 2018 - Covicat 2020 - Covicat-Content 2021 - Covicat-Content 2023
Collected data set	- Baseline Survey data set - EHR data sets (primary care spirometry, primary care vaccines, hospital care diagnoses, etc.) - Covicat 2020 Serology data set - Etc.

Table 2. *Description of the types of documents Mica hands with and their application in the DATOS-CAT project.*

Once each type of document was identified, the next step was to create the different types of documents. To streamline the process, we created a BSC Opal environment and uploaded the metadata information related to the hospital care (AH) EHRs. The datasets related to AH EHRs have the same data collection event, encompass the same population, within the same studio and the same network. The structure that has been developed for this Master's thesis can be seen in Table 3.

Document type	Master metadata structure
Network	GCAT project
Study	GCAT study
Population	GCAT cohort
Data collection event	Hospital Care Electronic Health Records (EHR AH)
Collected data set	- Hospital Care Diagnoses (AH_Diagnoses) - Hospital Care Procedures (AH_Procedures) - Hospital Care Episodes (AH_Episodes) - Hospital Care External Cause (AH_External_Cause)

Table 3. *Description of the types of documents used by Mica and their application in the master's final project.*

Once everything was created, each collected dataset was linked to the data tables previously organized in Opal. This linking allows the metadata of the GCAT study in Mica to be directly connected to the real metadata stored in Opal. After that, the metadata of the GCAT study was published in Mica, making the descriptions accessible to other researchers.

The ultimate goal of the platform is that users (i.e. researchers) can search and browse the different studies, data collection events, datasets, and variables described in Mica. For those with the necessary permissions, the data stored in Opal can be accessed, but this is covered in future sections.

4.3. Data Standardization

Data standardization is an essential process to ensure the harmonization (i.e. compatibility and consistency) of data, especially when working with data sets from a variety of sources. This is done by converting the different sources to a common data model. Among the different healthcare standards facilitating interoperability, this project specifically emphasizes the OMOP standard.

OMOP aims to establish a common standard to organize and store health data, allowing the integration of different sources and the comparison between them. It is an open-license standard that seeks to standardize and structure data. More concretely, the OMOP CDM is a person-centric model, that establishes a common vocabulary and data format, allowing for seamless integration across disparate heterogeneous sources.

The model is designed to encompass all relevant observational health data to enable research, adhering to the following principles [70]:

- **Suitability for purpose**. The common data model organizes the data in an optimal way for analysis.
- **Data protection**. Some of the data that could compromise the identity of the patient, such as name, precise date of birth, direction, etc. is limited. Exceptions can be made when the research requires such information.
- **Design of domains**. Domains are structured on a person-centered relational data model. Each record captures at least the person's identity and a date. It is important to mention that data is represented in tables that are interconnected by primary keys and foreign keys.
- **Rationale for domains**. Domains are created and defined separately in an entity-relationship model if they have a specific use for analysis and unique attributes. Other data can be stored as observations in an entity-attribute-value structure.
- **Standardized vocabularies**. The common data model relies on standardized vocabularies, which contain all the necessary standardized health concepts. More concretely, it employs widely recognized terminologies such as Systematized Nomenclature of Medicine – Clinical Terms (SNOMED-CT) for clinical conditions, Standardized Nomenclature of Medicine Clinical Terms (RxNorm) for medications, Logical Observation Identifiers Names and Codes (LOINC) for laboratory tests, and HL7/LOINC for clinical notes. These vocabularies provide a common language that facilitates the accurate exchange of health information and are available at Athena.
- **Reuse of existing vocabularies**. Many of the standardized vocabularies are leveraged from organizations or initiatives.
- **Maintaining source codes**. Although all codes are assigned to standardized vocabularies, the model also allows the original source code to be stored, ensuring that no information is lost.

- **Technology neutrality**. It does not require a specific technology.
- **Scalability**. It can accommodate data sources that vary in size.
- **Backwards compatibility**. All changes from the previous common data model versions are clearly delineated in the OHDSI GitHub repository.

In this project, the OMOP CDM version 5.4 is intended to be used as a standardized framework for data representation and analysis. It consists of 39 tables as can be seen in Figure 10.

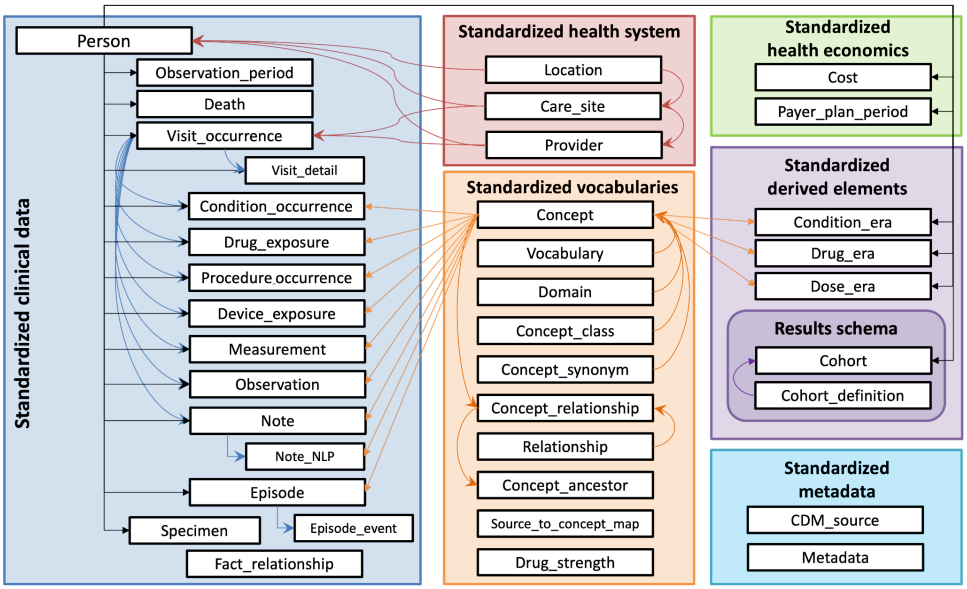


Figure 10. *Overview of all tables in the OMOP CDM v5.4. Source: https://ohdsi.github.io/CommonDataModel/index.html* [70]

The OMOP CDM groups the different tables in sections, as shown in Figure 10. **Standardized clinical data** group contains the different OMOP CDM tables related to clinical data. The purpose of each table within the model is presented below:

- **Person** - is the main table of the CDM. It serves as the central identity management for all the patients / persons in the database, i.e. all persons in the database need one record in this Person table. This table contains basic demographic information for each individual, such as the unique identifier, date of birth, sex, etc.
- **Observation period** - defines the time intervals over which information is available for each person.
- **Death** - contains the clinical event of how and when a person in the database dies.
- **Visit occurrence** - contains all events where patients interact with the health system, either to represent visits or other types of interaction (e.g. call). It contains information on the start and end date of the visit, the type of visit (e.g. hospitalization), etc. It always includes the patient identifier and a primary key to be used as a link to other tables.
- **Visit detail** - is an optional table used to represent details about a *visit_occurrence*, i.e. for each record in the *visit_occurrence* table there can be 0 or more records in *visit_detail*. This is useful to detail movements between units in a hospital.

- **Condition occurrence -** documents the presence of a disease or medical condition stated as a diagnosis, a sign or a symptom, observed during a visit. It is important to specify the date of diagnosis and the type of condition. The diagnosis is collected in the *condition_concept_id* field using SNOMED as standard vocabulary.

- **Drug exposure -** records events of exposure to drugs or other pharmacological substances during a patient's medical care. The concept of the drug or the substance is collected in the *drug_concept_id* field using RxNorm as standard vocabulary.

- **Procedure occurrence -** contains all medical procedures performed on patients, i.e. activities or processes ordered by / carried out by a healthcare provider.

- **Device exposure -** contains information about exposure to a foreign physical object (e.g. pacemakers, stents, etc.) or instrument (e.g. defibrillators) used for diagnostic or therapeutic purposes.

- **Measurement -** stores both orders and results of clinical measurements, such as laboratory tests. It is important to mention that measurements are recorded as attribute-value pairs, where the attribute is the concept measured and the value is the result. The *measurement_concept_id* is encoded using LOINC as standardized vocabulary.

- **Observation** - captures other clinical information that do not fit into other domains, such as patient-reported symptoms, family history, lifestyle data, etc.

- **Note -** is used to store unstructured information, such as notes or comments related to healthcare events. The *note_concept_id* is encoded using HL7 / LOINC CDO as standardized vocabulary.

- **Note NLP** - store the results of processing clinical notes, by using natural language processing (NLP) techniques.

- **Episode -** captures groups of related clinical events in a single episode of care, such as a complete cancer treatment or surgery.

- **Episode event -** documents the individual events within an Episode.

- **Specimen -** contains the records that identify a person's biological samples, such as blood samples. This table includes details on the type of sample, the date of collection, and the location where the sample was stored.

- **Fact relationship -** contains records about the relationships between facts stored as records in any table, such as indication relationship (i.e. relation between drug exposure and an associated condition).

Another integral component of the OMOP CDM is the use of standardized vocabularies to ensure consistency and interoperability across datasets. Therefore, there is the group of **standardized vocabularies** tables, detailed below:

- **Concept** - contains the unique identifiers for each medical concept, including descriptions and attributes. Each entry represents a single concept within the standardized vocabulary.

- **Vocabulary** - stores information about the different vocabularies used in the OMOP CDM, such as SNOMED, RxNorm, and LOINC. It helps in identifying which vocabulary a concept belongs to.

- **Domain** - categorizes concepts into different areas of healthcare, such as conditions, drugs, and procedures. This classification facilitates organized data storage and retrieval.

- **Concept class** - provides further classification within a domain, specifying the type or class of a concept (e.g. lab test, diagnosis). It helps in detailed categorization and analysis of concepts.

- **Concept synonym** - contains synonyms for each concept, ensuring that different terms referring to the same concept can be identified and mapped accurately.

- **Concept relationship** - defines relationships between concepts, such as hierarchical (e.g. parent-child) or associative relationships, helping in understanding the connections between different medical terms.

- **Relationship** - stores the types of relationships that can exist between concepts, providing a structured way to define how concepts are related.

- **Concept ancestor** - captures hierarchical relationships, detailing ancestor-descendant links between concepts. It is useful for queries that need to traverse these hierarchies.

- **Source to concept map** - maps source codes from external datasets to the standardized OMOP concepts, facilitating data integration and consistency.

- **Drug strength** - includes detailed information about the strength of drug concepts, such as dosage and units, ensuring precise representation of medication data.

Building an OMOP CDM database requires Extract - Transform - Load (ETL) or Extract - Load - Transform (ELT) processes that facilitate the data conversion to a standard model and terminology. In our project, we opted two different approaches: (i) a traditional ELT, to ensure data consistency and integrity through the integration process, enabling a smooth transition from heterogeneous data sets to a coherent and standardized structure; and (ii) a semantic ETL, to preserve the meaning and context of the information.

4.3.1. Traditional ELT

The traditional framework follows the OHDSI guidelines, since many of the tools that are used in the process were developed within the community. It encompasses a series of steps detailed below.

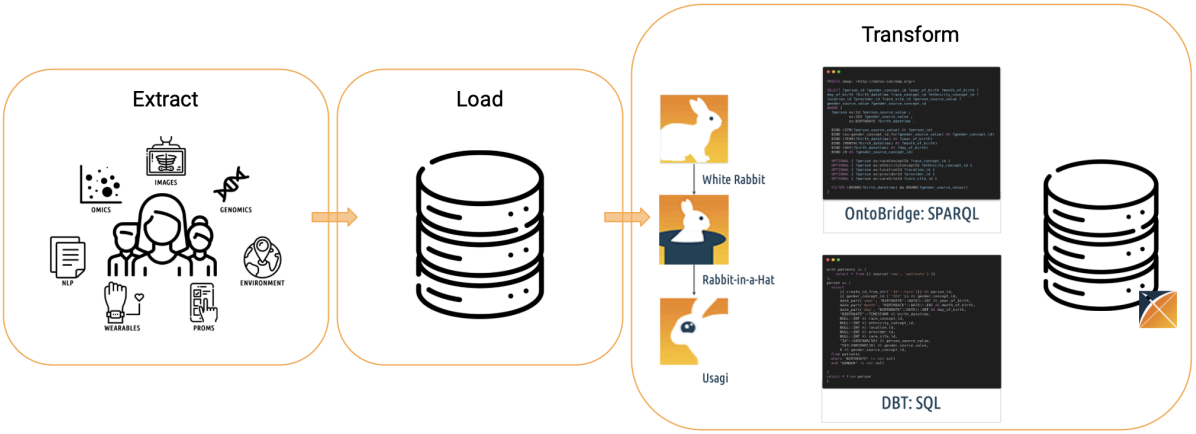


Figure 11. *Overview of the traditional ELT approach.*

Extract phase.

The first step involves identifying and extracting the data from the different data sources. In the DATOS-CAT project we have different sources and different types of data, which have been better identified thanks to the cataloging of the data, as we have seen in the previous section. We could say that the data come from two main sources: (i) research-collected data gathered both in the baseline and in the different follow-ups, and (ii) EHR obtained from PADRIS. In this final master's thesis, for reasons of data privacy, we generated synthetic datasets using the data profiling and characterization of the real datasets. This is explained in more detail in section 4.6.

Once we identified the data, we used Meltano for data ingestion. Meltano [71] is an open-source data movement tool that allows the creation of end-to-end data pipelines, including ELT process. Then, we created a Meltano project, and we added plugins to it. More concretely, we added an extractor called *tap-csv* to pull the data from the source (i.e. extract the data of CSV files).

Load phase.

Load phase enables the loading of raw data into a target data warehouse, in our case, a PostgreSQL. PostgreSQL [72] is an open-source object-relational database system that uses the Structured Query Language (SQL) standard language, and allows to organize the information in one or more schemas. For the DATOS-CAT project, we need at least four schemas: (i) ***raw*** schema, where the original information is stored; (ii) ***cdm*** schema, where the OMOP-transformed data is stored; (iii) ***vocabularies*** schema, where terminologies, standardized codes and their relationships are stored; and (iv) ***results*** schema, where the output of the Data Quality Dashboard is stored.

Therefore, prior to the load phase, we set up a PostgreSQL server using Docker, and created a database called *datoscat_tfm*. Then, we used the *psql* command, which is a terminal-based front-end to PostgreSQL, to connect to the database and create different schemas: (i) ***raw*** schema that will hold the source data, (ii) ***cdm*** schema that will hold de clinical tables from the OMOP-CDM model, (iii) ***vocabularies*** schemas to hold the vocabulary, and (iv) ***results*** schema.

Once the schemas were created, we needed to define and manage the structure of objects within the *cdm* and *vocabularies* tables. This process is known as SQL Data Definition Language (DDL). To do so, we cloned the OHDSI CommonDataModel repository from GitHub [73], which provides the DDL SQL queries to create both the *cdm* and *vocabularies* tables. The successful execution of this function ensures that the database structure aligns with the OMOP CDM specifications, laying the foundation for subsequent data operations.

Following the creation of the database schema, the next critical step is to load the vocabulary data into the *vocabularies* schema. The vocabulary data must be downloaded beforehand from Athena [74], the OHDSI tool for vocabulary management. Steps to download the vocabularies can be found in Appendix C. Athena provides a series of csv files containing the necessary vocabulary data for the OMOP CDM. These files are then imported into the *vocabularies* schema. With this process we ensured that all standard terminologies, required for the OMOP CDM, were correctly integrated into the database. Consequently, the database is equipped for standardized and consistent representation of the data, facilitating effective data analysis and research across various datasets.

After that, we added the second plugin to the Meltano project, a loader called *target-postgres*. The loader is responsible for uploading the raw data to the PostgreSQL database in the raw schema.

Transform phase.

In this last step, the different data formats, terminologies and structures are standardized and harmonized into a uniform format. The Transform phase includes a semantic and a syntactic mapping.

Within the semantic mapping phase, the different source-specific codes (e.g. ICD-10) are mapped to standard terminologies according to the domain (e.g. SNOMED for conditions, RxNorm for drugs, etc.). To facilitate the process, we can take advantage of two very useful tools of the OHDSI suite [75] : (i) Athena, which allows the search and load of standardized vocabularies, and (ii) USAGI, a tool that simplifies the code mapping process by suggesting mappings based on textual similarity of code descriptions. It also allows the user to search for more appropriate concepts if the suggestion is not correct.

On the other hand, within the syntactic mapping, the different data fields and values are mapped into the corresponding tables and fields of the OMOP CDM. For this purpose, we used White Rabbit and Rabbit-in-a-Hat [76], both OHDSI tools. First, with White Rabbit, we generated a scan report with detailed information about the raw data (i.e. tables, fields and values). Then, Rabbit-In-a-Hat tool, which is designed to read the White Rabbit's scan report, displayed the source data information through a graphical interface, allowing us to connect the source data to the tables and fields of the CDM. In this way, we obtained the alignment of the source data structure with the CDM framework.

Once both mappings were clear, we used DBT (Data Build Tool) [77] for the subsequent data transformation. Inside the Meltano project, we added a transformer called *dbt-postgres* as another plugging. The transformers allow the creation of new derived transformations from raw data sources, and, particularly, the *dbt-postgres* runs SQL-based transformations on data stored in the warehouse. Therefore, we created DBT models inside the Meltano project.

First, we created a file called *sources.yml* that specifies which sources of data will be used during the transformation. Here, we specified both the *raw* and the *vocabularies* tables. Then, we created the SQL-models in two different folders: (i) *cdm* folder to locate the OMOP CDM models, and (ii) *staging* folder to store intermediate models necessary for the transformation process. A SQL-model is defined as a *select* statement that applies SQL logic to transform data. For example, the SQL person model takes the raw demographic data and transforms it into records of the person table in OMOP CDM. It is important to mention that the model's name is inherited from the filename. After the creation of these models, we executed the DBT run command to execute the models and create the resulting tables with transformed data in the *datoscat_tfm* PostgreSQL.

4.3.2. Semantic ETL

On the other hand, there is a semantic approach to ETL technologies that focuses on the semantics of data (i.e. its meaning and context, as well as its structure and format).

Semantic ETL [78] involves extracting, transforming, and loading data in a way that preserves the meaning and context of the information. This approach goes beyond simple data format conversion, it ensures that the data's semantic integrity is maintained. By using standardized frameworks and models like Detailed Clinical Models (DCMs) and ISO13606, semantic ETL formalizes data operations to make them understandable, reproducible, and auditable. This method enables the creation of reusable datasets that are rich in context and meaning, essential for secondary uses such as clinical research and public health studies.

This approach allows standardized data together with their semantic meaning and syntactic structure, a key element for scalable and interoperable secondary data studies. To achieve semantic normalization, Hospital Clinic de Barcelona developed a tool called *OntoBridge* [79].

OntoBridge is a database normalization tool, based on ontologies, that uses semantic mappings to transform local databases into common data models standards such as OMOP CDM, in a secure, fast, scalable manner. There are different components necessary for the OntoBridge pipeline, the general idea is explained below.

- **Local model ontology** - is an ontology that models the structure of the raw data. It represents each table of the dataset as a class, and each variable as a property.
- **Clinical records ontology** - is an ontology that models the local codes used. These codes are represented as instances within the local model ontology.
- **Standard model ontology** - is an ontology representing the common data model standard, such as OMOP CDM. Here, the different OMOP tables and properties are defined.
- **Standard dictionary ontology** - is an ontology that contains the different standard codes that will be used.
- **Mapping ontology** - is an ontology that contains both the syntactic and semantic mappings between the raw data to the OMOP CDM. In general terms, the semantic mappings are represented by 'owl:sameAs' triplets, while in the syntactic mapping, the local properties (i.e. the database columns) are mapped to properties of the CDM structure, by using meta-classes.

OntoBridge uses an open-source tool called *Ontop* [80] to extract the data in the original data source (databases, spreadsheets, etc.) into a semantic structure in Resource Description Framework (RDF) [81] format, using a mapping language called R2RML [82] (RDB to RDF Mapping Language). Then those RDF elements are inserted into the local data model ontology.

After that, these ontologies are loaded into a Jena Fuseki [83] server, which is the semantic equivalent to a relational database, and the corresponding SPARQL [84] queries are executed to generate the OMOP CDM tables. SPARQL is a semantic query language, equivalent to SQL for relational databases. The general overview of the semantic pipeline can be seen in Figure 12.

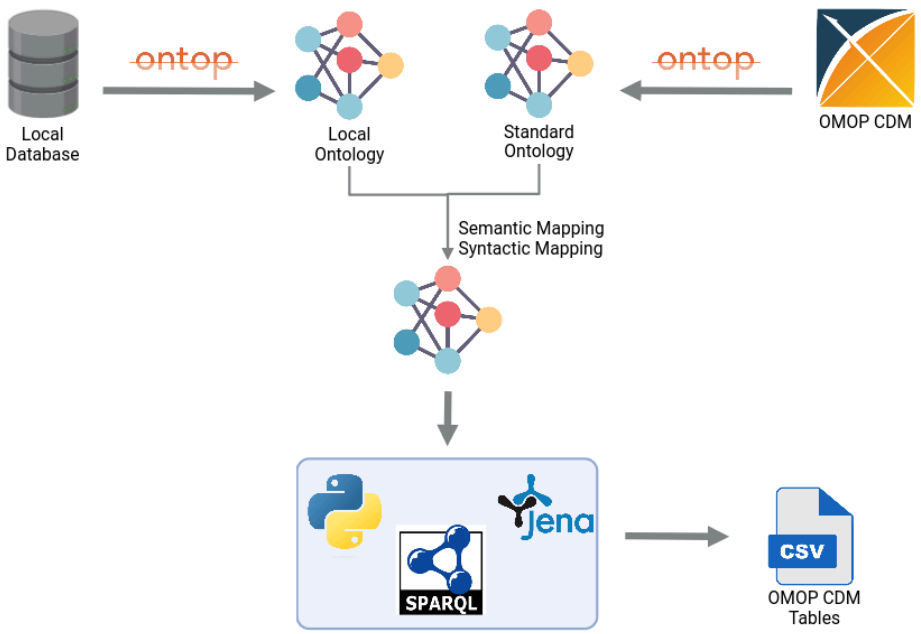


Figure 12. *Semantic ETL overview, using OntoBridge as the main tool of the process.*

For ease of use, Hospital Clinic de Barcelona has created a set of docker images that facilitate the deployment and configuration of the different parts that compose OntoBridge. Also, as part of this Master's thesis project, an automated local ontology was designed using Mica data dictionary as a basis. To do so, we generated a Python script that automates the generation of the local model ontology.

Then, as a proof-of-concept of the semantic approach we generated the OMOP CDM person table, showing that both approaches are equivalent.

4.3.3. ELT evaluation and validation

Once the traditional ELT process was completed, it was essential to assess data quality (DQ) to ensure the integrity and usability of the data. Data Quality is defined as "*the state of completeness, validity, consistency, timeliness and accuracy that makes data appropriate for a specific use*" (Roebuck, K. 2012) [85]. DQ assessment consists of three main components:

- **Conformance** - ensures that the data adheres to predefined standards and formats.
- **Completeness** - checks if all necessary data is present.
- **Plausibility** - verifies that the data values are realistic and logical.

To evaluate data quality in the context of the OHDSI framework, we used the **Data Quality Dashboard** (DQD) [86] tool. The DQD is used to systematically evaluate data quality across multiple dimensions within the OMOP CDM. It performs a series of automated checks to assess data conformance, completeness, and plausibility. The results are presented in an interactive dashboard, facilitating easy identification and remediation of data quality issues. DQD provides a robust framework for continuous data quality monitoring, enhancing the reliability and validity of the data for research purposes.

To assess quality, we first prepared the R environment by ensuring that all the required dependencies were installed. During the ELT process we added a model called *cdm_source*, an essential OMOP

CDM table that provides metadata about the dataset, including the version of the CDM, the date of the ELT process, etc.

After preparing the environment, we executed the Data Quality Dashboard, using a R script provided in the official OHDSI GitHub repository. We modified the script parameters to connect it to our specific database. This step involved running a series of predefined quality checks to validate data conformance, completeness, and plausibility.

4.4. Data Analysis

In the current era of information, data analysis has become a fundamental pillar for decision-making in a variety of fields, especially in medical research. However, accessing and sharing medical data still remains a challenge in terms of privacy and security, as it is extremely sensitive and subject to strict privacy regulations, such as HIPAA in the United States or GDPR in Europe.

Data federation solves the problem of data access. It emerges as an innovative solution to enable collaboration and data analysis between different institutions, without compromising the privacy of the patient. This approach allows researchers and healthcare professionals to perform statistical analysis of data sets distributed across different hospitals, research centers and organizations, without the need to share the actual data, as the data remains within appropriate jurisdictional boundaries. A prominent solution for conducting federated data analysis in a secure manner is DataSHIELD.

DataSHIELD allows federated analysis by enabling the analysis of sensitive data without requiring that data to be physically shared or moved. It employs a strategy where analysis code is sent to the data locations, and only the non-disclosive summary statistics are returned to the researcher. This method ensures compliance with privacy regulations and maintains the confidentiality of the data [87].

There are several packages available in DataSHIELD, of which we would like to highlight dsOMOP. **DsOMOP** [88] is a specialized package developed by ISGlobal-BRGE to facilitate the interaction with remote databases formatted in the OMOP CDM within a DataSHIELD environment. The dsOMOP ecosystem consists of two main components: (i) dsOMOP, which is the server-side package and (ii) dsOMOPClient [89, 90], which is the client-side package. In general terms, the server-side package handles the retrieval and transformation of data from OMOP CDM databases, ensuring that all operations comply with DataSHIELD's security model. Meanwhile the client-side package allows researchers to communicate with the server-side package, sending data requests and receiving processed data for analysis. This integration significantly enhances the ability to perform secure and compliant data analyses on OMOP CDM databases

In this Master's thesis, we wanted to test the federated analysis of the standardized synthetic data, using the dsOMOP package. Therefore, we first installed the server-side package in Opal. Then, within the Opal DATOS-CAT project, mentioned in the previous section, we added the PostgreSQL database as a resource. For more details on dsOMOP installation, see Appendix D.

4.5. Data Discovery

Data discovery is a critical process in biomedical research, enabling researchers to locate, access, and use relevant datasets efficiently. It involves identifying and understanding data sources, which can

significantly enhance research capabilities by providing new insights and facilitating data-driven decisions. In the context of the DATOS-CAT project, data discovery is essential for leveraging the extensive data collected from various cohorts and studies.

One of the key tools for data discovery is **Beacon v2** [91] , a protocol developed by the GA4GH. Beacon is a data discovery tool designed to enable researchers to identify datasets that contain specific genomic variants or pheno-clinic data. With this tool researchers are able to query multiple datasets to determine the presence of a specific record (e.g. query for specific allele), and Beacon will return whether any datasets within the network contain the queried record, without disclosing sensitive information.

It operates under strict privacy and security protocols, ensuring that no individual-level data is revealed, complying with data protection regulations like GDPR. There are several benefits associates, such as:

- Beacon streamlines the process of finding relevant datasets, saving researchers time and resources.
- It enables researchers to discover datasets across different institutions and countries. Therefore, it fosters collaboration and data sharing.
- It aligns with the principles of open science by making data more accessible and discoverable, thus advancing scientific research and innovation.

Beacon will play a pivotal role in the data discovery phase of the DATOS-CAT project by facilitating the identification of relevant genomic datasets. This capability is crucial for integrating and analyzing data across different cohorts, ultimately enhancing the quality and impact of biomedical research.

Moreover, leveraging the existing Python-based Beacon v2 reference implementation developed at CRG, from the BSC we created corresponding SQL queries to model the different Beacon endpoints to the expected results using aiosql library. This allowed us to transform the relational data into the Beacon v2 JSON specification [92, 93].

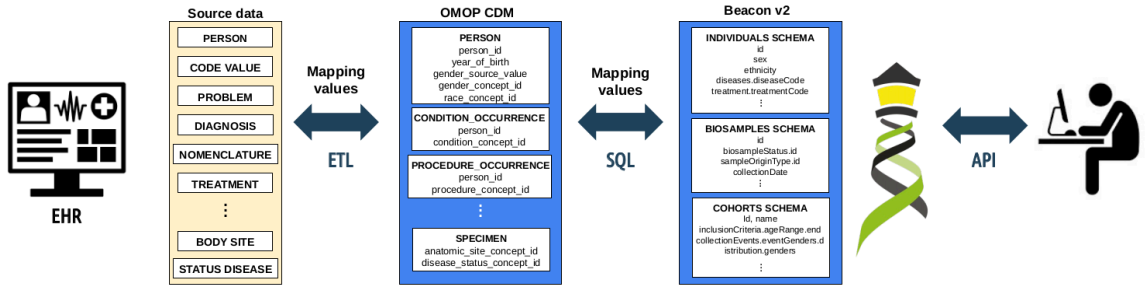


Figure 13. *Beacon v2 communication overview. Source: https://f1000research.com/posters/12-696*

Then, we exposed the whole platform, using the aiohttp Python library [94], as an API that can be queried or integrated into an existing Beacon network. By doing so, we were able to seamlessly integrate OMOP CDM databases into the Beacon v2 ecosystem, enabling researchers to discover and query genomic and clinical data in a secure and privacy-preserving manner.

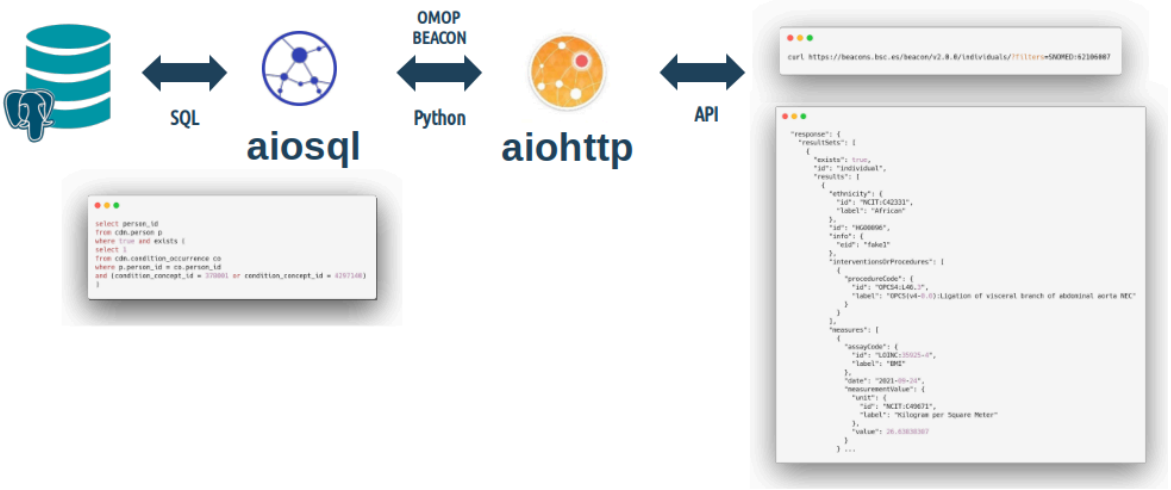


Figure 14. *Beacon v2 communication overview. Source: https://f1000research.com/posters/12-696*

Once a REST request is received, the corresponding model (individuals, cohorts, etc) is invoked with the query parameters, data is then retrieved from the database and converted to the Beacon v2 specification JSON response which is sent back to the client.

In the framework of this Master's thesis project, we decided to use the Beacon OMOP CDM on our synthetic *datoscat_tfm* database, to enable federated data discovery. First, we cloned the *impd-beacon_omopcdm* Gitlab repository and we set up the environment to run the script, specifying the parameters to connect with our PostgreSQL database.

4.6. Synthetic data

The DATOS-CAT project deals with data from the GCAT cohort and the COVICAT-CONTENT sub-cohort, which are considered personal data. Aware of the sensitivity and confidentiality involved in using this real data, we have decided to generate synthetic data to work on this Master project.

This decision allowed us to accurately simulate the medical information needed for the research, while protecting the privacy and security of the participants. The generation of synthetic data became an ethical and effective strategy to advance in the research project without compromising the confidentiality of individuals. At the same time, this decision represents a central point in the project. In a context where synthetic data generation is becoming an increasingly common option, its implementation emerged as a significant challenge.

Initially, our research group (BSC) had a software designed to generate synthetic data. This software uses events and transitions based on probabilities derived from the literature, allowing the creation of synthetic data sets that closely replicate the characteristics and patterns recognized in the literature and, therefore, in real life. However, due to access restrictions and to resemble the original cohort data more closely, an adaptation of the software to rely on another approach was required.

For the creation of synthetic data, detailed information about each variable is needed, such as its name, type of data, number of null values, and the frequency of each type of data, among others.

Since we cannot access the data directly, we need a tool that provides this information in an aggregated form. Aggregated information allows us to understand the general structure and characteristics of the database without revealing specific and sensitive data of individual records. To obtain this, we have considered two options:

- **White Rabbit Scan Report** - provides a detailed summary of a database/data set, giving information about the content and structure.

- **Python Profiling** - *'ydata-profiling'* is a Python library [95] that provides an efficient way to perform Exploratory Data Analysis (EDA). This tool generates a profile report of a pandas DataFrame, which includes descriptive statistics, information about the distribution of the data, relationships between variables, etc. It can be exported in various formats, including JSON, which allows the detailed information generated to be stored in a structured way, facilitating further processing and analysis.

Once we have the aggregated information file, we can generate synthetic data based on this information. For this purpose, we have developed a Python script called *reader.py* as part of the methodology, which is integrated into the software mentioned above.

This script was designed with the specific purpose of accepting as input aggregated data and transforming it into the events-transitions format required by the generator (i.e. convert the White Rabbit Scan Report or the Python Profiling to the structure required by the generator, the scheme can be seen below).

```
{
    'events': [
        {
            'name': 'profiling_EGA_id_AH_Procediments',
            'description': PosixPath('...'),
            'type': 'profiling_EGA_id_AH_Procediments',
            'values': [
                {
                    'name': 'EGA_id',
                    'type': 'list',
                    'params': [
                        {'val': 'EGA_00001', 'prob': 0.00013132666193890684},
                        {'val': 'EGA_00006', 'prob': 7.87959971633441e-05},
                      ...
                    ],
                }
            ],
        }
    ],
    'transitions':[
        {'a': 'demographic_report', 'b': 'profiling_EGA_id_AH_Episodis', 'p': 1},
        {'a': 'profiling_EGA_id_AH_Episodis','b': 'profiling_EGA_id_AH_Diagnostics','p': 1}
        ]
}
```

Figure 15. *Overview of the structure generated to produce synthetic data. Source: Figure created with Carbon.*

To simplify the process, all the variables were considered categorical with a certain frequency, obtained from the aggregated information (i.e. number of times this variable appears divided by the total). Therefore, all the variables are treated as independent variables, without considering any relation between them. This simplification constitutes a significant limitation, since many variables

are descriptions of others or are related by causality (for example, the probability of diabetes is greater if the individual has obesity). It may affect the precision of the posterior analysis, but it is useful to try all the data flow pipeline.

The synthetic data was generated based on the care hospital's EHRs, using the following data sets as a reference: (i) Diagnoses, (ii) Procedures, (iii) Episodes and (iv) External Cause. The OMOP CDM requires certain mandatory information about the individual, including year of birth and the biological sex at birth of the person. Since this information was not present in any of the available data sets, we generated a new data set that considered all identifiers from the hospital care data sets. Then, we established the age by assigning a random number between 40 and 65 (from which we will calculate the year of birth), and sex was determined by assigning a probability of 50% to be a male and 50% to be a female.

Once the demographic data set was created, we generated the Python Profiling from it. In this way, we were able to create the synthetic data for all the data sets (i.e. patient, diagnoses, procedures, episodes and external cause), establishing a transition from one event (i.e. data set) to another.

Chapter 5. Experiments and results

5.1. Data Catalog

5.1.1. Data Catalog Template

For data cataloging, Opal offered a structured template that we used for the creation of data dictionaries, consisting of two main tabs: (i) variables tab, offering a comprehensive description of the variable attributes contained in a data set; and (ii) categories tab, offering a detailed explanation of the categorical variables present in a data set, as well as information about missing or outlier variable attributes.

To ensure consistency between different datasets, we established internally clear guidelines on how to complete these dictionaries in a uniform way. The information contained in each of the tabs is detailed below in Table 4 and Table 5.

Columns	Description	Example
Table	Name of the data set file	COVICAT_serology_2020
Name	Name of the variable	IgM_NFL
ValueType	Type of variable	Text
Unit	Units of the variable (including none)	None
Annotations	Internal variable annotations	Seropositivity variables
Label:en	Description of the variable in English	IgM N-FL (full length nucleocapsid protein)
Label:es	Description of the variable in Spanish	IgM N-FL (proteína de nucleocápside de longitud completa)

Table 4. *Description of the structure used to create the "Variables" tab of the Mica data catalog. Each row represents a specific aspect of the variable, with its respective description and an example to illustrate its content.*

Columns	Description	Example
Table	Name of the data set file	COVICAT_serology_2020
Variable	Name of the variable to which the category belongs	IgM_NFL
Name	Name of the category or "Missing" category	Negative
Missing	1 for missing, and 0 for not	0
Label:en	Description of the category in English	Negative IgM N-FL
Label:es	Description of the category in Spanish	IgM N-FL Negativo

Table 5. *Description of the structure used to create the "Categories" tab of the Mica data catalog. Each row represents a specific entry within a categorical variable, with its respective description and an example to illustrate its content.*

The design of the structure and the diligence in filling in the fields on both tabs ensures clarity, consistency, and completeness of the metadata and, at the same time, helps the overall comprehension of the data collected.

5.1.2. Data Catalog Automation

In this section, we present an example of the data dictionary used to create the data catalog. Each file, as mentioned before, contains two main parts: (i) a variable descriptive tab and (ii) a category descriptive tab, both generated using an automated script that streamlines the process and guarantees consistency between documents by minimizing human error. The Python script takes the Scan Report, previously created with White Rabbit, as input to generate an Excel file with two different tabs.

An example of a Mica data catalog created with the script that automates this process, using the hospital care diagnostics dataset called *AH_Diagnostics*, is presented below.

Table	Variable	ValueType	unit	Label:en
AH_Diagnostics	EGA_id	text	none	EGA ID
AH_Diagnostics	id	text	none	Original cohort ID
AH_Diagnostics	DATA_INGRES	date	none	Date of entry
AH_Diagnostics	EDAT_INGRES	integer	none	Age at entry
AH_Diagnostics	pos_dx	integer	none	Position of diagnosis
AH_Diagnostics	ICD9	text	none	ICD-9
AH_Diagnostics	ICD9_desc	text	none	ICD-9 description
AH_Diagnostics	ICD10	text	none	ICD-10
AH_Diagnostics	ICD10_desc	text	none	ICD-10 description

Table 6. *Example of the "Variables" tab of the Mica data catalog for the hospital care diagnostics data set.*

Table	Variable	Name	Missing	Label:en
AH_Diagnostics	EGA_id	Missing	1	Missing
AH_Diagnostics	DATA_INGRES	Missing	1	Missing
AH_Diagnostics	EDAT_INGRES	Missing	1	Missing
AH_Diagnostics	pos_dx	1	0	1
AH_Diagnostics	pos_dx	2	0	2
		…		
AH_Diagnostics	pos_dx	15	0	15
AH_Diagnostics	ICD9	Missing	1	Missing

AH_Diagnostics	ICD9_desc	Missing	1	Missing
AH_Diagnostics	ICD10	Missing	1	Missing
AH_Diagnostics	ICD10_desc	Missing	1	Missing

Table 7. *Example of the "Categories" tab of the Mica data catalog for the hospital care diagnostics data set.*

5.1.3. Data Catalog Deployment

In this section, we present the results obtained from deploying a data catalog using Mica (OBiBa software). The results include a detailed overview of the project environment in Opal, the configuration steps taken in Mica, and the main features of the Mica web data portal. These results highlight the efficient organization and management of datasets, demonstrating the practical application of Mica for data integration and exploration.

Opal Project Overview

Figure 16 illustrates the project created in Opal and the uploaded data dictionaries as tables. This provides a clear visualization of the work environment and data organization, highlighting how data sets are structured and managed within the system.

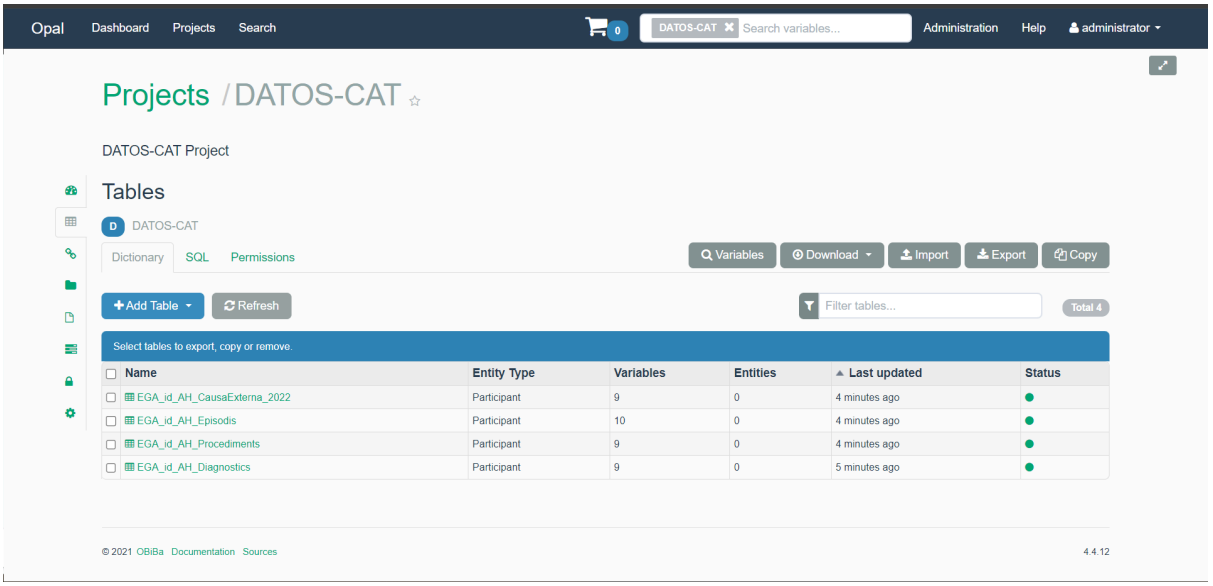


Figure 16. *Screenshot displaying the DATOS-CAT project overview in Opal along with the uploaded data dictionaries as tables.*

The Opal environment allows an efficient organization of data dictionaries into tables, facilitating data management and access. By structuring data in a tabular format, it becomes easier to maintain consistency and ensure data integrity across various datasets. This setup forms the foundation for a robust data catalog system, enabling seamless data integration.

Mica Configuration

The following section presents the Mica configuration, demonstrating the setup and relationships established within the system.

- **GCAT Network creation.** Figure 17 captures the draft view after the creation of the GCAT Network in Mica. The network creation is crucial to establish connections between different

studies and their respective datasets. More concretely, the GCAT Network serves as a central node, linking GCAT study with COVICAT-CONTENT study and their datasets.

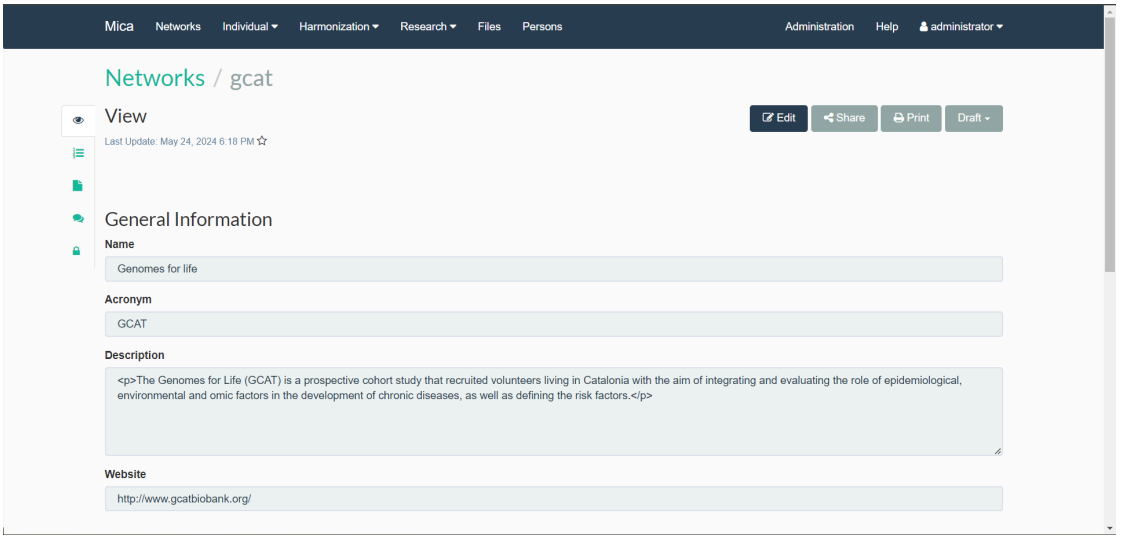


Figure 17. *Screenshot capturing the draft view after the creation of the GCAT Network in Mica.*

- **GCAT Study creation.** Figure 18 shows a draft view after the creation of the GCAT Study in Mica. The study creation process involves specifying its objectives, methodologies, and key variables, which are essential for conducting meaningful analyses. The GCAT Study provides a structured framework for collecting and organizing data, ensuring that all relevant information is systematically captured and accessible.
  In the DATOS-CAT project, another study reflecting the COVICAT-CONTENT sub-cohort should be created.

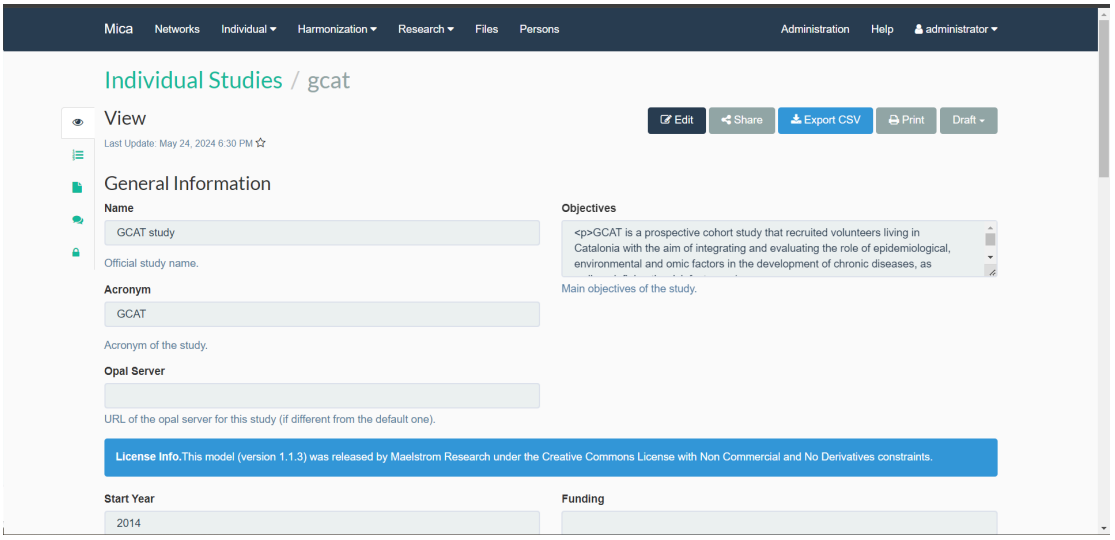


Figure 18. *Screenshot capturing the draft view after the creation of the GCAT Study in Mica.*

- **Collected Datasets List**. Figure 19 shows the list of all the collected datasets created in Mica. Each dataset is cataloged systematically by specifying different criteria, such as study name, population group, data collection event, etc. This organized approach enhances data discovery and supports efficient data management practices.

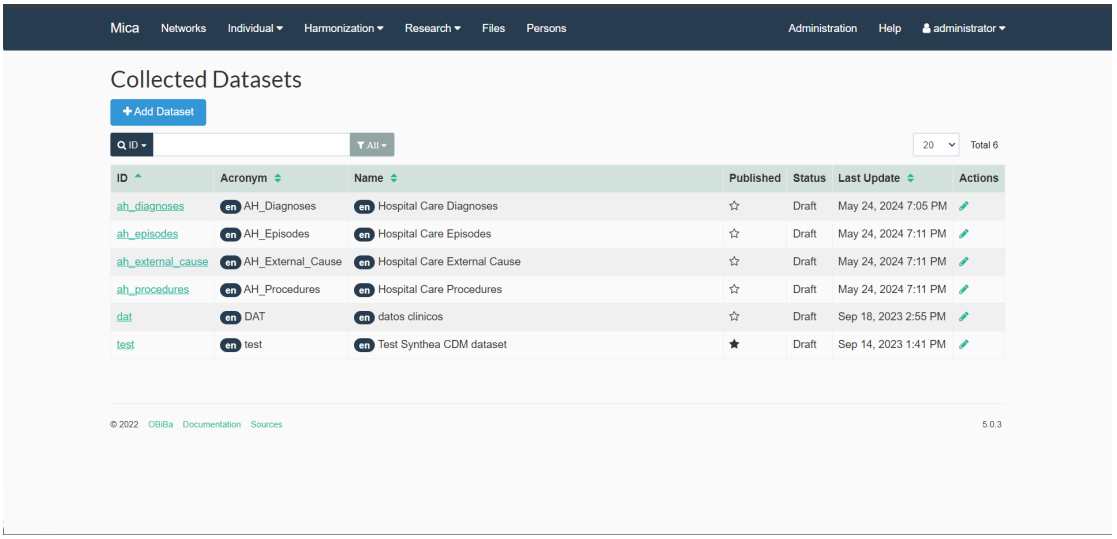


Figure 19. *Screenshot showing the list of all the Collected Datasets created in Mica.*

The administration view of the *AH_Diagnosis* collected dataset can be seen in Figure 20, where the relationships between the different components are visible. The dataset is linked to a study (GCAT study), a population (ID = 1), a data collection event (ID=01), an Opal project (DATOS-CAT Opal project), and an Opal table (*EGA_id_AH_Diagnostics*). This interconnected setup ensures comprehensive data integration and traceability.

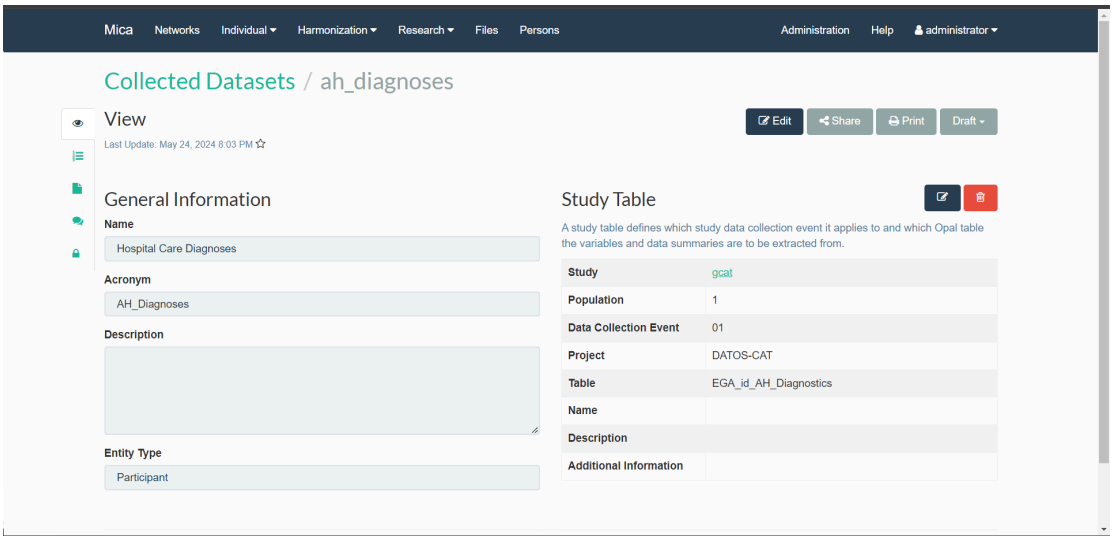


Figure 20. *Screenshot showing the draft view of the AH_Diagnosis collected dataset in Mica. The relationship that has been made with different components can be seen in the right part. AH_Diagnoses is linked with a Study (i.e. GCAT study), with a Population (i.e. ID = 1), with a Data collection Event (i.e. ID=01) and also with an Opal project (i.e. DATOS-CAT Opal project) and an Opal table (e.g. EGA_id_AH_Diagnostics).*

The relationship overview of the other collected data sets created, can be seen in Table 8.

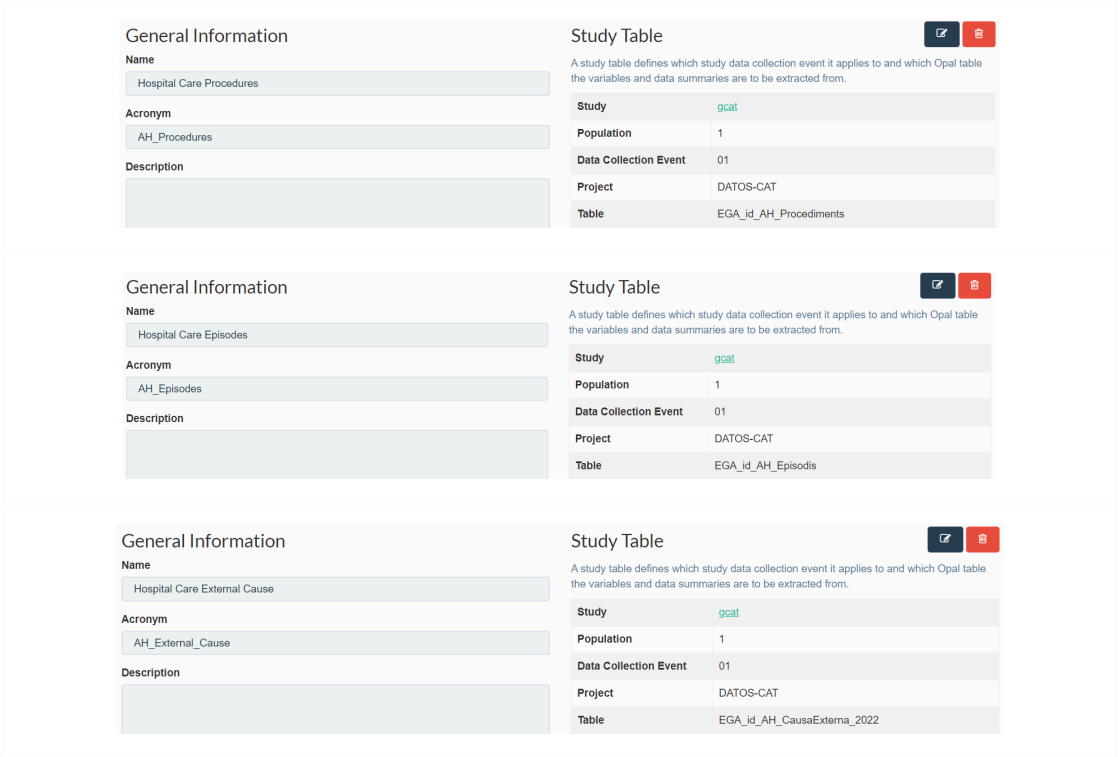


Table 8. *This table provides three screenshots showing the draft view of the other Collected Dataset in Mica. These views illustrate the relationship established with various components, ensuring a robust data linkage framework.*

Mica web data portal

The Mica web data portal offers an intuitive and user-friendly interface, allowing researchers to navigate through various sections and explore the data easily.

The main dashboard of the web data portal provides an overview of the available datasets, studies, and variables, making it simple for users to find the information they need. The dashboard is designed to be an accessible entry point for exploring the data, with clear navigation options and search functionalities.

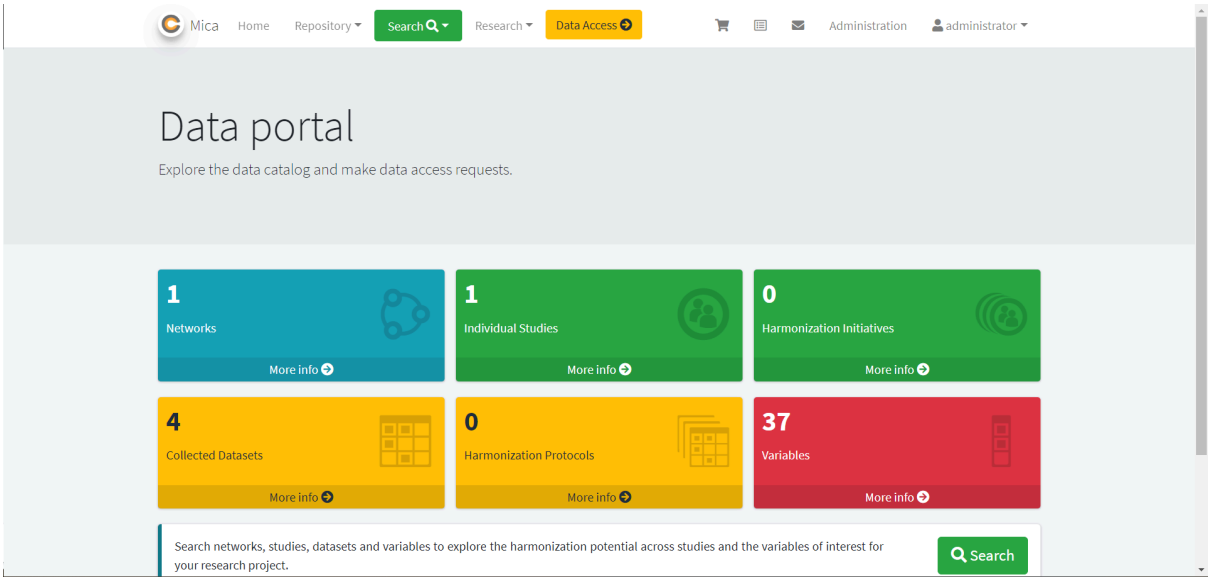


Figure 21. *Screenshot featuring the main dashboard of Mica, displaying the Web Data Portal.*

The platform allows searching by Network, Individual Studies, Collected Datasets, or Variables. This functionality makes it easy for researchers to search specific datasets or variables relevant to their research needs.

Figure 22 shows the list of available Networks, meanwhile Figure 23 shows a Network view. This feature allows users to easily access the metadata, such as datasets, protocols, variables, etc. that are part of specific research networks. This is particularly useful for collaborative projects where data from multiple sources is integrated and analyzed.

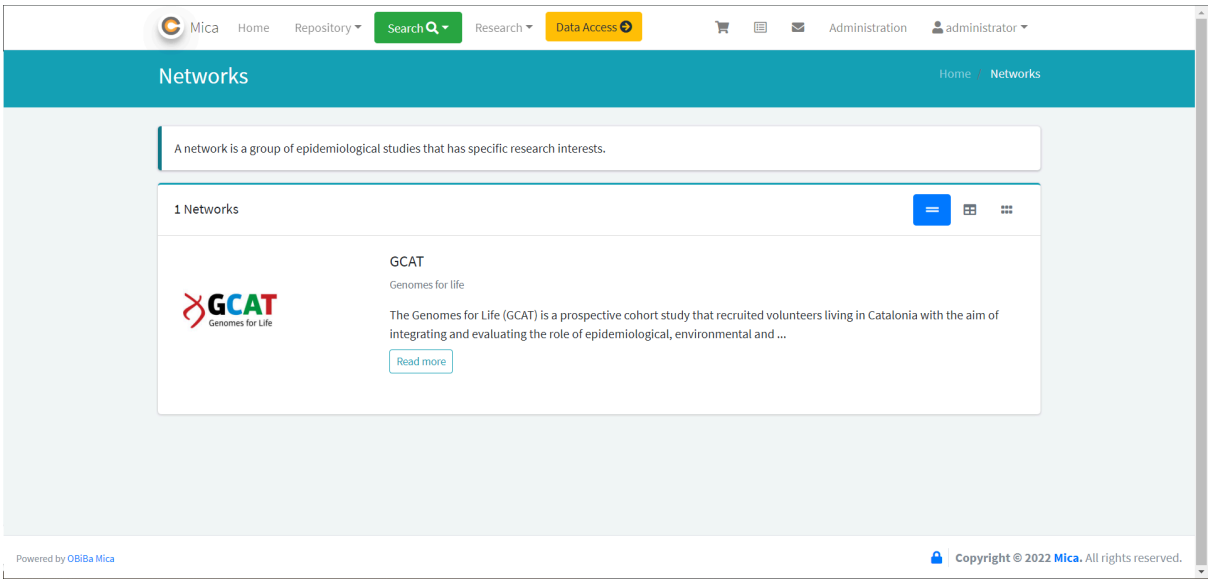


Figure 22. *Screenshot showing the search functionality by Network in the Mica web data portal.*

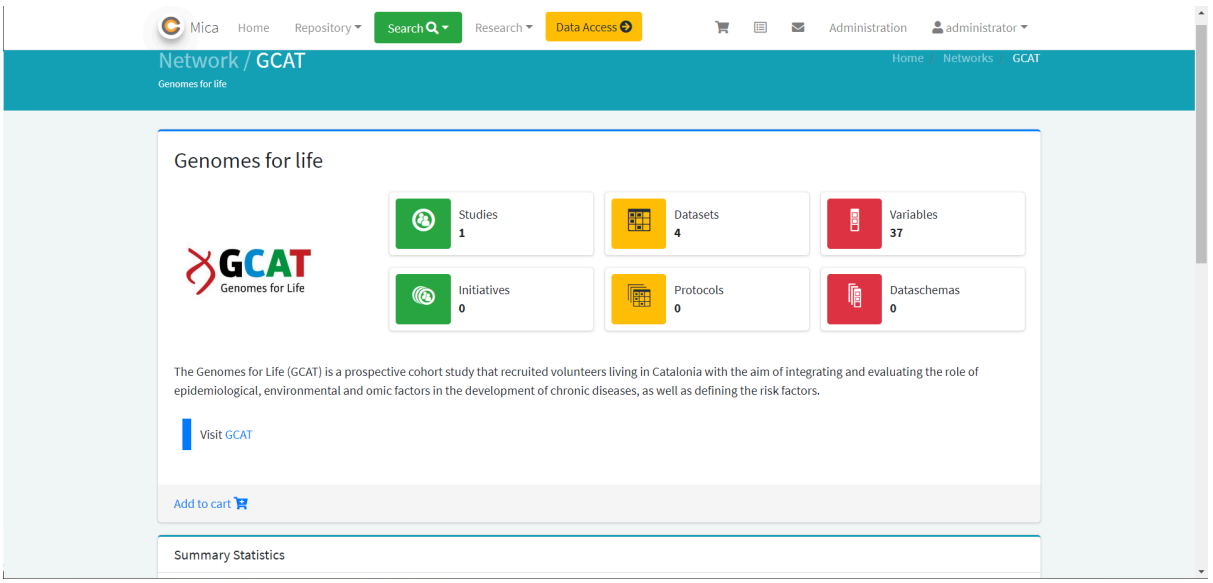


Figure 23. *Screenshot showing the GCAT Network view.*

The search feature, as can be seen in Figure 24, provides a comprehensive list of all datasets collected within the system. Users can filter and sort these datasets based on various criteria, facilitating efficient data retrieval.

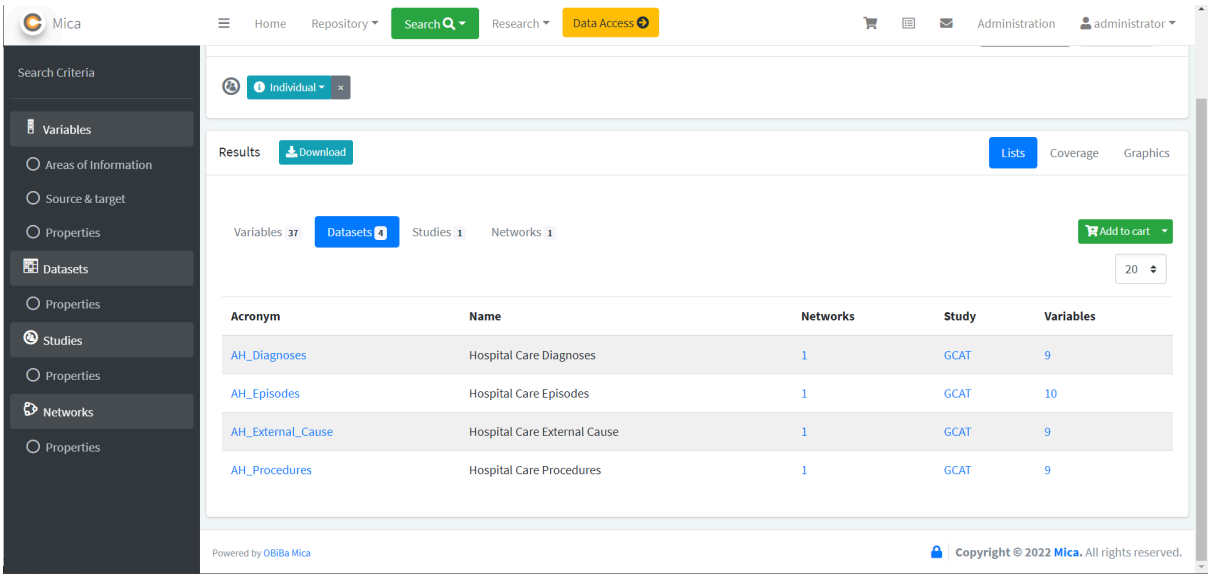


Figure 24. *Screenshot showing the search functionality by Collected Datasets in the Mica web data portal.*

The metadata associated with each collected data set can also be viewed by selecting one of them from the list. As for example, in Figure 25, the information related to the *AH_Episodis* collected dataset is available.

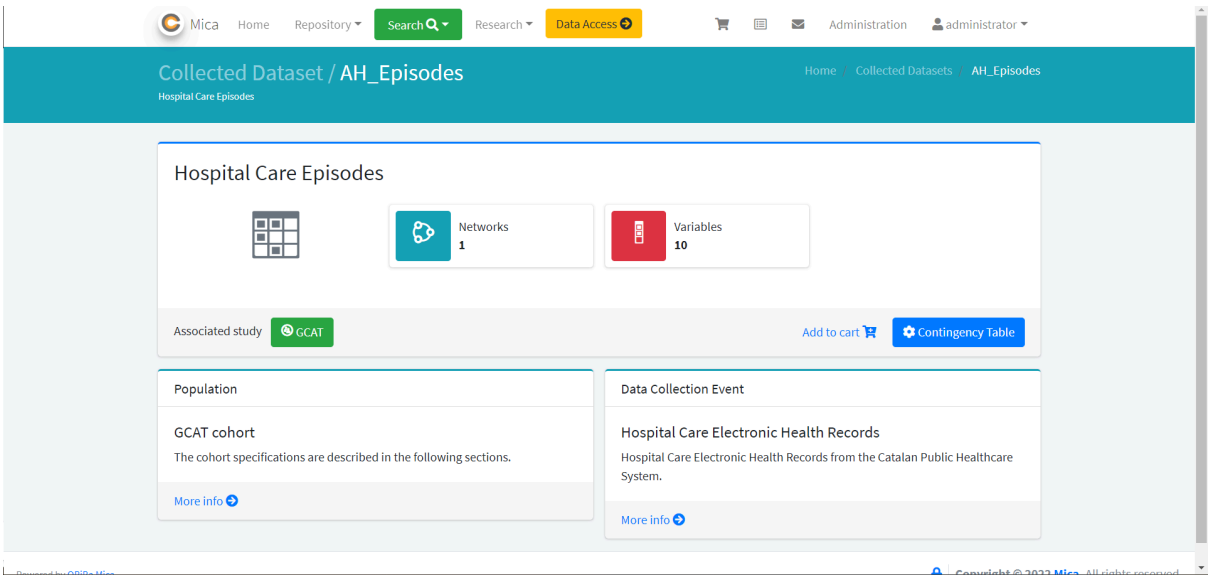


Figure 25. *Screenshot showing the AH_Episodis collected dataset.*

Among all the features, we can see the detail of a variable as shown in Figure 26. All the information related to the variable "*circ_alta_desc*" is displayed, such as the valueType, the units, the categories it has (if it is a categorical variable, as it is the case). Furthermore, it is possible to see in which dataset it is located, to which population group, etc.

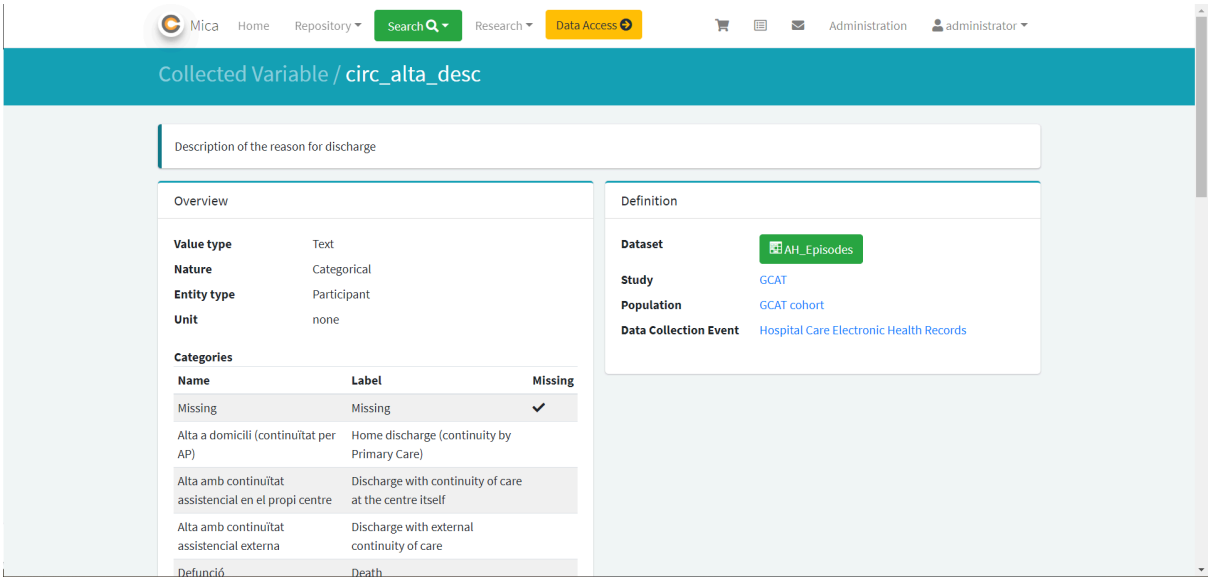


Figure 26. *Screenshot showing all the information related to the variable circ_alta_desc.*

All these features available in the Mica web data portal collectively enhance the data exploration experience, allowing researchers to efficiently access and use the data collected from the Catalan cohort. By providing multiple search options, Mica ensures that researchers can find the data they need in the most convenient and effective manner.

5.2. Synthetic data

In the following section, the results of the synthetic data generation are presented. Due to access restrictions to the original database, synthetic data were generated from aggregate data from the original data. This approach allows replicating the key statistical properties of the original data set while protecting the privacy and confidentiality of the original information.

Given the long list of datasets available for the cohort, we decided to focus only on basic demographic data and on the care hospital's EHR. Each of the generated data sets is detailed below, as well as the assumptions we have considered for the subsequent steps:

- **Demographic data set** - contains basic information about individuals, such as (i) the age, which ranges from 40 to 65 years, and (ii) the biological sex at birth of the person, being considered male or female.
- **Diagnostic data set** - contains information on patients' medical diagnoses, including the variables in Table 9. This data set is essential to understand the prevalence and type of diseases treated.

Variable	Description
*EGA_id**	Unique identifier assigned to each individual in the EGA system
id	Unique identifier of the record in the database
*DATA_INGRES**	Date of entry (i.e. date of admission of the patient to the care hospital)
EDAT_INGRES	Age at entry (i.e. age of the patient at admission)
*pos_dx**	Position of diagnosis in the list of diagnoses associated with the patient (main diagnose, secondary, etc.)
ICD9	Diagnosis code according to the International Classification of Diseases, 9th edition
ICD9_desc	Description of the diagnosis associated with the ICD-9 code
*ICD10**	Diagnosis code according to the International Classification of Diseases, 10th edition
ICD10_desc	Description of the diagnosis associated with the ICD-10 code

Table 9. *Description of the variables of the diagnoses data set. The table presents the names of the variables together with their descriptions.*

- **Procedure data set** - records medical interventions performed on patients during their care, crucial for analyzing the treatments and resources used.

Variable	Description
*EGA_id**	Unique identifier assigned to each individual in the EGA system.
id	Unique identifier of the record in the database
*DATA_INGRES**	Date of entry (i.e. date of admission of the patient to the care hospital)
EDAT_INGRES	Age at entry (i.e. age of the patient at admission)
*pos_proc**	Position of procedure in the list of procedures associated with the patient (main procedure, secondary, etc.)
ICD9	Procedure code according to the International Classification of Diseases, 9th edition
ICD9_desc	Description of the procedure associated with the ICD-9 code
*ICD10**	Procedure code according to the International Classification of Diseases, 10th edition
ICD10_desc	Description of the procedure associated with the ICD-10 code

Table 10. *Description of the variables of the procedure data set. The table presents the names of the variables together with their descriptions.*

- **Event data set** - documents significant events during the patient's hospital stay, such as admissions to the Intensive Care Unit (ICU) and the reason for discharge from the hospital.

Variable	Description
*EGA_id**	Unique identifier assigned to each individual in the EGA system.
id	Unique identifier of the record in the database
*DATA_INGRES**	Date of entry (i.e. date of admission of the patient to the care hospital)
*DATA_ALTA**	Date of discharge of the patient from the care hospital
*INGRES_UCI**	Entry to Intensive Care Unit
*HORA_INGRES_UCI**	Time of entry to the Intensive Care Unit
*DATA_ALTA_UCI**	Date of discharge of the patient from the Intensive Care Unit
circ_alta	Reason for discharge from hospital (internal numerical classification)
*circ_alta_desc**	Description of reason for discharge from hospital
EDAT_INGRES	Age at entry (i.e. age of the patient at admission

Table 11. *Description of the variables of the episodes data set. The table presents the names of the variables together with their descriptions.*

- **External cause data set** - specifies the external causes that have led to patients' medical conditions or injuries. It includes codes and descriptions of external causes, such as accidents, poisonings, and acts of violence, along with relevant identifiers and dates. This data set is important for analyzing external factors affecting health and developing prevention strategies.

Variable	Description
*EGA_id**	Unique identifier assigned to each individual in the EGA system.
id	Unique identifier of the record in the database
any_visita	Year of external cause visit
*DATA_INGRES**	Date of entry (i.e. date of admission of the patient to the care hospital)
*DATA_ALTA**	Date of discharge of the patient from the care hospital
EDAT_INGRES	Age at entry (i.e. age of the patient at admission)
*posicio**	Position of external cause in the list of external causes associated with the patient
*codi**	Diagnosis code according to the International Classification of Diseases, 10th edition
descripcio_codi	Description of the diagnosis associated with the ICD-10 code

Table 12. *Description of the variables of the external cause data set. The table presents the names of the variables together with their descriptions.*

During the analysis of the data sets, we found that many of the variables present are redundant (i.e. they contain the same or similar information), repeating unnecessarily. Furthermore, due to the way this synthetic dataset has been generated, there is no relation between the variables. In other words, it has not been specified that a specific variable is related to another one. On the one hand, it can be seen that all data sets contain two identifiers: (i) EGA_id and (ii) id. However, we have not defined that an EGA identifier corresponds to a cohort identifier. Therefore, we decided to keep only one of them, EGA_id.

On the other hand, in both the diagnosis and procedure data sets, we have identified ICD-9 and ICD-10. Although both serve to classify diseases and medical conditions, they are redundant in this context. Moreover, when creating the synthetic data, we realized that the codes would be different and, it does not provide no additional value when both are included. For this reason, we decided to keep only the ICD-10 codes, which is the most up-to-date and detailed version of the coding system. We also chose to remove the descriptions associated with these codes, as they are not essential for the purpose of the analysis and their inclusion only contributes to redundancy.

For this reason, we have decided to remove the ICD-9 diagnosis codes and keep only the ICD-10 codes, as the two codes would be different and would not add value to the analysis. We have also chosen to remove the descriptions associated with these codes, as they are not essential for the purpose of the analysis and contribute to redundancy.
Also, we noticed that the age at entry variable (*EDAT_INGRES*) is redundant if we already have the year of birth and date of entry, as this variable will not be consistent in a synthetic data set. Therefore,

we have also decided to drop this variable. To this last point we also include the variable year of visit (*any_visita*) of the external cause data set, which will probably not match with the year specified in the date of entry (*DATA_INGRES*).

In the Event dataset, there are two variables that reflect the reason for discharge from hospital: (i) *circ_alta* and (ii) *circ_alta_desc*. However, we decided to keep only the second one, because the other one is an internal numerical classification that does not add value in our context.

To facilitate the understanding of this last point, we have marked all the variables we decided to use with an asterisk.

5.3. Data Standardization

This section presents the results of the procedures carried out to convert the raw data into the OMOP CDM format.

5.3.1. Traditional ELT

On the one hand, the traditional ELT was divided into three main phases: (i) extract, (ii) load and (iii) transform.

In the extraction phase, we first identified the five synthetic datasets that we wanted to transform. Then, we created a Meltano project and added the *tap-csv* extractor to extract these datasets in csv format. The configuration of the extractor was as follows:

```
plugins:
  extractors:
  - name: tap-csv
    variant: meltano
    pip_url: git+https://gitlab.com/meltano/tap-csv.git
    config:
      files:
        - entity: patient
          file: /.../demographic_report.csv
          keys: EGA_id

        - entity: diagnoses
          file: /.../AH_Diagnoses.csv
          keys: EGA_id

        - entity: procedures
          file: /.../AH_Procedures.csv
          keys: EGA_id

        - entity: episodes
          file: /.../AH_Episodes.csv
          keys: EGA_id

        - entity: external_cause
          file: /.../AH_External_Cause.csv
          keys: EGA_id

    add_metadata_columns: False
```

Figure 27. *Meltano's tap-csv extractor configuration. Figure created with Carbon.*

After that, for the load phase we created a PostgreSQL database using Docker, as well as we created the raw schema of the database. Later, Meltano's *target-postgres* loader was used to transfer the raw data to the PostgreSQL data warehouse. The configuration of the loader was as follows:

```
plugins:
  loaders:
  - name: target-postgres
    variant: meltanolabs
    pip_url: meltanolabs-target-postgres
    config:
      host: localhost
      port: 5432
      user: postgres
      database: datoscat_tfm
      default_target_schema: raw
      activate_version: False
      add_record_metadata: False
```

Figure 28. *Meltano's target-postgres loader configuration. Figure created with Carbon.*

The database loaded in the raw schema includes the tables: *patient*, *diagnoses*, *procedures*, *episodes* and *external_cause*. Figure 29 represents a diagram of these raw tables, illustrating their initial structure and variables prior to transformation. This diagram provides a clear view of the organization.

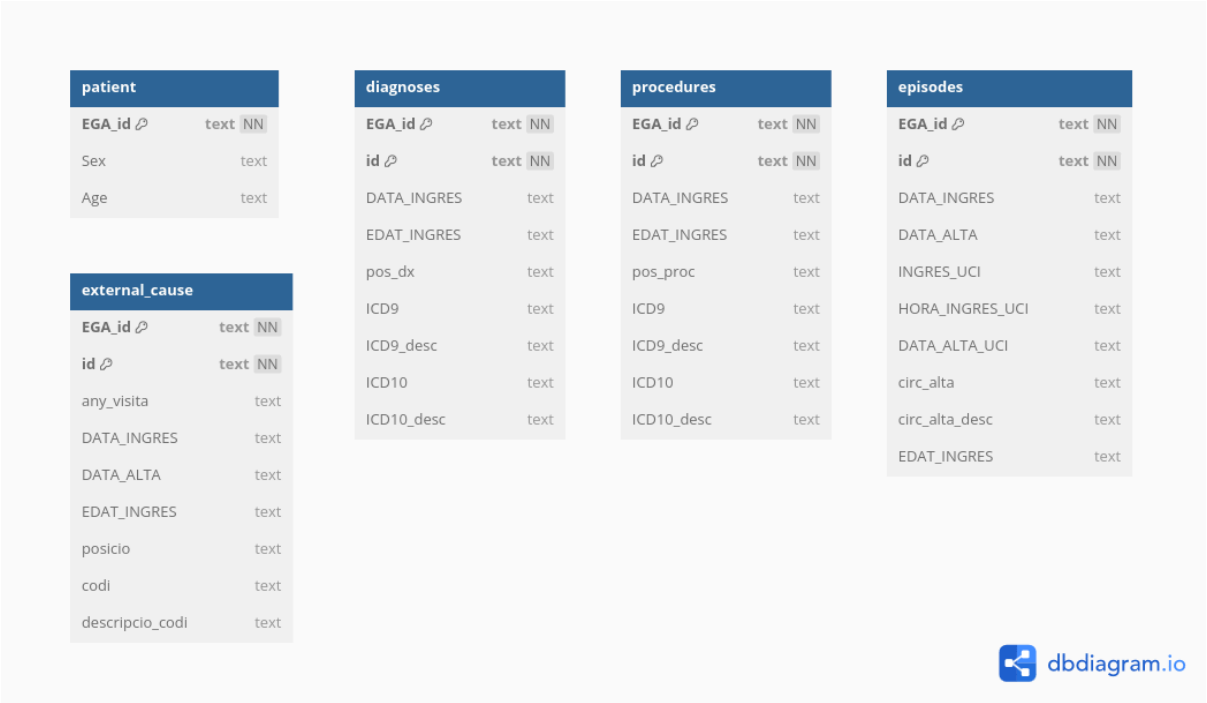


Figure 29. *Diagram of the raw data tables (patient, diagnoses, procedures, episodes and external_cause) illustrating the initial structure and variables, prior to the transformation to the OMOP CDM. Source: Created using dbdiagram.io*

Another key aspect was the **OMOP CDM vocabularies**, used to achieve semantic interoperability (i.e. to standardize raw data terminology to the OMOP CDM vocabulary). Therefore, we downloaded a list of vocabularies from Athena. Figure 30 shows the selected vocabularies from Athena.

Show all ▼ SHOW HISTORY DOWNLOAD VOCABULARIES

ID (CDM V4.5)	CODE (CDM V5)	NAME	REQUIRED	LATEST UPDATE
1	SNOMED	Systematic Nomenclature of Medicine - Clinical Terms (IHTSDO)		27-Sep-2023
2	ICD9CM	International Classification of Diseases, Ninth Revision, Clinical Modification, Volume 1 and 2 (NCHS)		01-Oct-2014
3	ICD9Proc	International Classification of Diseases, Ninth Revision, Clinical Modification, Volume 3 (NCHS)		01-Oct-2014
4	CPT4	Current Procedural Terminology version 4 (AMA)	EULA required	01-May-2023
5	HCPCS	Healthcare Common Procedure Coding System (CMS)		01-Jan-2024
6	LOINC	Logical Observation Identifiers Names and Codes (Regenstrief Institute)		18-Sep-2023
7	NDFRT	National Drug File - Reference Terminology (VA)		06-Aug-2018
8	RxNorm	RxNorm (NLM)		02-Jan-2024
9	NDC	National Drug Code (FDA and manufacturers)		25-Feb-2024

Figure 30. *Athena screenshot showing the list of downloaded vocabularies.*

Prior to the transformation process, it is important to understand the structure of the OMOP CDM vocabulary schema. Figure 31 provides a visual representation of the tables within the schema.

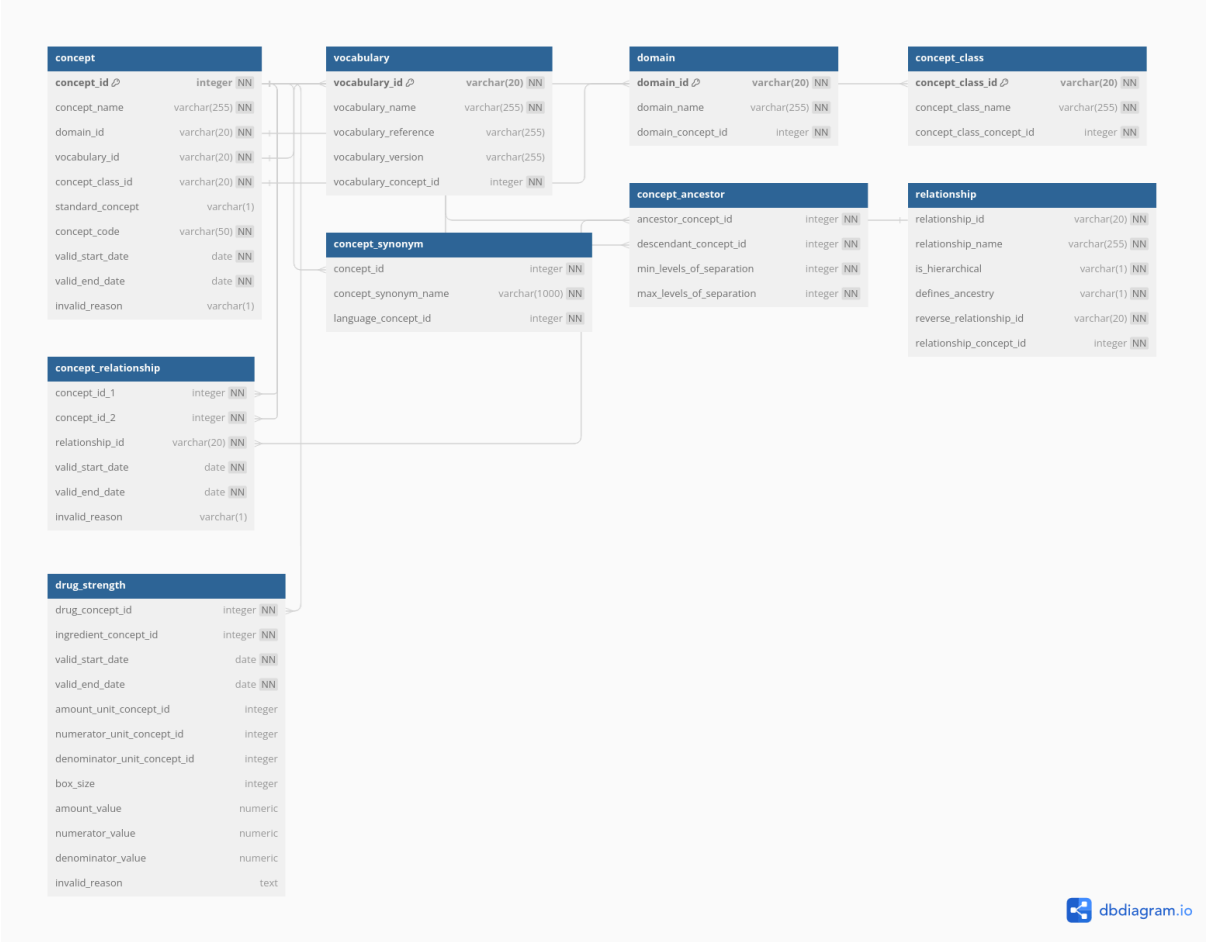


Figure 31. *Diagram illustrating the tables within the OMOP CDM vocabulary schema, detailing the structure and the relationship between tables. Source: Created using dbdiagram.io*

***Transform* phase**.

For the Transformation phase, we started with the syntactic mapping. Syntactic mapping is an important step in the data transformation process, due to it ensures that the data conforms to the structure required by the OMOP CDM. We generated a scan report of the *datoscat_tfm* postgreSQL database with the **White Rabbit** tool. In this process the different tables, columns and data types present in the *raw* schema are identified and documented.

First, White Rabbit was configured to connect to the *datoscat_tfm* postgreSQL database. Figure 32 shows the configuration screen, where the connection details are specified.

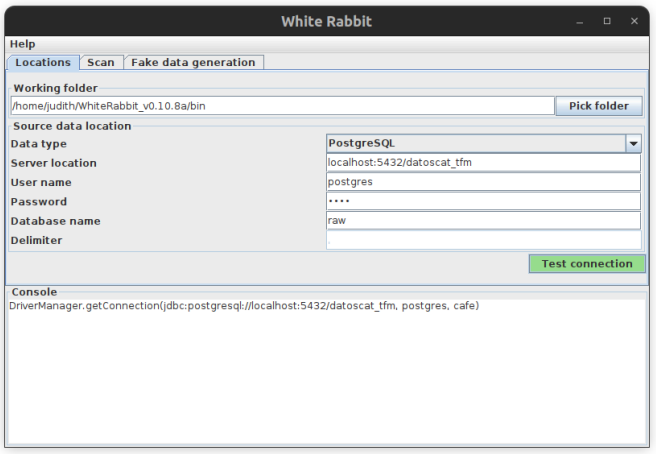


Figure 32. *White Rabbit configuration screen for connecting to the datoscat_tfm PostgreSQL database. Different details, including the data type, server location, user name, password and database name (i.e. schema) are specified.*

Once the connection was configured, White Rabbit was executed to scan the database and generate the report. Figure 33 displays the tables scanned at the top, and in the console part, it shows that the Scan Report has been generated.

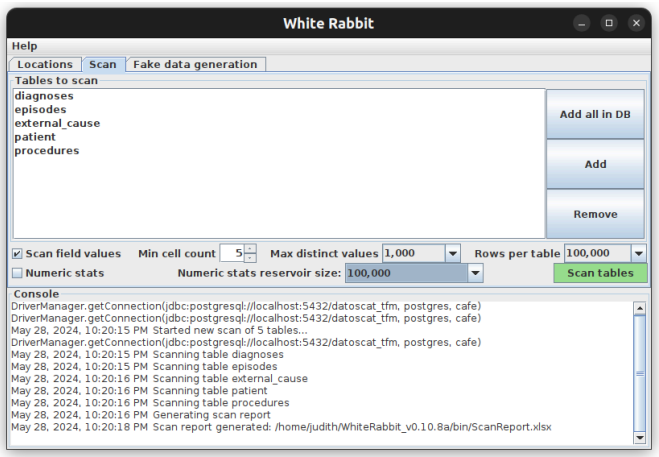


Figure 33. *White Rabbit scan screen. The top section displays the scanned tables, while the console section confirms the generation of the Scan Report.*

The Scan Report is used to understand the composition of the raw database, as it includes several tabs with key information. Each table in the raw schema of the database was examined.

- **Field Overview tab**. It offers a summary of the entire database. Table 13 summarizes part of the structure of the *Field Overview* tab obtained by scanning the raw schema of the *datoscat_tfm* database.

Table	Field	Type	N rows checked	N unique values	Fraction unique
patient	EGA_id	text	5216	5216	100%
patient	Sex	text	5216	2	0.0%
patient	Age	text	5216	26	0.5%
…	…	…	…	….	…
diagnoses	EGA_id	text	3949	3949	100%
diagnoses	DATA_INGRES	text	3949	2359	59.7%

Table 13. *Field overview tab of the Scan Report of the PostgreSQL database. Only some of the variables offered by this tab are displayed in the Table.*

- **Table Overview tab**. It lists all tables in the database along with the number of rows, number of rows checked, number of fields, and number of empty fields in each table. Table 14 summarizes the *Table Overview* tab obtained by scanning the raw schema of the *datoscat_tfm* database.

Table	Description	N rows	N rows checked	N Fields	N Fields Empty
patient	-	5216	5216	10	3
diagnoses	-	3949	3949	16	3
procedures	-	4296	4296	16	3
episodes	-	4636	4636	17	3
external_cause	-	335	335	16	3

Table 14. *Table overview tab of the Scan Report of the PostgreSQL database.*

- **Each table tab**. In the Scan Report there is a tab for each of the tables, i.e. one for *patient*, one for *diagnoses*, one for *procedures*, one for *episodes* and one for *external_cause*. In each of the tabs there is information on the variables, the options available for each variable along with their frequency. Table 15 summarizes the *patient* tab obtained by scanning the raw schema of the *datoscat_tfm* database.

EGA_id	Frequency	Sex	Frequency	Age	Frequency
List truncated…	…	Female	2629	56	235
		Male	2587	54	230
				…	…
				47	173
				40	169

Table 15. *Patient tab of the Scan Report of the PostgreSQL database. Only some of the categories offered by this tab are displayed in the Table.*

Once the scan report was generated, we used **Rabbit-in-a-Hat** tool to visually map the terms to the OMOP CDM. This process involved linking each variable of each table to its corresponding entity within the OMOP CDM, ensuring a clear and accurate representation of the data structure. More concretely, the datasets were mapped as follows:

- **Patient data set** - mapped to the “Person” table. This table contains information about the characteristics of individuals, such as age and sex.
- **Diagnoses data set** - mapped to “Visit Occurrence" and “Condition Occurrence" tables. These tables record medical visits and the conditions diagnosed during those visits.
- **Procedures data set** - mapped to “Visit Occurrence” and “Procedure Occurrence” tables. These tables document medical visits and the procedures performed during those visits.
- **Episodes data set** - mapped to “Visit Occurrence”, “Visit Detail” and “Death” tables. These tables capture details of medical visits and related events, including ICU stays and deaths.
- **External Cause data set** - mapped to “Visit Occurrence”, “Condition Occurrence” and “Observation” tables. These tables record medical visits, conditions and observations related to external causes.

Figure 34 shows a screenshot of the Rabbit-in-a-Hat diagram illustrating the relationships between the *datoscat_tfm* PostgreSQL database and the OMOP CDM v5.4. The diagram helps visualize the connections between tables, facilitating an understanding of the data structure.

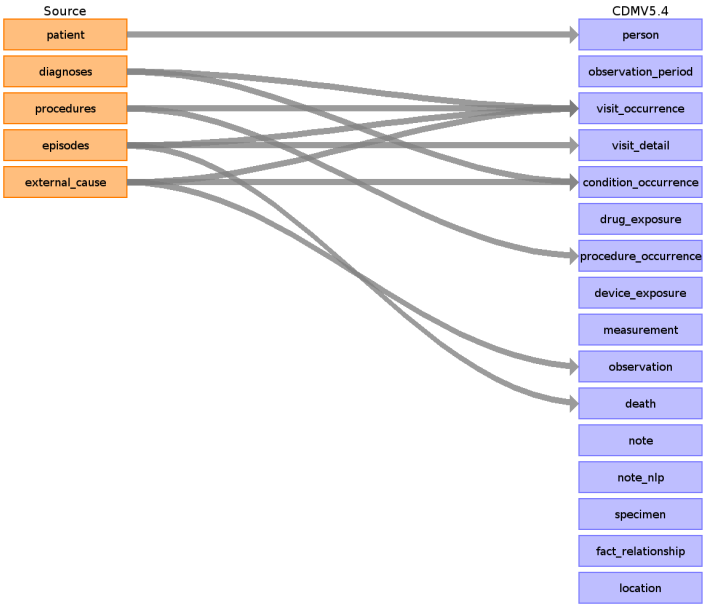


Figure 34. *Screenshot of the Rabbit-in-a-Hat diagram illustrating the relationships between the datoscat_tfm PostgreSQL database and the OMOP CDM v5.4. This visualization aids in understanding the connections between tables and the overall data structure.*

In addition, Figure 35 provides a detailed view of the *patients* table and how its variables are mapped to the *person* table in the OMOP CDM. This detailed mapping helps clarify how patient-specific information is transformed into the standardized format.

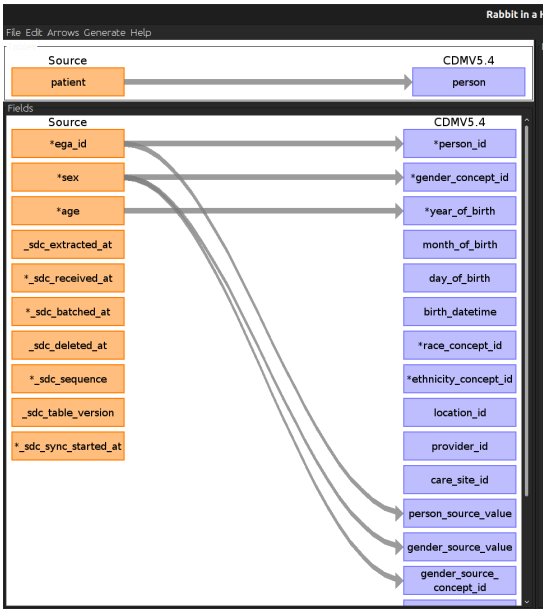


Figure 35. *Screenshot of the Rabbit-in-a-Hat with a detailed mapping of the "patient" table variables to the "person" table in the OMOP CDM v5.4. This detailed view illustrates the specific relationships and transformations applied to patient data.*

On the other hand, we did the semantic mapping. Semantic mapping ensures that the data also aligns with OMOP CDM vocabulary standards. This involves mapping source data concepts to standardized vocabularies.

For most of the concepts, we were able to directly map the source data using the *concept* and *concept_relationship* tables from the OMOP vocabulary. For other concepts, we decided to look directly for their equivalence in Athena, such as the variable *sex*. However, for the variable *circ_alta*, we saw that it was more difficult to find its standard codes and decided to use the **USAGI** tool.

USAGI helps in mapping non-standardized to standardized vocabularies by providing a user-friendly interface. Image X shows the process of importing the codes into USAGI to map the *circ_alta* variable, which represents the discharge circumstance from the *episodes* dataset.

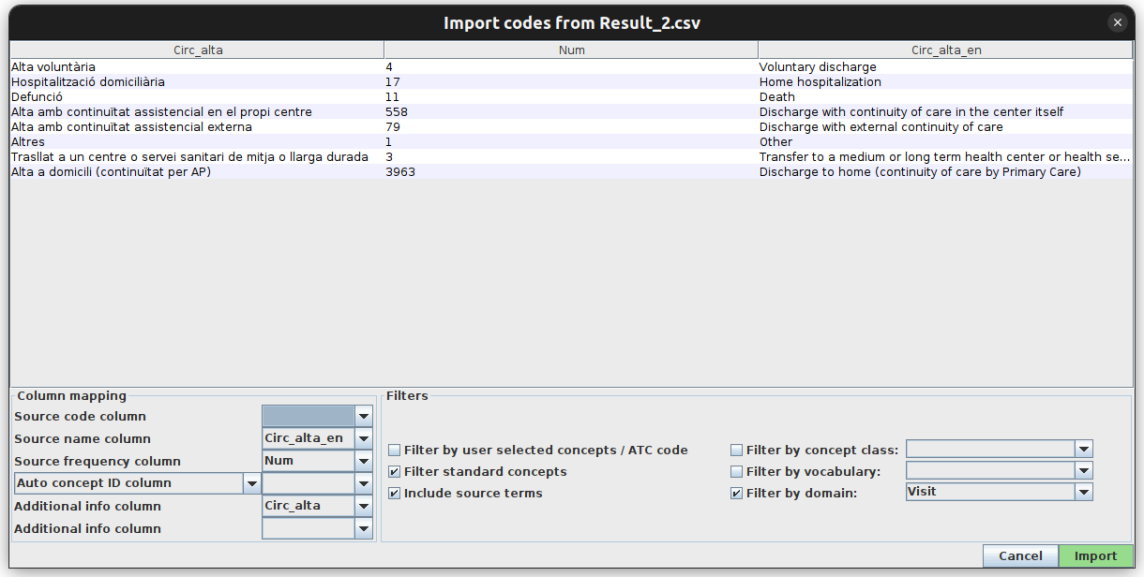


Figure 36. *Importing codes for the variable circ_alta into USAGI and the subsequent configuration.*

This tool also suggests a possible standard code based on the source data. Image X demonstrates the suggested mapping that Usagi performs, using the match with the highest score.

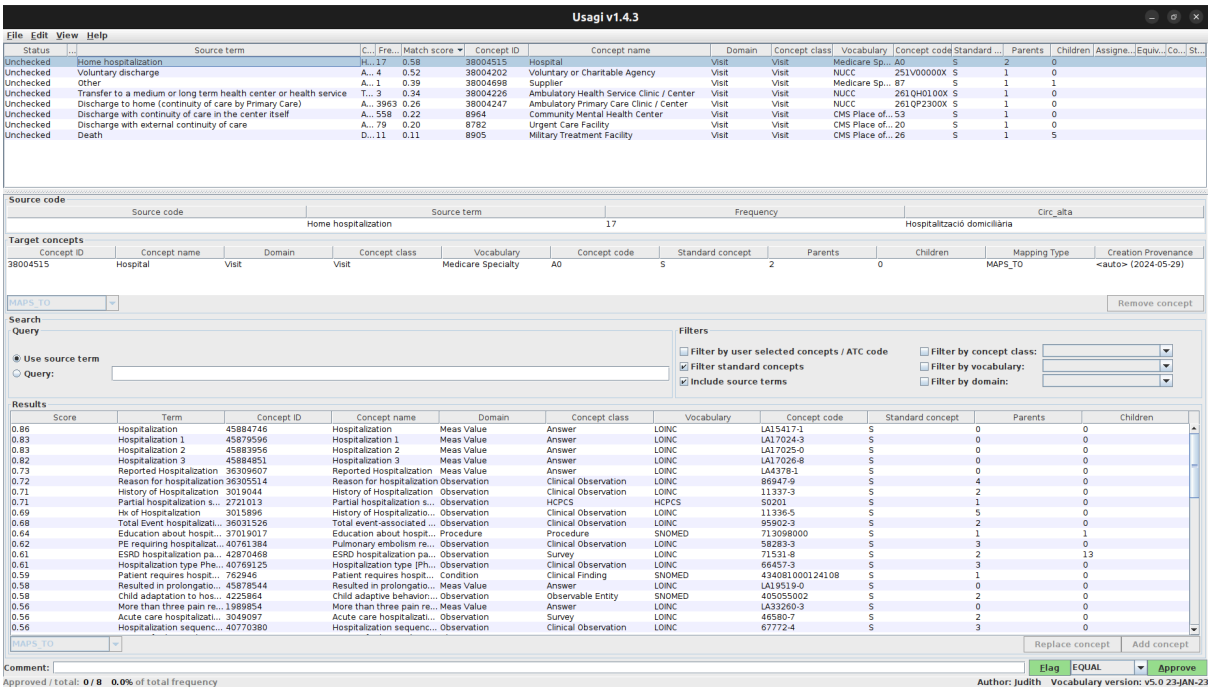


Figure 37. *USAGI mapping results showing the highest scoring match for the circ_alta variable.*

It was found that the codes for *circ_alta* do not have a standard mapping. This means that the specific codes from the dataset could not be directly aligned with the standardized terminologies in the OMOP CDM. Therefore, we have decided not to establish the reason for hospital discharge for the proof-of-concept. Consequently, further efforts may be needed to either refine these codes or develop custom mappings to ensure they are appropriately integrated and interoperable within the OMOP CDM.

As we mentioned in the methods section, **dbt** is a command-line tool that enables data analysts and engineers to transform data in their database more effectively. Dbt does this by combining SQL with software engineering best practices to manage and automate transformations.

Dbt employs Jinja, a template engine for Python, to add logic and dynamic behavior to SQL queries. Using Jinja, we defined the macros (essentially reusable SQL snippets), variables, control structures (such as loops and conditionals), and more within the dbt models.

Tables are represented as **models**, SQL queries that define a transformation from the source data to the target table. The content of these files is regular SQL, but with Jinja templates to add logic. Macros in dbt are reusable SQL code snippets also written in Jinja. They can be used to encapsulate logic that can be reused across multiple models. They can be also used to define and use **variables** within SQL queries to make them dynamical and use control structures like loops and conditionals to add logic to queries.

We generated the dbt project for the transformation of our data into the OMOP CDM. The following code, displayed in Figure 38, represents the transformation model for the OMOP CDM PERSON table.

```
-- person

{{ config(
    materialized='table',
    enabled=true
)
}}

with patients as (

    select *, row_number() over (partition by "EGA_id" order by "Age" desc) rn
    from {{ source('raw', 'patient') }}

),
person as (
    select
        {{ create_id_from_str('"EGA_id"::text')}} AS person_id,
        {{ gender_concept_id ('"Sex"') }} AS gender_concept_id,
        {{calculate_year('"Age"::int')}}::INT AS year_of_birth,
        NULL::INT AS month_of_birth,
        NULL::INT AS day_of_birth,
        NULL::TIMESTAMP AS birth_datetime,
        0::INT AS race_concept_id,
        0::INT AS ethnicity_concept_id,
        NULL::INT AS location_id,
        NULL::INT AS provider_id,
        NULL::INT AS care_site_id,
        "EGA_id"::VARCHAR(50) AS person_source_value,
        "Sex"::VARCHAR(50) AS gender_source_value,
        0::INT AS gender_source_concept_id,
        NULL::VARCHAR(50) AS race_source_value,
        0::INT AS race_source_concept_id,
        NULL::VARCHAR(50) AS ethnicity_source_value,
        0::INT AS ethnicity_source_concept_id
    from patients
    where "Age" is not null -- Don't load patients who do not have birthdate and sex
    and "Sex" is not null
    and rn = 1
)
select * from person
```

Figure 38. *SQL code to transfer raw data from the patient table to the OMOP CDM person table. The model is called person.sql. Source: Created using Carbon*

A breakdown of the person model can be seen below. More concretely, Figure 39 displays a dbt-specific macro used to configure the model. It sets the materialization strategy to "table", creating a physical table, and enables the model.

```
-- person

{{ config(
    materialized='table',
    enabled=true
)
}}
```

Figure 39. *Configuration settings of the person model within the SQL code used to map the patient table to the OMOP CDM person table. Source: Created using Carbon*

Figure 40 displays *patients* Common Table Expression (CTE), which is a temporary result set in SQL that you can reference within a SELECT, INSERT, UPDATE, or DELETE statement. CTEs are defined using the WITH clause followed by a query that generates the result set. CTEs are used to

improve the readability and maintainability of SQL code by breaking complex queries into simpler and more manageable parts.

```
with patients as (

    select *, row_number() over (partition by "EGA_id" order by "Age" desc) rn
    from {{ source('raw', 'patient') }}

),
```

Figure 40. *Segment of the person model within the SQL code, showing the patients CTE. Source: Created using Carbon*

Within the *patients* CTE there are two important parts:

- `{{ source('raw', 'patient') }}`. This is a dbt macro that references the "patient" table from the "raw" schema. It ensures the source table is correctly referenced and tracked by dbt.
- `row_number() over (partition by "EGA_id" order by "Age" desc)`. This assigns a unique row number to each partition of "EGA_id" ordered by "Age" in descending order to generate a unique identifier for the table.

```
person as (
    select
        {{ create_id_from_str('"EGA_id"::text')}} AS person_id,
        {{ gender_concept_id ('"Sex"') }} AS gender_concept_id,
        {{calculate_year('"Age"::int')}}::INT AS year_of_birth,
        NULL::INT AS month_of_birth,
        NULL::INT AS day_of_birth,
        NULL::TIMESTAMP AS birth_datetime,
        0::INT AS race_concept_id,
        0::INT AS ethnicity_concept_id,
        NULL::INT AS location_id,
        NULL::INT AS provider_id,
        NULL::INT AS care_site_id,
        "EGA_id"::VARCHAR(50) AS person_source_value,
        "Sex"::VARCHAR(50) AS gender_source_value,
        0::INT AS gender_source_concept_id,
        NULL::VARCHAR(50) AS race_source_value,
        0::INT AS race_source_concept_id,
        NULL::VARCHAR(50) AS ethnicity_source_value,
        0::INT AS ethnicity_source_concept_id
    from patients
    where "Age" is not null -- Don't load patients who do not have birthdate and sex
    and "Sex" is not null
    and rn = 1
)
select * from person
```

Figure 41. *Segment of the person model within the SQL code, showing the person CTE. Source: Created using Carbon*

The next CTE is *person*, represented in Figure 41. This query maps the source data into the corresponding fields of the OMOP CDM PERSON table. Several macros have been defined:

- `{ create_id_from_str('"EGA_id"::text')}}`. This generates a unique "person_id" from the "EGA_id" field.
- `{{ gender_concept_id('"Sex"') }}`. This maps the "Sex" field to a gender concept ID.

- `{{ calculate_year('"Age"::int') }}`. This calculates the year of birth from the "Age" field.

These macros can be seen in Figure 42. Moreover, the query filters out rows where "Age" or "Sex" is null, due to these are required fields in the OMOP CDM. At the same time, the query only keeps the first row (rn = 1) for each "EGA_id". A final `select * from person` selection creates the table.

```sql
-- Macro to create a bigint type ID from a string in absolute value
{% macro create_id_from_str(text) %}
    abs(('x' || substr(md5({{ text }}), 1, 16))::bit(64)::bigint)
{% endmacro %}

-- OMOP TABLE: person
--- Macro to transform 'M' and 'F' sex values into their concept_id
{% macro gender_concept_id(sex) %}
(CASE WHEN {{ sex }} = 'Male' THEN 8507::int -- Male
      WHEN {{ sex }} = 'Female' THEN 8532::int -- Female
      WHEN {{ sex }} is null THEN 0::int -- No data
      ELSE 8551::int -- Unknown
      END)
{% endmacro %}

-- OMOP TABLE: person
--- Macro to calculate the year of birth from age
{%macro calculate_year(age) %}
(2014 - {{age}})
{% endmacro %}
```

Figure 42. *SQL code responsible for executing the macros used in the transformation process. Source: Created using Carbon*

The project can be executed using dbt or meltano. Figure 43 shows the result of a successful execution of the transformation using the command `meltano invoke dbt-postgres:run`. The code for the whole DBT project can be found in the Supplementary Materials.

```
judith@judith-HP: ~/Documents/Judith/Master/TFM/TFM
16:46:19    compiled Code at ../.meltano/transformers/dbt/target/run/my_meltano_project/models/cdm/procedure_occurrence.sql
16:46:19
16:46:19  Done. PASS=13 WARN=0 ERROR=1 SKIP=0 TOTAL=14
(.venv) judith@judith-HP:~/Documents/Judith/Master/TFM/TFM$ meltano invoke dbt-postgres:run
2024-06-04T16:46:39.903025Z [info     ] Environment 'dev' is active
16:46:43  Running with dbt=1.3.7
16:46:43  Found 14 models, 0 tests, 0 snapshots, 0 analyses, 295 macros, 0 operations, 0 seed files, 7 sources, 0 exposures, 0 metrics
16:46:43
16:46:44  Concurrency: 2 threads (target='dev')
16:46:44
16:46:44  1 of 14 START sql table model cdm.cdm_source ................................... [RUN]
16:46:44  2 of 14 START sql table model cdm.death ........................................ [RUN]
16:46:44  1 of 14 OK created sql table model cdm.cdm_source .............................. [SELECT 1 in 0.33s]
16:46:44  2 of 14 OK created sql table model cdm.death ................................... [SELECT 11 in 0.34s]
16:46:44  3 of 14 START sql table model cdm.person ....................................... [RUN]
16:46:44  4 of 14 START sql table model cdm.visit_detail ................................. [RUN]
16:46:44  4 of 14 OK created sql table model cdm.visit_detail ............................ [SELECT 153 in 0.13s]
16:46:44  3 of 14 OK created sql table model cdm.person .................................. [SELECT 5216 in 0.20s]
16:46:44  5 of 14 START sql table model cdm_staging.diagnoses_condition_occurrence ....... [RUN]
16:46:44  6 of 14 START sql table model cdm_staging.diagnoses_visit_occurrence ........... [RUN]
16:46:44  6 of 14 OK created sql table model cdm_staging.diagnoses_visit_occurrence ...... [SELECT 2317 in 0.14s]
16:46:44  7 of 14 START sql table model cdm_staging.episode_visit_occurrence ............. [RUN]
16:46:45  7 of 14 OK created sql table model cdm_staging.episode_visit_occurrence ........ [SELECT 2726 in 0.17s]
16:46:45  8 of 14 START sql table model cdm_staging.external_cause_condition_occurrence .. [RUN]
16:46:47  5 of 14 OK created sql table model cdm_staging.diagnoses_condition_occurrence .. [SELECT 2107 in 2.85s]
16:46:47  9 of 14 START sql table model cdm_staging.external_cause_visit_occurrence ...... [RUN]
16:46:47  9 of 14 OK created sql table model cdm_staging.external_cause_visit_occurrence . [SELECT 193 in 0.13s]
16:46:47  10 of 14 START sql table model cdm.observation ................................. [RUN]
16:46:58  8 of 14 OK created sql table model cdm_staging.external_cause_condition_occurrence  [SELECT 52 in 13.50s]
16:46:58  11 of 14 START sql table model cdm.procedure_occurrence ........................ [RUN]
16:47:00  10 of 14 OK created sql table model cdm.observation ............................ [SELECT 90 in 13.29s]
16:47:00  12 of 14 START sql table model cdm_staging.procedures_visit_occurrence ......... [RUN]
16:47:01  12 of 14 OK created sql table model cdm_staging.procedures_visit_occurrence .... [SELECT 2554 in 0.14s]
16:47:01  13 of 14 START sql table model cdm.condition_occurrence ........................ [RUN]
16:47:01  13 of 14 OK created sql table model cdm.condition_occurrence ................... [SELECT 2159 in 0.10s]
16:47:01  14 of 14 START sql table model cdm.visit_occurrence ............................ [RUN]
16:47:01  14 of 14 OK created sql table model cdm.visit_occurrence ....................... [SELECT 7789 in 0.14s]
16:47:12  11 of 14 OK created sql table model cdm.procedure_occurrence ................... [SELECT 2554 in 14.01s]
16:47:12
16:47:12  Finished running 14 table models in 0 hours 0 minutes and 28.71 seconds (28.71s).
16:47:12
16:47:12  Completed successfully
16:47:12
16:47:12  Done. PASS=14 WARN=0 ERROR=0 SKIP=0 TOTAL=14
(.venv) judith@judith-HP:~/Documents/Judith/Master/TFM/TFM$
```

Figure 43. *Screenshot showing the result of the execution of the transformation process from the raw data to the OMOP CDM.*

Table 16 details the number of records that have been transformed into each respective OMOP CDM table. This detailed breakdown illustrates how the raw data has been accurately processed to a OMOP CDM standard. Each row represents a specific OMOP CDM table, along with the count of records that have been successfully migrated.

Table	Number of records
person	5216
visit_occurrence	7789
visit_detail	153
death	11
condition_occurrence	2159
procedure_occurrence	2554
observation	90

Table 16. *Number of records transformed into each respective OMOP CDM table from raw data, following the execution of Meltano.*

It is essential to understand the structure and relationships of the OMOP CDM to which data has been mapped. Figure 44 represents the diagram of the OMOP CDM, focusing on the tables we have mapped to (*person*, *visit_occurrence*, *visit_detail*, *death*, *condition_occurrence*, *procedure_occurrence*, *observation*, *condition_era* and *observation_period*). This diagram provides a comprehensive overview of these tables and their interconnections, ensuring clarity in how the various tables interact and support the standardized framework.

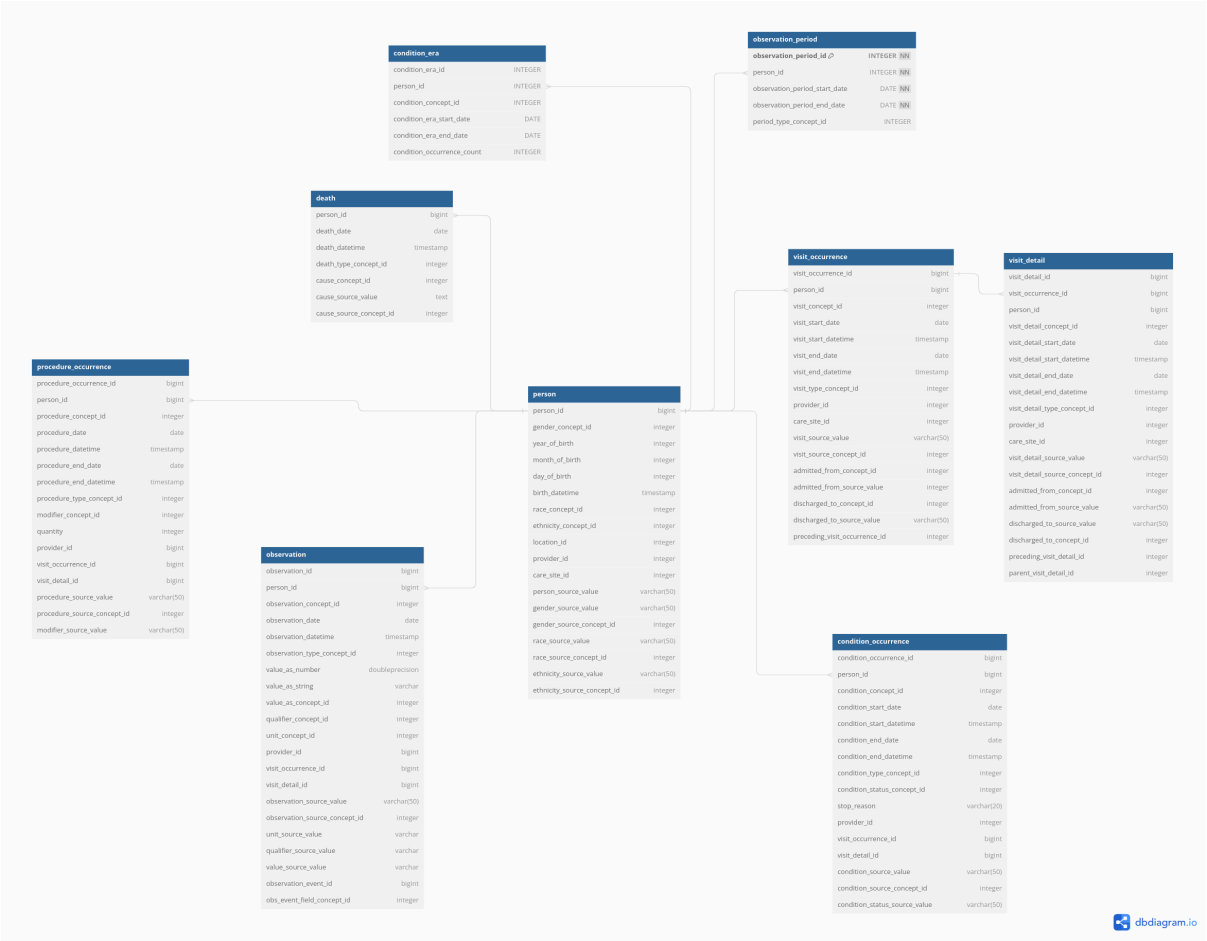
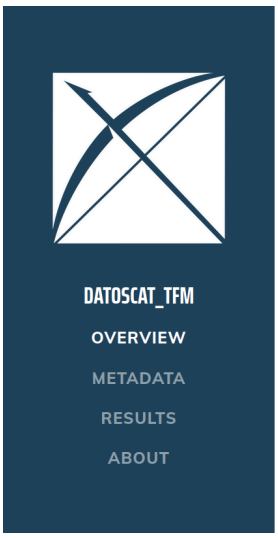


Figure 44. *Diagram of the OMOP CMD created. It illustrates the relationships and structure of the key tables: person, visit_occurrence, visit_detail, death, condition_occurrence, procedure_occurrence, observation, condition_era and observation_period. Source: Created using dbdiagram.io*

5.3.2. ELT evaluation and validation

The Data Quality Dashboard was used to assess the data quality of our synthetic dataset, which had been transformed to the OMOP CDM v5.4. The execution of the Data Quality Dashboard provided a comprehensive analysis of data conformance, completeness, and plausibility.

The overall data quality score obtained was 45%. This relatively low score is mainly due to the synthetic dataset used, which does not contain all the tables required by the OMOP CDM. The lack of tables and incomplete data entries significantly affected the completeness and conformance metrics, which are crucial components of the overall data quality score.



DATA QUALITY ASSESSMENT

DATOSCAT_TFM

DataQualityDashboard Version: 2.6.0
Results generated at 2024-06-04 22:06:19 in 7 mins

	Verification				Validation				Total			
	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass
Plausibility	123	375	498	25%	262	29	291	90%	385	404	789	49%
Conformance	320	427	747	43%	41	81	122	34%	361	508	869	42%
Completeness	155	202	357	43%	5	11	16	31%	160	213	373	43%
Total	598	1004	1602	37%	308	121	429	72%	906	1125	2031	**45%**

330 out of 906 passed checks are Not Applicable, due to empty tables or fields.
1058 out of 1125 failed checks are SQL errors.
Corrected pass percentage for NA and Errors: 90% (576/643).

Figure 45. *Results of the Data Quality Dashboard Assessment.*

If we make a detailed breakdown of the results:

- **Plausibility** (49% score). The nature of the synthetic data resulted in certain values and patterns that did not align with real-world data expectations.
- **Conformance** (42% score). The absence of several mandatory tables and fields has affected the conformance.
- **Completeness** (43% score). Many required data fields were missing, contributing to a low score.

These results highlight the limitations of using a synthetic dataset for comprehensive data quality assessment. Despite the low overall score, this exercise provided valuable insights into the areas needing attention for data quality improvement and the importance of using complete and accurate datasets for meaningful analysis.

5.3.3. Semantic ETL

The semantic ETL is based on the OntoBridge tool developed by the Hospital Clinic. We show here the semantic transformation of the patient dataset, which contains information on the age and sex of the individual.

One of the first things we needed for standardization was a local ontology, which represents in a semantic way the source dataset. Figure 46 shows a diagram illustrating the general structure of the local ontology and how it maps to the standardized ontology.

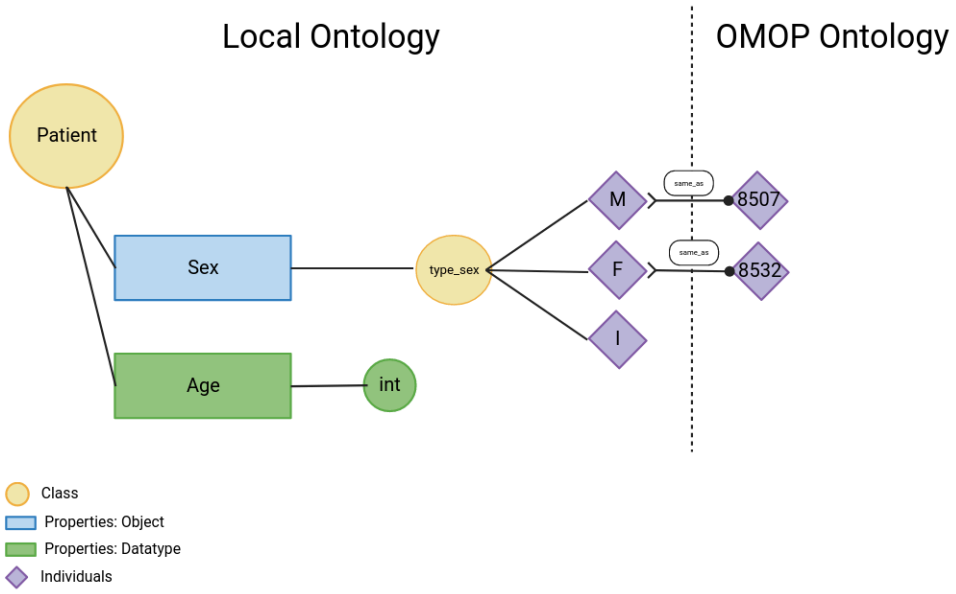


Figure 46. *Diagram illustrating the structure of the local ontology developed, showing the relationships and clinical concepts modeled from the Mica data dictionary.*

To automate the generation of this local ontology, a custom script was developed to generate the ontology from a data dictionary. This script takes as input the data dictionary and produces the local ontology in OWL format. An extract of the generated OWL, focusing on the variable "sex", is presented in Figure 47.

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
         xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
         xmlns:owl="http://www.w3.org/2002/07/owl#"
         xml:base="http://example_structure.org/onto.owl"
         xmlns="http://example_structure.org/onto.owl#">

<owl:Ontology rdf:about="http://example_structure.org/onto.owl"/>

<owl:ObjectProperty rdf:about="#demographic_Mica_dictionary_sex">
  <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#FunctionalProperty"/>
  <rdfs:domain rdf:resource="#demographic_Mica_dictionary"/>
  <rdfs:range rdf:resource="#demographic_Mica_dictionary_Sex"/>
</owl:ObjectProperty>

...

<owl:Class rdf:about="#demographic_Mica_dictionary_Sex">
  <rdfs:subClassOf rdf:resource="#Value_range"/>
</owl:Class>

<owl:NamedIndividual rdf:about="#demographic_Mica_dictionary_Sex_Male">
  <rdf:type rdf:resource="#demographic_Mica_dictionary_Sex"/>
</owl:NamedIndividual>

<owl:NamedIndividual rdf:about="#demographic_Mica_dictionary_Sex_Female">
  <rdf:type rdf:resource="#demographic_Mica_dictionary_Sex"/>
</owl:NamedIndividual>

</rdf:RDF>
```

Figure 47. *Fragment of the automatically generated local ontology OWL file, highlighting the semantic representation of the variable sex. Source: Created using Carbon*

Once the local ontology was completed, the R2RML (RDB to RDF Mapping Language) mapping was generated. This mapping allows the transformation of relational data into a semantic format compatible with RDF using *Ontop*. The script developed for this task produces files in Turtle format that represent the R2RML mapping.

```
@prefix rr: <http://www.w3.org/ns/r2rml#>.
@prefix xsd: <http://www.w3.org/2001/XMLSchema#>.
@prefix ex: <http://example_structure.org/onto.owl#>.
<#TriplesMapDemographic>
    a rr:TriplesMap;
    rr:logicalTable [ rr:sqlQuery "SELECT EGA_id, Sex, Age FROM patient" ];

    rr:subjectMap [
        rr:template "http://example_structure.org/onto.owl#demographic_Mica_dictionary#{}";
        rr:class ex:patients;
    ];
    # Triplet declaration for each property: ontology_variable
    rr:predicateObjectMap [
        rr:predicate ex:demographic_Mica_dictionary_EGA_id;
        rr:objectMap [ rr:column "EGA_id"; rr:datatype xsd:str ]
    ];

    rr:predicateObjectMap [
        rr:predicate ex:demographic_Mica_dictionary_Sex;
        rr:objectMap [ rr:column "Sex" ]
    ];

    rr:predicateObjectMap [
        rr:predicate ex:demographic_Mica_dictionary_Age;
        rr:objectMap [ rr:column "Age"; rr:datatype xsd:int
    ].
```

Figure 48. *Fragment of the R2RML code in Turtle format, used to map relational data to RDF, facilitating integration with the OWL local ontology. Source: Created using Carbon*

Then, the generated RDF data was integrated into the previously modeled OWL ontology, populating the data structures with patient instances and local clinical concepts. This process is repeated for the standard concept database. Moreover, semantic equivalences are established in the mapping ontology.

Figure 49 shows how the extension of the OMOP CDM ontology looks like with the elements corresponding to this table.

```
<omop:Field_type rdf:ID="gender_concept_id">
    <omop:type rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >INTEGER</omop:type>
    <rdfs:label xml:lang="en">gender_concept_id</rdfs:label>
    <omop:required rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >Yes</omop:required>
    <omop:description rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >A foreign key that refers to an identifier in the CONCEPT table for the unique gender of the person.
</omop:description>
    <rdfs:subClassOf rdf:resource="#PERSON_field"/>
  </omop:Field_type>al>

<owl:NamedIndividual rdf:about="#demographic_Mica_dictionary_Sex_Female">
  <rdf:type rdf:resource="#demographic_Mica_dictionary_Sex"/>
</owl:NamedIndividual>

</rdf:RDF>
```

Figure 49. *Fragment of the OWL file of the standard dictionary ontology, related to gender_concept_id. Source: https://github.com/InformaticaClinica/OntoBridge/tree/main. Source: Image created using Carbon*

Finally, a Python script was run that uploaded the resulting ontologies onto a local Jena Fuseki server, performed SPARQL data extraction queries and post-processed the resulting CSV file to ensure compliance with the CDM. SPARQL queries and post-processing in Python are tailored to each CDM and therefore do not depend on the source data model, ensuring full scalability.

Although we have only performed the transformation for the PERSON table, the effectiveness of semantic ETL in facilitating interoperability and data analysis in the context of OMOP CDM has been demonstrated, providing a coherent and flexible semantic representation of health data.

5.4. Data Analysis

The data analysis section was conducted using the dsOMOP package, a powerful tool which enables federated data analysis with datasets standardized to the OMOP CDM.

Prior to analysis, the dsOMOP package was configured on both the server and client sides. We first configured the package on the server-client environment (i.e. the Opal server). Figure 50 shows the list of packages available in the DataSHIELD R Server, dsOMOP was added successfully.

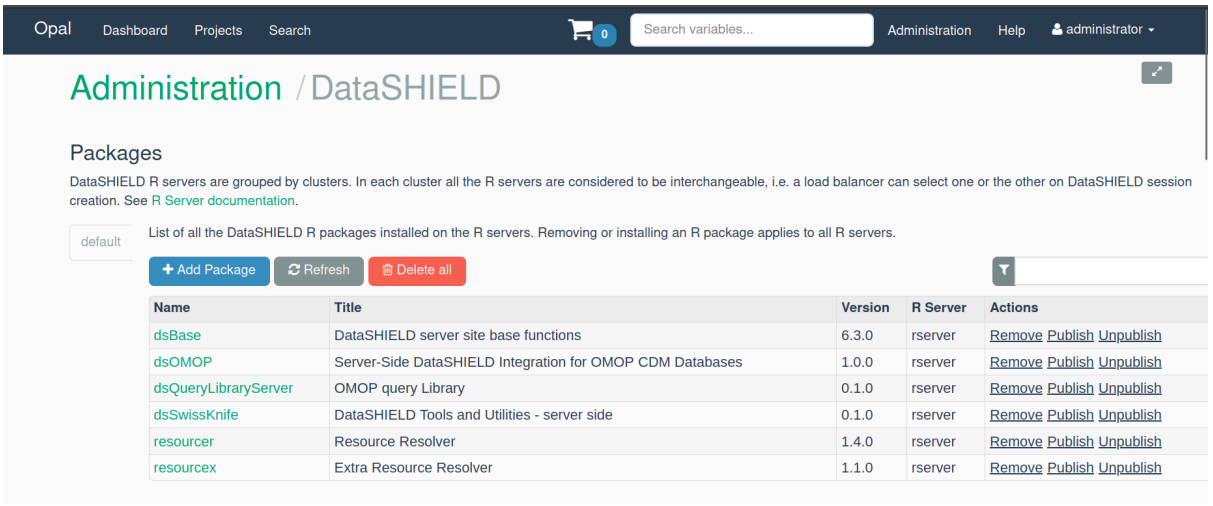


Figure 50. *Screenshot of the DataSHIELD administration in the Opal server.*

After that, we needed to upload the standardized database. To do so, we created a resource in the Opal project, which connects to the *cdm* schema of our *datoscat_tfm* PostgreSQL database. Figure 51 shows the main view of the DATOS-CAT resource section. The process also involved defining access permissions to ensure data privacy and integrity.

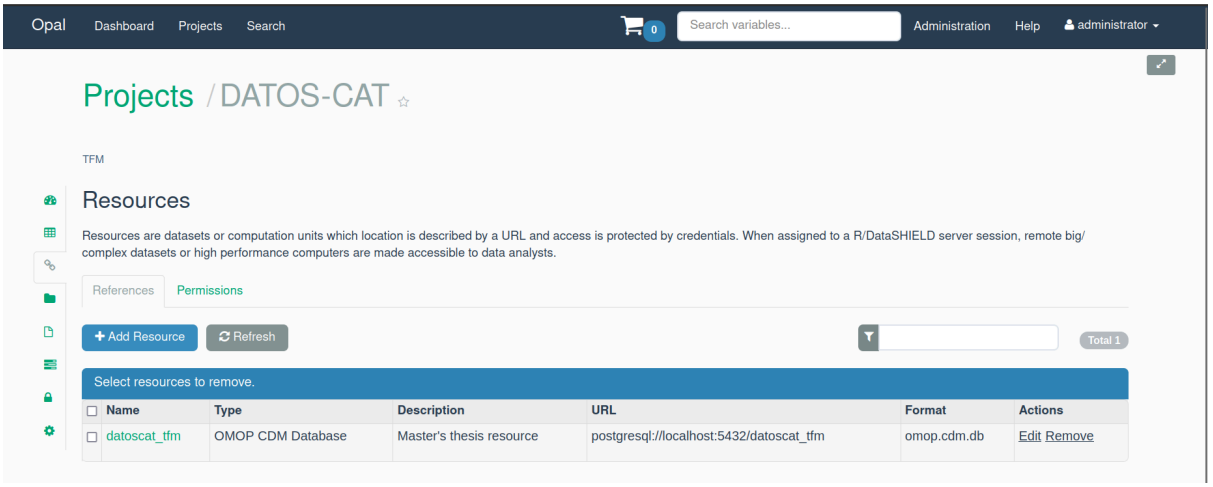


Figure 51. *Screenshot of the DATOS-CAT resources section in the Opal server.*

Once configured, the package facilitated seamless access to the OMOP CDM created in the previous section, while preserving data confidentiality. Through federated queries, OMOP CDM tables such as diagnoses, procedures, medication and observations can be accessed. Therefore, we made a couple of queries to check that we could access our PostgreSQL database and see how the tool worked.

The code for connecting to the database using dsOMOPClient can be found in the Supplementary Material. As dsOMOP allows to integrate other libraries of the DataSHIELD ecosystem, we used dsBaseClient package to analyze the sex distribution, by accessing the *gender_concept_id* value of the OMOP CDM. Figure 52 shows the query to access the *gender_concept_id* value, while Figure 53 shows the result. If we look at the last part of the query result, we see that there is a count for females (n = 2629) and another for male (n =2587). These values are consistent with the values in Table 15 of the scan report.

```
ds.table("omop$gender_concept_id")
```

Figure 52. *DataSHIELD command to access "gender_concept_id". Source: Created using Carbon*

```
Aggregated (exists("gender_concept_id", omop)) [====================================] 100% / 0s
Aggregated (asFactorDS1("omop$gender_concept_id")) [================================] 100% / 3s
Aggregated (tableDS(rvar.transmit = "omop$gender_concept_id", cvar.transmit = NULL, ) [=] 100% ...

 Data in all studies were valid

Study 1 :  No errors reported from this study

$output.list
$output.list$TABLE_rvar.by.study_row.props
                    study
omop$gender_concept_id tfm
               female   1
               male     1
               NA     NaN

$output.list$TABLE_rvar.by.study_col.props
                    study
omop$gender_concept_id       tfm
               female 0.5040261
               male   0.4959739
               NA     0.0000000

$output.list$TABLE_rvar.by.study_counts
                    study
omop$gender_concept_id  tfm
               female 2629
               male   2587
               NA        0
```

Figure 53. *Output of the DataSHIELD command showing the "gender_concept_id" results.*

We also checked the age distribution of the individuals, accessing year_of_birth. The histogram in Figure 54 shows the distribution of the year of birth, while Figure 55 presents the quantiles and the mean of the same variable.

Figure 54. *Histogram of the year_of_birth distribution. This histogram illustrates the distribution of the year of birth among the individuals, indicating the frequency for a group of years.*

```
  Aggregated (exists("year_of_birth", omop)) [========================================] 100% / 0s
  Aggregated (classDS("omop$year_of_birth")) [========================================] 100% / 0s
  Aggregated (isValidDS(omop$year_of_birth)) [========================================] 100% / 2s
  Aggregated (lengthDS("omop$year_of_birth")) [========================================] 100% / 0s
  Aggregated (quantileMeanDS(omop$year_of_birth)) [========================================] 100% / 0s
$tfm
$tfm$class
[1] "integer"

$tfm$length
[1] 5216

$tfm$`quantiles & mean`
      5%      10%      25%      50%      75%      90%      95%     Mean
1950.000 1951.000 1955.000 1961.000 1968.000 1972.000 1973.000 1961.246
```

Figure 55. *Descriptive Statistics of the year_of_birth. The query answer provides the quantiles (5%, 10%, 25%, 50%, 75%, 90%, 95%) and the mean of the year of birth data.*

There is the option to retrieve specific tables, concepts, or columns, as the initial one only includes data from the Person table. In Figure 56, we access *concept_id 320128* which refers to essential hypertension from the *diagnosis_occurrence* table. The results can be seen in Figure 57.

```
o$auto(tables = c("condition_occurrence"),
       concepts= c(320128))
ds.summary("omop")
```

Figure 56. *DataSHIELD command to retrieve the concept_id 320128 from the condition_occurrence table. Source: Created using Carbon*

```
  Assigned all resource (dsO.wLY8 <- ...) [=============================================] 100% / 4s
  Aggregated (...) [=============================================================] 100% / 2s
  Assigned all resource (dsO.wLY8 <- ...) [=============================================] 100% / 2s
  Assigned expr. (dsOH.HDy4 <- ...) [============================================] 100% / 3s
  Aggregated (exists("omop")) [==================================================] 100% / 0s
  Aggregated (exists("dsOH.HDy4")) [=============================================] 100% / 0s
  Assigned expr. (omop <- mergeDS("omop", "dsOH.HDy4", "person_id", "person_id", TRUE, ) [=] 100%...
  Aggregated (testObjExistsDS("omop")) [=========================================] 100% / 0s
  Aggregated (messageDS("omop")) [===============================================] 100% / 3s
  Aggregated (exists("omop")) [==================================================] 100% / 0s
  Aggregated (classDS("omop")) [=================================================] 100% / 0s
  Aggregated (isValidDS(omop)) [=================================================] 100% / 1s
  Aggregated (dimDS("omop")) [===================================================] 100% / 1s
  Aggregated (colnamesDS("omop")) [==============================================] 100% / 0s
$tfm
$tfm$class
[1] "data.frame"

$tfm$`number of rows`
[1] 5216

$tfm$`number of columns`
[1] 11

$tfm$`variables held`
 [1] "person_id"                                          "gender_concept_id"
 [3] "year_of_birth"                                      "race_concept_id"
 [5] "ethnicity_concept_id"                               "essential_hypertension.condition_occurrence_id"
 [7] "essential_hypertension.condition_start_date"        "essential_hypertension.condition_start_datetime"
 [9] "essential_hypertension.condition_type_concept_id"   "essential_hypertension.condition_status_concept_id"
[11] "essential_hypertension.visit_occurrence_id"
```

Figure 57. *Output of the DataSHIELD command to retrieve the concept_id 320128 from the condition_occurrence table.*

The analyses presented demonstrate the effectiveness and utility of using the dsOMOP package to access and retrieve standardized data within the OMOP CDM. The ability to efficiently extract demographic information such as *gender_concept_id* and specific conditions from the *condition_occurrence* table demonstrates the potential of this tool. This standardized approach can facilitate interoperability and harmonization of data from a variety of health care sources.

5.5. Data Discovery

The data discovery was implemented using a Beacon v2 API on top of our OMOP CDM. To do so, we downloaded the Gitlab repository and configured the environment to connect to our standard database. The main Uniform Resource Identifier (URI) to access the API (Application Programming Interface) is as follows: `localhost:5050/api/individuals`, Figure 58 shows that we have been able to implement the Beacon v2 in our database.

Figure 58. *Implementation of Beacon v2 in our database. This figure shows how we have managed to integrate Beacon v2 into our database.*

To validate the correct implementation of Beacon v2 in our database, we performed a specific query to count the number of *male* (i.e. gender_concept_id = 8507) records, specifying the following URI: `localhost:5050/api/individuals?filters=Gender:M`. The results obtained are shown in Figure 59. The query produced a *resultsCount* of 2587, which exactly matches the number of male records in our database. This result confirms that the implementation of Beacon v2 has been successful.

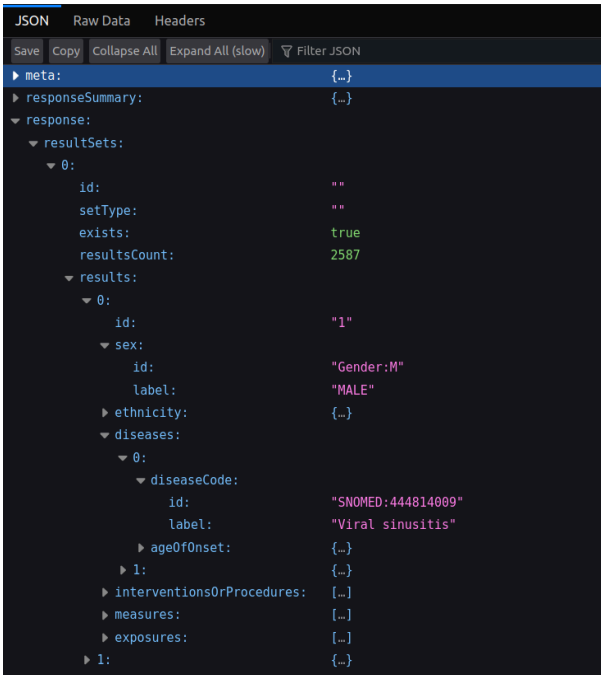


Figure 59. *Output after applying a filter to count the number of male records in our database.*

5.6. Code and data availability

All the code generated for this project is available in the Supplementary Material, or in the following GitHub repository: https://github.com/JudMartinez/TFM.

Chapter 6. Discussion and conclusions

The main objective of this Master's thesis is to contribute to the DATOS-CAT project by establishing guidelines, tools, and protocols to facilitate the implementation and use of various technologies, as well as to improve the interoperability and visibility of the Catalan population-based cohort.

6.1. Data Catalog

In line with the FAIR principles, Data Catalog is a fundamental tool for all projects working with data. The data catalog facilitates the location of metadata information, in a coherent, ordered and structured way.

The results presented in this Master thesis highlight the effectiveness of using Mica (OBiBa software) for data catalog deployment and management. The integration of Opal and Mica has demonstrated several significant advantages, which are crucial for the efficient handling of large-scale research datasets.
One of the key findings of OBiBa software is the streamlined organization and management of datasets within the Opal environment. The ability to upload and structure data dictionaries as tables in Opal provides a solid foundation for data integrity and consistency.

The configuration in Mica underscores the software's ability to establish intricate relationships between data sets. These relationships are essential for comprehensive data integration, enabling a holistic view of the data landscape. The visual representation of these connections in Mica helps to understand the interdependencies and linkages between the various components of the data.

The collected datasets' detailed documentation and the establishment of relationships with studies, populations, and data collection events further enhance the reliability and traceability of the data. This interconnected setup ensures that data provenance is maintained, as well as supports data consistency and reliability.

Another notable strength is the user-friendly design of the Mica web data portal. The main dashboard and versatile search functionalities provide an intuitive interface for data exploration. Researchers can efficiently search specific datasets or variables, enhancing the data discovery by offering flexibility and convenience.

Despite these strengths, there are some limitations to consider. One limitation is the potential complexity involved in setting up and configuring the Mica system. While the software offers powerful features, the initial setup may require a steep learning curve for new users. Additionally, while the relationships between datasets and other components are well-documented, ensuring that these connections remain up-to-date and accurate over time can be challenging.

In conclusion, the deployment of a data catalog using Mica has proven to be an effective solution for organizing, managing, and exploring large-scale research datasets. The integration of Opal and Mica offers a comprehensive framework that supports data integrity, accessibility, and analysis. While there are areas for improvement, the strengths of the system provide a strong foundation for efficient data management.

6.2. Synthetic Data

The generation of synthetic data was a crucial element of this Master's thesis, as it enabled the design of the DATOS-CAT project data flow while preserving data privacy. Synthetic data was generated based on aggregated data, more specifically, from the raw hospital care data of the GCAT cohort. This approach offers several distinct advantages over using pre-existing synthetic datasets like Synthea, particularly when developing a proof-of-concept for a project.

Generating synthetic data from aggregated data allows for the creation of datasets that closely mimic the actual characteristics of the target population. This allowed us to control the types and variables included in the dataset, ensuring that the relevant factors were represented. The solution adopted helped us to understand the conditions and the challenges the DATOS-CAT project aims to address. Therefore, the entire process can be adapted to the specific needs of the project, ensuring greater relevance and applicability.

At the same time, using aggregated data as the basis for synthetic data generation also enhances data privacy and security. Since the synthetic data is derived from aggregate statistics rather than individual records, it inherently minimizes the risk of re-identification and breaches of privacy. This aspect is crucial for projects dealing with sensitive information, such as the DATOS-CAT project. In this way, in this Master's thesis, we are ensuring compliance with data protection regulations without compromising the utility of the data.

However it has some limitations that can be critical for the analysis part. Synthetic data has been generated without considering any relationship between variables, which can lead to discrepancies with the real data.

In conclusion, generating synthetic data from aggregated data provided a more relevant, customizable, and secure approach for developing this Master's thesis and the proof-of-concept for the DATOS-CAT project. The realism of the variables helped us to better understand the impact and effectiveness of the proposed solutions, facilitating more informed decision-making and support.

6.3. Data Standardization

In this work, we have highlighted the importance of data standardization in facilitating data integration, analysis, and discovery. The use of common data models, such as the OMOP CDM, plays a critical role in enabling interoperability and comparability of data from different sources.

The OMOP CDM provides a standardized framework for representing healthcare data, allowing data from various EHRs and other sources to be harmonized and integrated. This standardization enables the application of consistent analytical methods and tools across different datasets, facilitating the generation of reliable and reproducible research findings.

To meet our objectives of comparing and evaluating two different Extract-Load-Transform (ELT) processes, assessing their effectiveness, efficiency, and compatibility with the established data model,

we have conducted an in-depth analysis. Through our investigation, we have determined that both approaches exhibit strengths in various aspects, highlighting their potential utility in transforming data into the OMOP CDM format. Table 17 provides a comprehensive comparison of both approaches outlining key aspects such as flexibility, reusability, scalability and required tools and knowledge of each method. The results of this comparison will be presented at the 34th Medical Informatics Europe Conference [96], and serve as a valuable resource for researchers and practitioners seeking insights into the selection of an appropriate ELT process for their data standardization endeavors within the healthcare domain.

Characteristic	OntoBridge	Traditional ELT
Flexibility	High: changes in variables require minimal efforts	Medium: changes in variables require some code adjustment.
Reusability	High: ontologies, SPARQL queries, and mappings are only performed once per data source/data model.	High: mappings and SQL queries are performed once per data source / data model.
Scalability	High: the same methodology can be used for different CDMs and different versions of each CDM with minimal adjustments to previous efforts.	Low: the methodology can only be used with OMOP CDM and on occasions might depend on the updates to OHDSI tools
Required tools and knowledge	Tools: Ontop, Protégé; Knowledge: ontologies, R2RLM mappings	Tools: White Rabbit, Rabbit-in-a-hat, Usagi, Meltano; Knowledge: SQL.

Table 17. *Comparison of the two Extract-Load-Transform (ELT) approaches, evaluating their flexibility, reusability, scalability, and required knowledge and tools. OntoBridge is used as a semantic approach.*

The results presented in this work demonstrate the successful implementation of the OMOP CDM for part of the GCAT cohort data. Once it will be implemented for the entire cohort, we will be able to seamlessly integrate data from multiple sources, including clinical, administrative, and biobank data, creating a comprehensive and cohesive dataset.

By leveraging the OMOP CDM, scientists will be able to use a wide range of existing OHDSI tools (e.g. ATLAs) and other resources for data analysis and visualization. These tools will enable us to conduct various analyses, including cohort identification, outcome assessment, and risk factor analysis, in a standardized and efficient manner.

The adoption of the OMOP CDM within the OHDSI environment has several benefits. First, it facilitates the replication and validation of research findings across different datasets, enhancing the robustness and reliability of the results. Second, it promotes collaboration and knowledge sharing among researchers, as standardized data and tools enable seamless data exchange and joint analysis. Third, it reduces the time and effort required for data preparation and analysis, allowing researchers to focus on higher-level scientific questions.

A key aspect is the standardization of data prior to collection, i.e., defining the collection in terms of standard data. During the project, we encountered several variables that do not have an established standard. This lack of common vocabularies can make standardization difficult. Although there is the option of creating new vocabularies, it does not make sense to standardize something that only we

could use. Therefore, addressing this issue is crucial to ensure the interoperability and usefulness of the data collected.

In conclusion, the standardization of a synthetic dataset representing part of the GCAT cohort, using the OMOP CDM, has significantly enhanced data integration, analysis, and discovery. The adoption of common data models and tools will enable the activation of a broader community of scientists around the GCAT cohort data, fostering collaborative research and accelerating the generation of new knowledge.

6.4. Data Analysis

Federated analysis methods provide a secure and privacy-preserving approach to analyze data across multiple institutions without the need for physical transfer. This is especially valuable for sensitive healthcare data, as it minimizes the risk of data breaches and re-identification.

The DATOS-CAT project will use these methods to enable researchers to conduct collaborative studies on chronic diseases and their risk factors, while adhering to strict data protection regulations. Specifically, it intends to use DataSHIELD, a solution that allows researchers to perform federated queries and analyze data from multiple sources without compromising patient privacy. This approach eliminates the need for data transfer, ensuring data security.

To enable querying across different institutions or databases, data harmonization is essential. Therefore, healthcare data is standardized to a common data format, such as the OMOP CDM. To facilitate federated cross-institutional studies through standardized databases, ISGlobal has developed dsOMOP, a DataSHIELD package tailored for the OMOP CDM.

Looking ahead, federated analysis methods have the potential to revolutionize the way healthcare data is shared and analyzed. By enabling researchers to collaborate across institutional boundaries, these methods can accelerate the pace of medical research and lead to new discoveries that improve patients' outcomes.

In conclusion, the deployment of federated analysis methods in this Master's thesis project has demonstrated the feasibility and effectiveness of conducting collaborative research on sensitive healthcare data while maintaining patient privacy. These methods will empower researchers to access and analyze data from multiple institutions, including the GCAT cohort and the COVICAT-CONTENT sub-cohort, leading to a more comprehensive understanding of chronic diseases and their risk factors, as well as enhancing their scientific visibility.

6.5. Data Discovery

The implementation of Beacon v2 in our synthetic database has been carried out as a proof-of-concept in the context of this Master's thesis project. Through this implementation, we have evaluated its functionality and demonstrated its ability to manage and query data efficiently.

To validate the correct implementation, we performed a series of queries and tests. A query applied to filter and count the number of male records yielded a *resultsCount* of 2587. This value perfectly

matched the expected number of male records in the database, confirming that the implementation of Beacon v2 has been successful. The accuracy of these results reinforces our confidence in the integrity and accuracy of the implemented system.

During the deployment process, we have gained valuable knowledge about the configuration and optimization of Beacon v2. This knowledge will be crucial when we scale up the use of this tool to the DATOS-CAT project, a larger and more ambitious project.

In conclusion, Beacon will be instrumental in the DATOS-CAT project for identifying and accessing relevant datasets. Beacon facilitates the discovery of datasets that are FAIR (Findable, Accessible, Interoperable, and Reusable). By adhering to FAIR principles, Beacon ensures that datasets are easily discoverable, accessible, and reusable by researchers, enabling efficient and effective data sharing and collaboration. Beacon therefore empowers researchers to carry out innovative research and contribute to the advancement of scientific knowledge.

Chapter 7. Ethical-social impact, sustainability and diversity

7.1. Ethical-social impact

The DATOS-CAT project, and therefore this Master's thesis, has significant ethical and social implications, particularly in the context of handling sensitive health data. Below are the key considerations:

- **Privacy and data security**. It is essential to ensure the protection of patient privacy when working with healthcare data. This involves implementing robust measures to anonymize and protect sensitive personal data at all levels, including tools and technologies, while complying with applicable laws and regulations, such as the **GDPR**. The project adheres to the GDPR regulations, which mandates strict guidelines for data protection and privacy, and has a Data Management Plan (DMP) available.

  Moreover, the use of DataSHIELD for data analysis and the use of Beacon v2 for data discovery, allows non-disclosive federated solutions. This means that the data remains at the source and only aggregated results are shared, minimizing the risks of data breaches.

- **Informed Consent**. All the GCAT cohort participants, as well as the one from the COVICANT-CONTENT sub-cohort, have provided informed consent, understanding that their data will be used for research purposes. The document includes detailed explanations of how data is collected, stored, and used, ensuring transparency and trust.

- **Open science**. The project maintains transparency by making its data management practices, methodologies, and tools publicly available.

- **Impact on Healthcare**. Although analysis is not the main objective of the project, it is to make the Catalan cohort visible. Making available data to the scientific community offers new research opportunities that may lead to a better understanding of chronic diseases and their risk factors, ultimately improving public health outcomes. Standardizing the data allows us to integrate the cohort with other standardized cohorts, such as the UKBiobank. In this way, it increases the ability to conduct comprehensive research, which can lead to new treatments and interventions.

  Therefore, creating an accessible and standardized data infrastructure can help reduce disparities in health care.

7.2. Sustainability

The sustainability of the DATOS-CAT project is crucial for its long-term success and impact. Key strategies include:

- **Long-term Data Management**. The project has a Data Management Plan, essential to implement robust data management practices. A DMP is key to the life cycle of research

data, not only to explain data collection and description, but also for data preservation and access. It ensures that the data will be properly documented and available to other researchers in the future for analysis purposes.
The DATOS-CAT DMP describes how data will be collected and managed during the project. It also describes how the data will be stored and reused in the future, after the project is completed. Two DTMs are planned. The first DTM was completed at the beginning of the project and another one will be drafted at the end project to address all updates and changes in data management.

- **Financial sustainability**. Implementing efficient data management processes and automation, such as the generation of data dictionaries, can reduce operational costs in the long term. In addition, by improving interoperability and data reuse, resources are optimized, and unnecessary duplications are avoided.
  Securing ongoing funding through grants, collaborations, and partnerships to support the project's activities and objectives.

- **Operational sustainability**. Developing a data infrastructure that is scalable and adaptable ensures that the project can evolve and sustain itself over time. Moreover, it allows the inclusion of additional cohorts and datasets in the future. We also ensure that the methodologies and tools developed can be adapted and used by other research projects worldwide.

  Adopting the described workflow for data management avoids duplication of efforts and unnecessary storage of data in multiple systems. This not only optimizes the use of technological resources, but also ensures that data is always organized and accessible in a single centralized repository.

- **Environmental sustainability**. Although environmental sustainability is not the primary focus in data-related projects, by minimizing data duplication and centralizing storage, the number of computational resources required is reduced. This translates into lower energy demand and a reduced carbon footprint associated with the operation of data centers and IT equipment.

7.3. Diversity

The DATOS-CAT project places a strong emphasis on diversity, recognizing the importance of inclusive research practices:

- **Inclusive Participant Recruitment.** The cohort ensured diverse representation of the Catalan population by recruiting participants from diverse backgrounds, genders, ethnicities, and socioeconomic status. One of the requirements was that age was between 40 and 65 years.

- **Multidisciplinary Collaboration**. The DATOS-CAT project involves collaboration between different institutions, bringing together experts from various fields to benefit from a wide range of perspectives and enrich research. This interdisciplinary approach facilitates the integration of diverse expertise, enabling innovative solutions to complex problems.

Through regular meetings and workshops, the DATOS-CAT project fosters continuous dialogue and exchange of ideas among its members. This ongoing interaction not only keeps the research dynamic and adaptive, but also builds a strong network of professionals committed to advancing the field of data science.

- **Training and Education**. From the very beginning, there has been a strong commitment to the exchange of knowledge and the development of skills through the organization and participation in workshops, seminars, and conferences, both nationally and internationally. These activities not only allow for mutual enrichment among researchers but also foster interdisciplinary collaboration and the development of new ideas and approaches.

  Our commitment to diversity in knowledge dissemination is reflected in active participation in various academic and scientific symposiums, where ideas are shared and discussed with the global scientific community. This approach has enabled the integration of diverse perspectives and fostered a continuous innovation environment.

  Furthermore, we have promoted the presentation of research findings in internationally recognized data communities. Appendix E includes posters presented at some of these prominent communities, highlighting the importance of visibility and communication of the results obtained in diverse and plural contexts.

  A significant example of our commitment to diversity and scientific outreach is our participation in the **#100tífiques** initiative. This project aims to highlight the work of women in science and promote role models for future generations. Through #100tífiques, we have been able to connect with a network of outstanding female scientists and contribute to the empowerment and representation of women in science.

  These efforts have significantly contributed to creating an inclusive and dynamic knowledge network, where the diversity of ideas and approaches strengthens the quality and impact of the research conducted.

By addressing these areas, the DATOS-CAT project aims to create a positive ethical, social, and environmental impact while fostering diversity and inclusivity in biomedical research.

Chapter 8. Future work

Looking forward, several key areas need to be addressed to meet the objectives of this Master's Thesis and the DATOS-CAT project.

Among the key areas for future work, the first action is to develop a comprehensive **data catalog** encompassing the data sets from both the GCAT cohort and the COVICAT-CONTENT sub-cohort. This catalog should be available on the Mica web data portal, following the established guidelines.

An important aspect related is ensuring the continuous updating and maintenance of the catalog. This includes integrating new data entries as follow-ups are generated, or as new data sources are added. Implementing automated processes for updating and verifying the data will enhance system reliability and reduce data inconsistencies. Moreover, expanding the web data portal's capabilities with more advanced search and filtering options, as well as improving the visual interface, could further streamline data exploration.

After completing the data catalog, we will proceed with **data standardization**. The standardization of the different sources to the OMOP CDM will be done with one of the approaches, the traditional ELT or the semantic ETL. We need to decide between both, meanwhile processes to automate both approaches will be done.

While the generation of **synthetic data** (from aggregated data) will not be used beyond the proof-of-concept for the DATOS-CAT project, it is essential within my research group to introduce some changes in the implementation, such as the addition of some constraints and rules. This synthetic data can be a valuable tool in the initial stages of other projects, where real data is not yet available, helping to establish the data flow and facilitate decision-making.

Deploying a platform that allows for federated data discovery and analysis across multiple datasets within the Catalan cohorts is a priority. Ensuring data privacy and security is paramount in this process. Since we are working within the OMOP environment, Beacon v2 with the OMOP CDM plugin can be very useful in facilitating **data discovery**. This tool will allow researchers to quickly identify relevant datasets and improve the efficiency of their studies. On the other hand, we observed that the OBiBa environment can be very useful for the development of the DATOS-CAT project. This environment not only facilitates the creation of a data catalog, but also allows for **federated data analysis** through DataSHIELD. The idea is that researchers will have access to a set of standardized OMOP resources, specifically for federated analysis using the dsOMOP package.

By addressing these areas, we aim to fulfill the specific objectives of the project, enhancing the implementation and use of technologies within the DATOS-CAT project and improving the interoperability and visibility of the Catalan population-based cohort.

Bibliography

[1] *Personalized medicine*. (n.d.). Genome.gov. https://www.genome.gov/genetics-glossary/Personalized-Medicine

[2] Olson, J., Bielinski, S., Ryu, E., Winkler, E., Takahashi, P., Pathak, J., & Cerhan, J. (2014). Biobanks and personalized medicine. *Clinical Genetics*, *86*(1), 50–55. https://doi.org/10.1111/cge.12370

[3] Sudlow, C., Gallacher, J. E., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., & Collins, R. (2015). UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Medicine*, *12*(3), e1001779. https://doi.org/10.1371/journal.pmed.1001779

[4] *GCAT Genomes for Life - GCAT*. (n.d.). http://www.gcatbiobank.org/

[5] Obón-Santacana, M., Vilardell, M., Carreras, A., Duran, X., Velasco, J., Galván-Femenía, I., Alonso, T., Puig, L., Sumoy, L., Duell, E. J., Perucho, M., Moreno, V., & De Cid, R. (2018). GCAT|Genomes for life: a prospective cohort study of the genomes of Catalonia. *BMJ Open*, *8*(3), e018324. https://doi.org/10.1136/bmjopen-2017-018324

[6] *Programa d'Analítica de Dades per a la Recerca i la Innovació en Salut*. (n.d.). Agència De Qualitat I Avaluació Sanitàries De Catalunya (AQuAS). https://aquas.gencat.cat/ca/fem/intelligencia-analitica/padris/index.html

[7] *The Impact of COVID-19 Will Be Explored through an Epidemiological Study on 24,000 Catalan Volunteers - ISGLOBAL*. (n.d.). ISGLOBAL. https://www.isglobal.org/en/-/un-estudio-epidemiologico-evalua-el-impacto-de-la-covid-19-en-una-poblacion-de-24-000-voluntarios-catalanes

[8] Kogevinas, M., Karachaliou, M., Espinosa, A., Aguilar, R., Castaño-Vinyals, G., Garcia-Aymerich, J., Carreras, A., Cortés, B., Pleguezuelos, V., Papantoniou, K., Rubio, R., Jiménez, A., Vidal, M., Serra, P., Parras, D., Santamaría, P., Izquierdo, L., Cirach, M., Nieuwenhuijsen, M., . . . Tonne, C. (2023). Long-Term exposure to air pollution and COVID-19 vaccine antibody response in a general population cohort (COVICAT study, Catalonia). *Environmental Health Perspectives*, *131*(4). https://doi.org/10.1289/ehp11989

[9] *Estudio multi-caso control poblacional, incluyendo tumores de alta incidencia en España*. (n.d.). https://www.mccspain.org/

[10] *INMA - Environment and Childhood Project*. (n.d.). ISGlobal. https://www.isglobal.org/en/-/proyecto-inma-infancia-y-medio-ambiente

[11] *HOME | ECRHS*. (n.d.). ECRHS. https://www.ecrhs.org/

[12] *URBAN training in COPD - ISGLOBAL*. (n.d.). ISGLOBAL. https://www.isglobal.org/en/-/urban-training-in-copd

[13] *LeRAgs - ISGLOBAL*. (n.d.). ISGLOBAL. https://www.isglobal.org/en/-/lerags

[14] *GCAT SUMMARY AGGREGATE DATA - GCAT*. (n.d.). http://www.gcatbiobank.org/investigadors/en_gcat-summary-aggregate-data/

[15] Gibney, E., & Van Noorden, R. (2013). Scientists losing data at a rapid rate. *Nature*. https://doi.org/10.1038/nature.2013.14416

[16] Korbmacher, M., Azevedo, F., Pennington, C. R., Hartmann, H., Pownall, M., Schmidt, K., Elsherif, M., Breznau, N., Robertson, O., Kalandadze, T., Yu, S., Baker, B. J., O'Mahony, A., Olsnes, J. Ø.-., Shaw, J. J., Gjoneska, B., Yamada, Y., Röer, J. P., Murphy, J., . . . Evans, T. (2023). The replication crisis has led to positive structural, procedural, and community changes. *Communications Psychology*, *1*(1). https://doi.org/10.1038/s44271-023-00003-2

[17] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J., Da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., . . . Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1). https://doi.org/10.1038/sdata.2016.18

[18] GO FAIR initiative. (2022, January 21). *FAIR principles - GO FAIR*. GO FAIR. https://www.go-fair.org/fair-principles/

[19] *What are the FAIR Principles?* (n.d.). https://faircookbook.elixir-europe.org/content/recipes/introduction/brief-FAIR-principles.html

[20] *Open science*. (2024, May 16). Research and Innovation. https://research-and-innovation.ec.europa.eu/strategy/strategy-2020-2024/our-digital-future/open-science_en

[21] *General Data Protection Regulation (GDPR) – legal text*. (2024, April 22). General Data Protection Regulation (GDPR). https://gdpr-info.eu/

[22] *Health Insurance Portability and Accountability Act of 1996 (HIPAA)*. (2024, May 15). Public Health Law. https://www.cdc.gov/phlp/php/resources/health-insurance-portability-and-accountability-act-of-1996-hipaa.html#:~:text=The%20Health%20Insurance%20Portability%20and,the%20patient's%20consent%20or%20knowledge

[23] Gaimster, H., PhD. (2023, June 20). The five benefits of federated data analysis. *lifebit*. https://www.lifebit.ai/federated-data-analysis/the-five-benefits-of-federated-data

[24] Data sharing in the age of deep learning. (2023). *Nature Biotechnology*, *41*(4), 433. https://doi.org/10.1038/s41587-023-01770-3

[25] Casaletto, J., Bernier, A., McDougall, R., & Cline, M. S. (2023). Federated Analysis for Privacy-Preserving Data Sharing: A Technical and Legal primer. *Annual Review of Genomics and Human Genetics*, *24*(1), 347–368. https://doi.org/10.1146/annurev-genom-110122-084756

[26] *Data harmonization*. (n.d.). https://www.icpsr.umich.edu/web/pages/DSDR/harmonization.html

[27] https://www.sciencedirect.com/topics/computer-science/data-standardization

[28] Bönisch, C., Kesztyüs, D., & Kesztyüs, T. (2022). Harvesting metadata in clinical care: a crosswalk between FHIR, OMOP, CDISC and openEHR metadata. *Scientific Data*, *9*(1). https://doi.org/10.1038/s41597-022-01792-7

[29] Kohler, S., Boscá, D., Kärcher, F., Haarbrandt, B., Prinz, M., Marschollek, M., & Eils, R. (2023). Eos and OMOCL: Towards a seamless integration of openEHR records into the OMOP Common Data Model. *Journal of Biomedical Informatics*, *144*, 104437. https://doi.org/10.1016/j.jbi.2023.104437

[30] Rinaldi, E., & Thun, S. (2021). From OpenEHR to FHIR and OMOP data model for microbiology findings. In *Studies in health technology and informatics*. https://doi.org/10.3233/shti210189

[31] [Outline of Health Level Seven (HL7) standard]. (1999, December 1). PubMed. https://pubmed.ncbi.nlm.nih.gov/10639827/

[32] *DATOS-CAT – Planes complementarios Salud*. (n.d.). https://planescomplementariossalud.es/datos-cat/

[33] *Home*. (n.d.). https://www.ga4gh.org/

[34] *Phenopackets*. (n.d.). https://www.ga4gh.org/product/phenopackets/

[35] *Beacon*. (n.d.). https://www.ga4gh.org/product/beacon-api/

[36] *OHDSI – Observational Health Data Sciences and Informatics*. (n.d.). https://ohdsi.org/

[37] *OMOP Common Data Model*. (n.d.). https://ohdsi.github.io/CommonDataModel/

[38] *European Open Science Cloud (EOSC)*. (2023, February 10). Research and Innovation. https://research-and-innovation.ec.europa.eu/strategy/strategy-2020-2024/our-digital-future/open-science/european-open-science-cloud-eosc_en#what-the-european-open-science-cloud-is

[39] *European Health Data Space*. (2024, April 24). Public Health. https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space_en

[40] *1 + million genomes – IMPACT*. (n.d.). https://impact.isciii.es/1-million-genomes/

[41] *European "1+ Million Genomes" initiative*. (2023, December 14). Shaping Europe's Digital Future. https://digital-strategy.ec.europa.eu/en/policies/1-million-genomes

[42] *1+MG Framework*. (n.d.). 1+MG Framework. https://framework.onemilliongenomes.eu/about-the-framework

[43] *Beyond One Million Genomes (B1MG) project*. (n.d.). https://b1mg-project.eu/

[44] *GDI project*. (n.d.). https://gdi.onemilliongenomes.eu/

[45] *ELIXIR*. (n.d.). ELIXIR. https://elixir-europe.org/

[46] *EGA European Genome-Phenome Archive - EGA European Genome-Phenome Archive*. (2017, May 1). The European Bioinformatics Institute (EMBL-EBI). https://ega-archive.org/

[47] Embl-Ebi. (n.d.). *What is EGA? | European Genome-phenome Archive*. https://www.ebi.ac.uk/training/online/courses/ega-quick-tour/what-is-ega/

[48] Freeberg, M. A., Fromont, L. A., D'Altri, T., Romero, A. F., Ciges, J. I., Jene, A., Kerry, G., Moldes, M., Ariosa, R., Bahena, S., Barrowdale, D., Barbero, M. C., Fernandez-Orth, D., Garcia-Linares, C., Garcia-Rios, E., Haziza, F., Juhasz, B., Llobet, O. M., Milla, G., . . . Rambla, J. (2021). The European Genome-Phenome Archive in 2021. *Nucleic Acids Research*, *50*(D1), D980–D987. https://doi.org/10.1093/nar/gkab1059

[49] EHDEN. (2024, February 20). *European Health Data Evidence Network – ehden.eu*. ehden.eu. https://www.ehden.eu/#

[50] Cordis, C. (2019, January 10). *European Health Data and Evidence Network*. CORDIS | European Commission. https://cordis.europa.eu/project/id/806968

[51] *Plan Estratégico – IMPACT*. (n.d.). https://impact.isciii.es/plan-estrategico/

[52] *IMPaCT - Data | IMPaCT-Data. Infraestructura de Medicina de Precisión asociada a la Ciencia y la Tecnología*. (n.d.). https://impact-data.bsc.es/about/impact-data/

[53] *Planes complementarios Salud*. (n.d.). https://planescomplementariossalud.es/

[54] *Plan Estratégico IMPACT*. (n.d.). https://www.isciii.es/QueHacemos/Financiacion/IMPaCT/Paginas/Plan.aspx

[55] *About INB | INB*. (n.d.). https://inb-elixir.es/about-inb

[56] *UK Biobank - UK Biobank*. (n.d.). https://www.ukbiobank.ac.uk/

[57] Sudlow, C., Gallacher, J. E., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., & Collins, R. (2015b). UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Medicine*, *12*(3), e1001779. https://doi.org/10.1371/journal.pmed.1001779

[58] Papez, V., Moinat, M., Voss, E. A., Bazakou, S., Van Winzum, A., Peviani, A., Payralbe, S., Kallfelz, M., Asselbergs, F. W., Prieto-Alhambra, D., Dobson, R. J. B., & Denaxas, S. (2022). Transforming and evaluating the UK Biobank to the OMOP Common Data Model for COVID-19 research and beyond. *Journal of the American Medical Informatics Association*, *30*(1), 103–111. https://doi.org/10.1093/jamia/ocac203

[59] *Cohorte impact*. (n.d.). https://cohorte-impact.es/

[60] *EuCanSHare*. (n.d.). euCanSHare. http://www.eucanshare.eu/

[61] OBiBa. (n.d.). *OBIBA*. https://www.obiba.org/

[62] *What is DataSHIELD?* (2022, October 31). DataSHIELD. https://www.datashield.org/about/about-datashield-collated

[63] *Coral Docker Stack Deployment*. (n.d.). https://www.eucanconnect.eu/coral-docker-stack-deployment/

[64] *Building a catalogue of datasets*. (n.d.). https://faircookbook.elixir-europe.org/content/recipes/infrastructure/data-catalog.html

[65] Ohdsi. (n.d.). *GitHub - OHDSI/WhiteRabbit:* https://github.com/OHDSI/WhiteRabbit

[66] *WhiteRabbit for ETL design – OHDSI*. (n.d.). https://www.ohdsi.org/analytic-tools/whiterabbit-for-etl-design/

[67] *DATOS-CAT*. (n.d.). GitHub. https://github.com/DATOS-CAT

[68] *Resources — Opal documentation*. (n.d.). https://opaldoc.obiba.org/en/dev/resources.html

[69] *Documents — Mica documentation*. (n.d.). https://micadoc.obiba.org/en/latest/documents.html

[70] *OMOP CDM Background*. (n.d.). https://ohdsi.github.io/CommonDataModel/background.html

[71] *Meltano Documentation | Meltano Documentation*. (n.d.). https://docs.meltano.com/

[72] *PostgreSQL: about*. (n.d.). The PostgreSQL Global Development Group. https://www.postgresql.org/about/

[73] Ohdsi. (n.d.-a). *GitHub - OHDSI/CommonDataModel: Definition and DDLS for the OMOP Common Data Model (CDM)*. GitHub. https://github.com/OHDSI/CommonDataModel

[74] *Athena*. (n.d.). https://athena.ohdsi.org/search-terms/start

[75] *Software Tools – OHDSI*. (n.d.). https://www.ohdsi.org/software-tools/

[76] *Rabbit in a hat*. (n.d.). https://ohdsi.github.io/WhiteRabbit/RabbitInAHat.html

[77] *About dbt run command | dbt Developer Hub*. (2024, June 4). https://docs.getdbt.com/reference/commands/run

[78] Pedrera, M., Garcia, N., Rubio, P., Cruz, J. L., Bernal, J. L., & Serrano, P. (2021). Making EHRs reusable: A common framework of data operations. In *Studies in health technology and informatics*. https://doi.org/10.3233/shti210831

[79] InformaticaClinica. (n.d.). *GitHub - InformaticaClinica/OntoBridge*. GitHub. https://github.com/InformaticaClinica/OntoBridge

[80] *Introduction | Ontop*. (n.d.). https://ontop-vkg.org/guide/

[81] *RDF - Semantic Web Standards*. (n.d.). https://www.w3.org/RDF/

[82] *R2RML: RDB to RDF Mapping Language*. (n.d.). https://www.w3.org/TR/r2rml/

[83] *Apache jena - Apache jena fuseki*. (n.d.). https://jena.apache.org/documentation/fuseki2/

[84] *SPARQL 1.1 Query language*. (n.d.). https://www.w3.org/TR/sparql11-query/

[85] Roebuck, Kevin. 2012. Data Quality: High-Impact Strategies-What You Need to Know: Definitions, Adoptions, Impact, Benefits, Maturity, Vendors. Emereo Publishing.

[86] Ohdsi. (n.d.-b). *GitHub - OHDSI/DataQualityDashboard: A tool to help improve data quality standards in observational data science.* GitHub. https://github.com/OHDSI/DataQualityDashboard

[87] *6 The Resources | Orchestrating privacy-protected big data analyses of data from different resources with R and DataSHIELD*. (2021, January 19). https://isglobal-brge.github.io/resource_bookdown/resources.html#resources

[88] Isglobal-Brge. (n.d.). *GitHub - isglobal-brge/dsOMOP*. GitHub. https://github.com/isglobal-brge/dsOMOP

[89] Isglobal-Brge. (n.d.-b). *GitHub - isglobal-brge/dsOMOPClient: To be supplied*. GitHub. https://github.com/isglobal-brge/dsOMOPClient

[90] Isglobal-Brge. (n.d.-c). *GitHub - isglobal-brge/dsOMOPHelper: To be supplied*. GitHub. https://github.com/isglobal-brge/dsOMOPHelper

[91] Developers, M. B. L. F. &. B. (n.d.). *Beacon v2 Project Website*. https://beacon-project.io/

[92] *IMPaCT-Data / impd-beacon_omopcdm · GitLab*. (n.d.). GitLab. https://gitlab.bsc.es/impact-data/impd-beacon_omopcdm

[93] *aiosql*. (2024, May 30). PyPI. https://pypi.org/project/aiosql/

[94] *Welcome to AIOHTTP — aiohttp 3.9.5 documentation*. (n.d.). https://docs.aiohttp.org/en/stable/

[95] *ydata-profiling*. (2024, May 7). PyPI. https://pypi.org/project/ydata-profiling/

[96] MIE2024. (2024, June 5). *Home - MIE2024*. MIE2024 - Congress of the European Federation for Medical Informatics. https://mie2024.org/

Appendices

Appendix A. Opal deployment

In this section, we explain in detail the steps for creating a project and uploading the different dictionaries as tables in Opal.

1. Open a web browser and access the Opal URL (e.g. https://coral.bsc.es/repo). To log in you need administrator credentials.

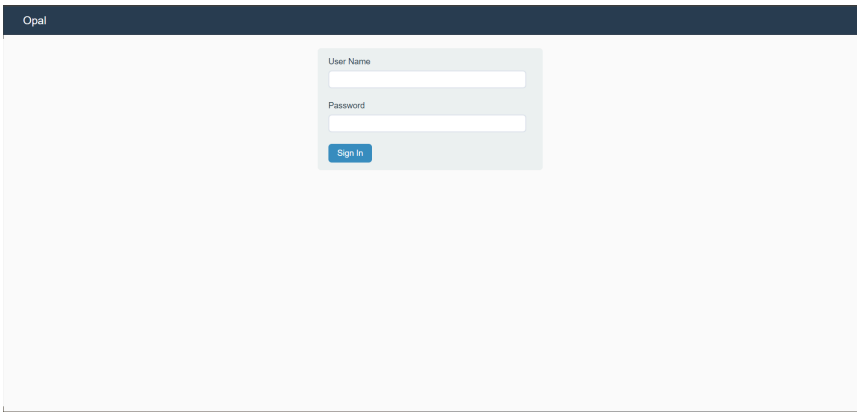


Figure 60. *Opal homepage where you are asked for credentials.*

2. Once logged in, you will be redirected to the home screen.

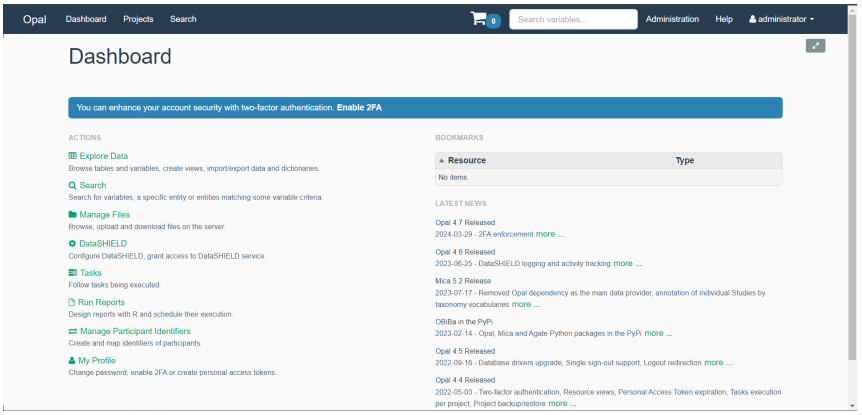


Figure 61. *Opal administration main page.*

3. Select the "Projects" tab to view existing projects and create new ones. When you click on the "+ Add Project" button, a pop-up window will appear as in Figure 62, asking you to enter information about the project, such as the name, title, database to store the information, etc. Once you have completed the content, click "Save".

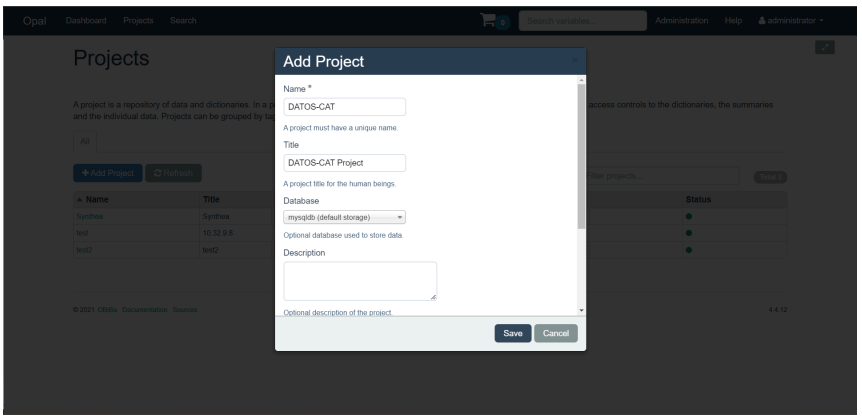


Figure 62. *Pop-up window for adding a new project in Opal.*

4. Once created, you will be redirected to the project home screen.

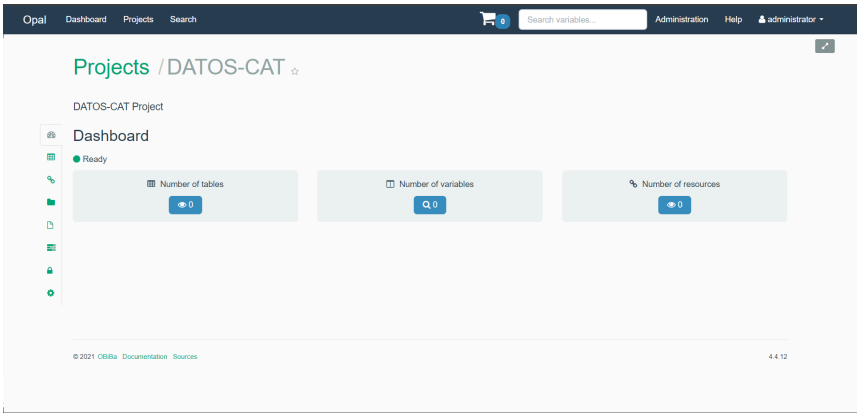


Figure 63. *DATOS-CAT project home screen in Opal.*

5. To upload the data dictionaries, go to the "Tables" section and click on "+ Add Table" button. A drop-down will appear, and you should select "Add/update tables from dictionary…".

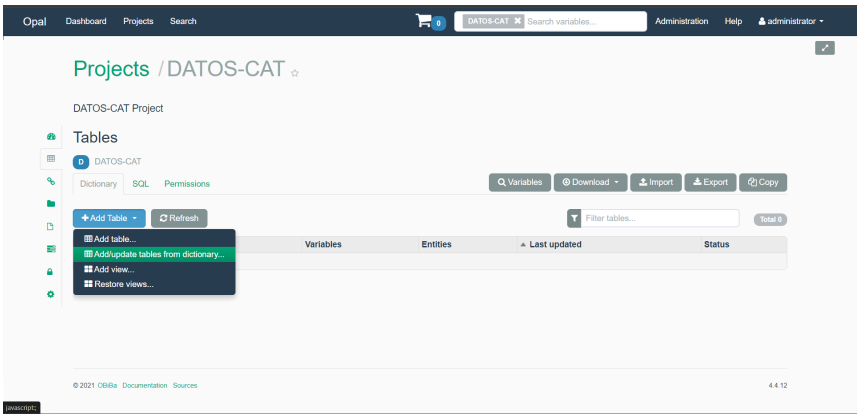


Figure 64. *DATOS-CAT project home screen in Opal II.*

6. A pop-up window will appear where you can download an Excel template (we have used this as a guide to create the data dictionary). In addition, you can select and upload files from your local system. Remember that the file must be in Excel or CSV format.

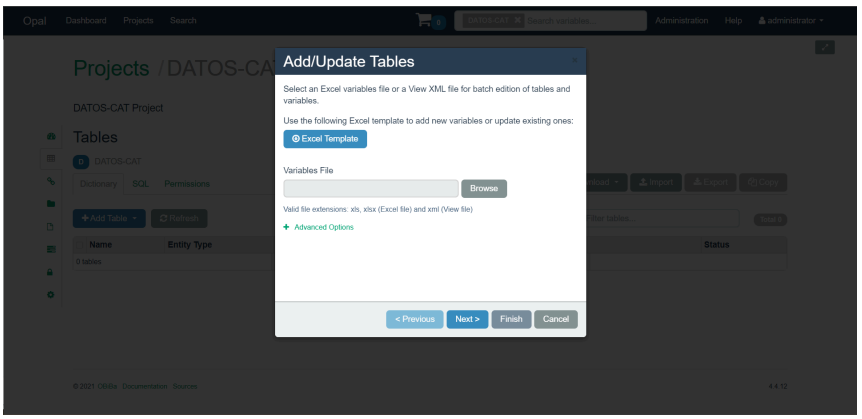


Figure 65. *Pop-up window for adding/updating tables from a dictionary in Opal.*

7. Once you select the data dictionaries, you can select the file you wish to upload.

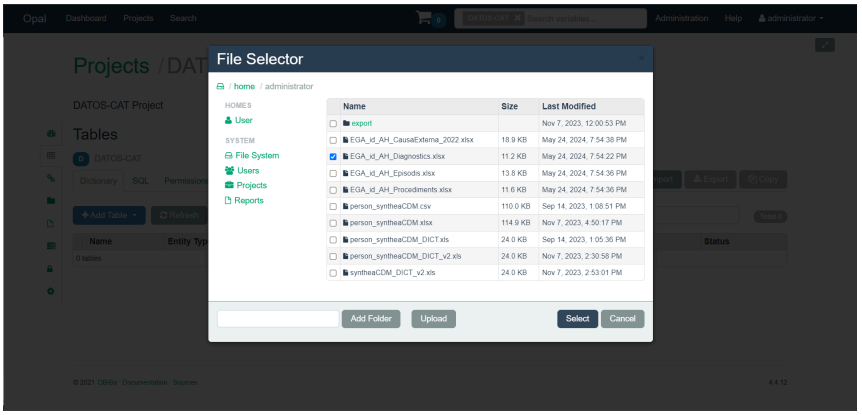


Figure 66. *Pop-up window for adding/updating tables from a dictionary in Opal II.*

8. There are some configuration options for the file you are uploading. Once you have configured the options, click the “Finish” button.

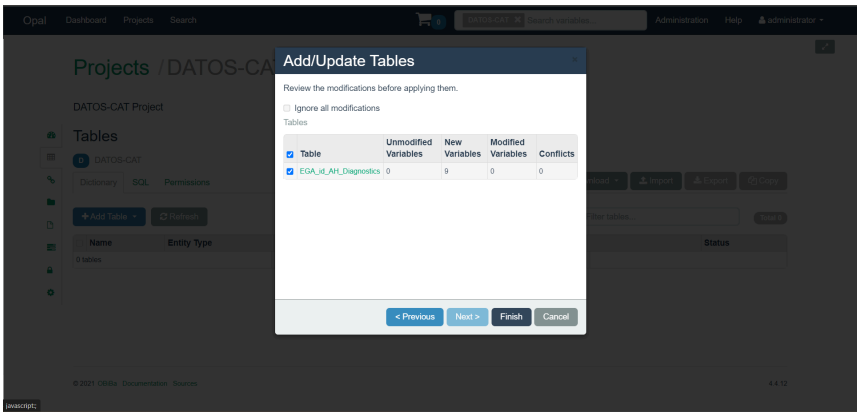


Figure 67. *Pop-up window for adding/updating tables from a dictionary in Opal III.*

9. Once created, you will be redirected to the project tables view, where you will see the data dictionary uploaded as a table.

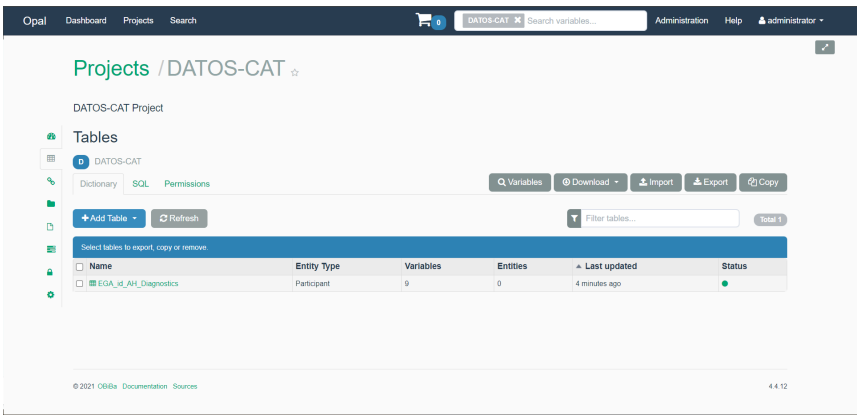


Figure 68. *DATOS-CAT project home screen with a table uploaded in Opal.*

10. The process can be repeated as many times as necessary.

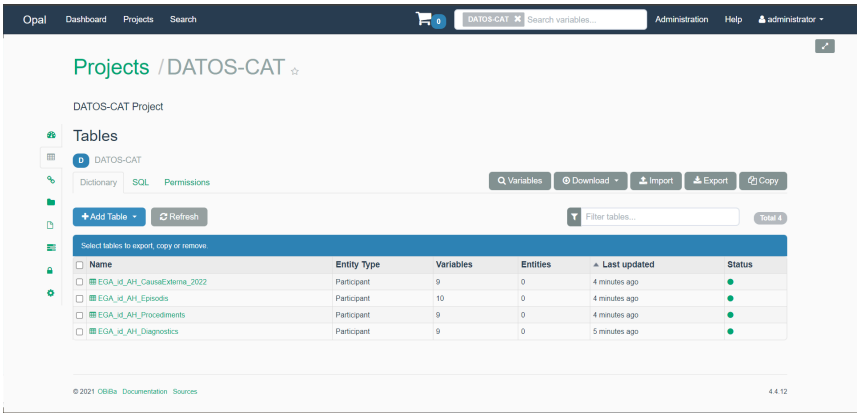


Figure 69. *DATOS-CAT project home screen with different tables uploaded in Opal.*

Appendix B. Mica deployment

In this section, we explain in detail the steps to create the web data portal. It is important to mention that it is the continuation of the previous section.

1. Open a web browser and access the Mica URL (e.g. https://coral.bsc.es/pub).

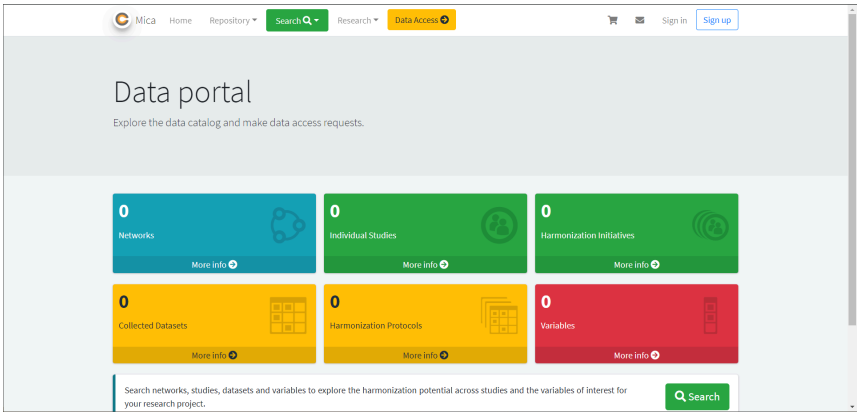


Figure 70. *Main screen of the web data portal in Mica.*

2. As the URL redirects to the main screen of the web data portal, you need to log in with administrator credentials to configure the catalog.

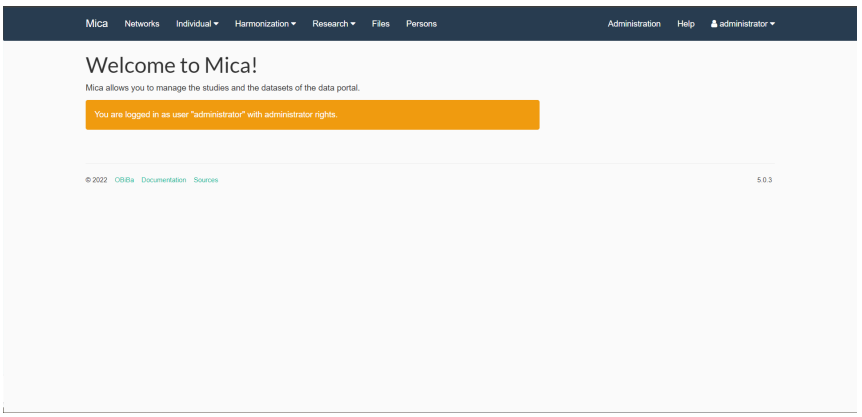


Figure 71. *Main screen of the administration part of Mica.*

3. **Create a network**. To create a network, choose the “Networks” option and select “+ Add Network”. By doing so, a new screen opens, allowing you to enter the information of the new network, such as the name, acronym, description, etc. Once completed, click on the "Save" button to create and apply the new network configuration.

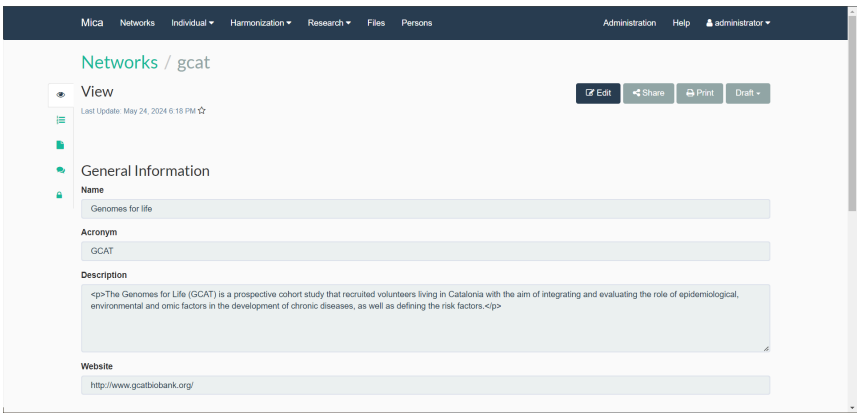


Figure 72. *GCAT Network draft view in Mica.*

4. **Individual Studies.** Go to the "Individual" tab, select "Individual Studies" and click on "+ Add Study". This will redirect to a new window, where you should also provide all the information regarding the study. Once created, you can link it to a Network if necessary.

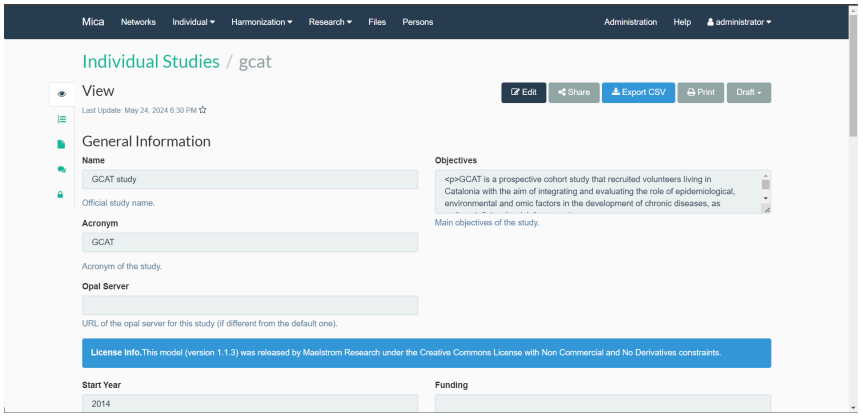


Figure 73. *GCAT Individual Studies draft view in Mica.*

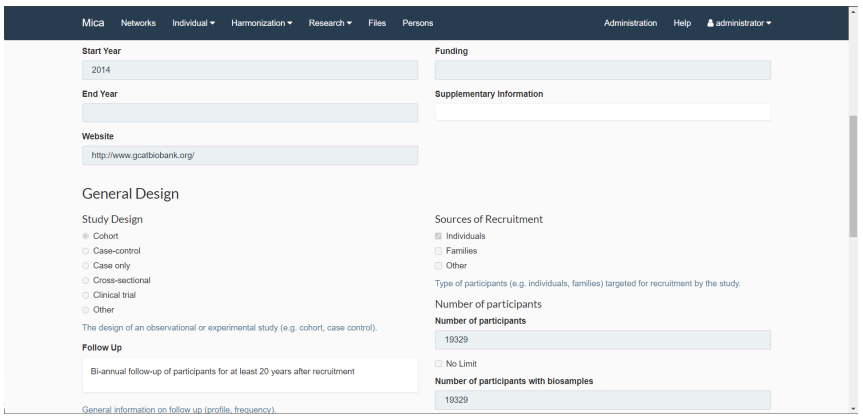


Figure 74. *GCAT Individual Studies draft view in Mica II.*

5. **Population.** At the end of any individual study, you could add different population groups. Click on the "+ Add Population" button and complete the required information.

6. **Collected Datasets**. To create a collected dataset, go to the "Individual" tab, select "Collected Datasets" and click on "+ Add Dataset". Once the collected dataset is created, you should link it to: (i) a Study, (ii) a Population group, and (iii) a Data Collection Event.

Moreover, to create the data catalog you should also link it to: (i) an Opal project and (ii) a table.

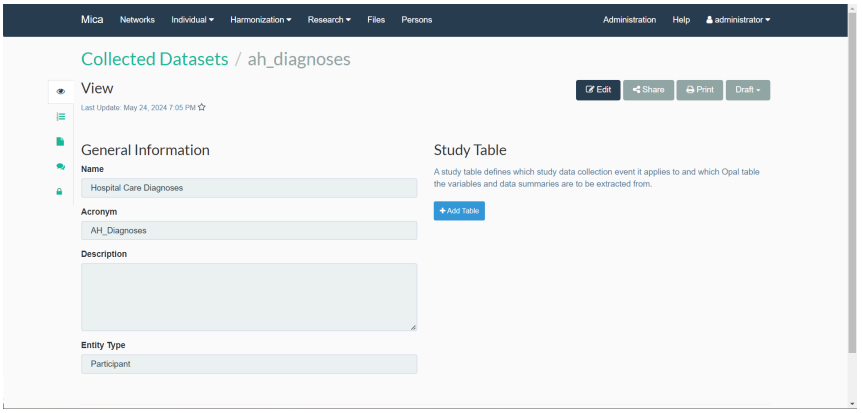


Figure 75. *Diagnosis AH Collected Dataset draft view in Mica.*

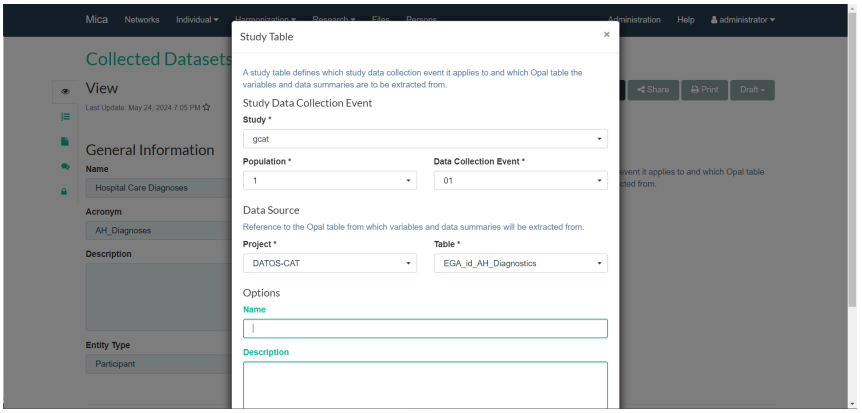


Figure 76. *Pop-up window to link Study Table in the Diagnosis AH Collected Dataset in Mica.*

11. Once you have created the entire scheme and established all the relationships, you can publish the information and it will appear on Mica's home screen.

Appendix C. Athena vocabularies

In this section, we explain in detail the steps to follow to download the vocabularies from Athena, necessary for the transformation from raw data to OMOP CDM.

1. Open a web browser and access the Athena URL (https://athena.ohdsi.org/).

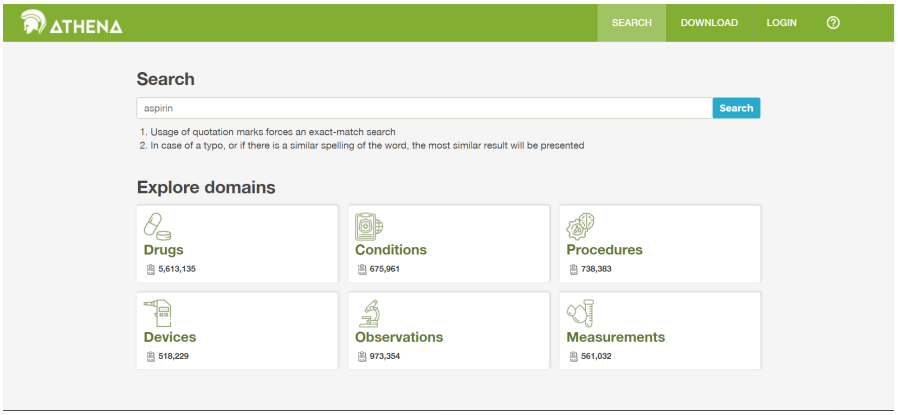


Figure 77. *Athena main page*

2. You must log in or register to have access.

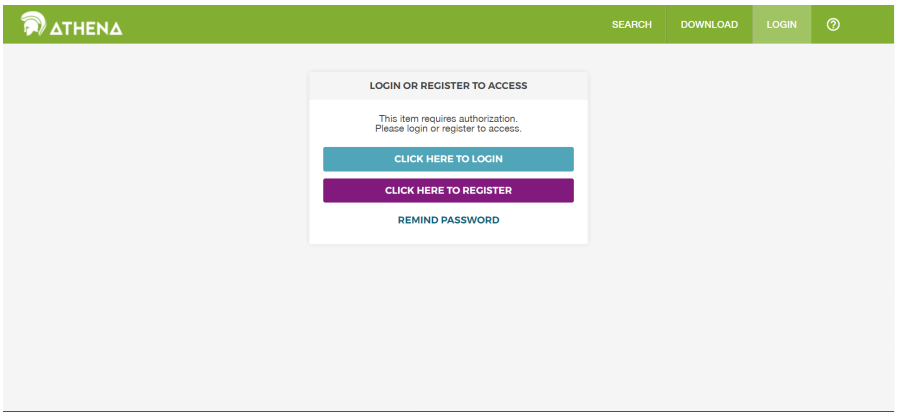


Figure 78. *Athena login page*

3. Once logged in, go to the "Download" tab at the top right and there you can select all the vocabularies you need for the transformation. After that, press the "Download vocabularies" button.

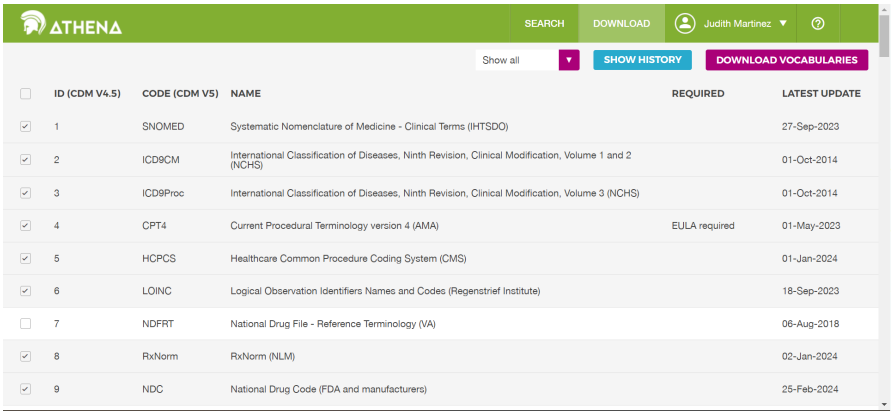


Figure 79. *Download tab in Athena for selecting vocabularies*

4. The download is not automatic, you must wait for an email with a link to download the standardized vocabularies.

Appendix D. DataSHIELD

In this section, we explain in detail the steps to follow to add dsOMOP package in the Opal server-side and how to upload a database as a resource.

1. **Add dsOMOP in the server-side packages.** Go to the "Administration" tab and select "DataSHIELD". In this view, there is a section titled "Packages". Click the "+ Add Package" button and complete the information related to the Package. After that, click the "Install" button.

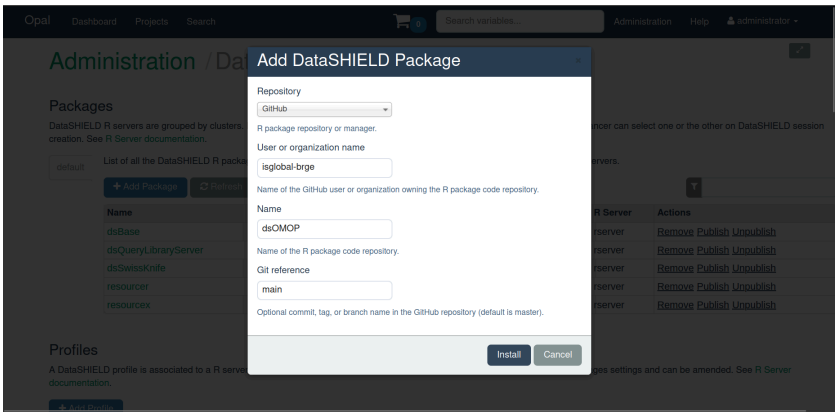


Figure 80. *Pop-up to add dsOMOP package in the DataSHIELD Administration part of Opal*

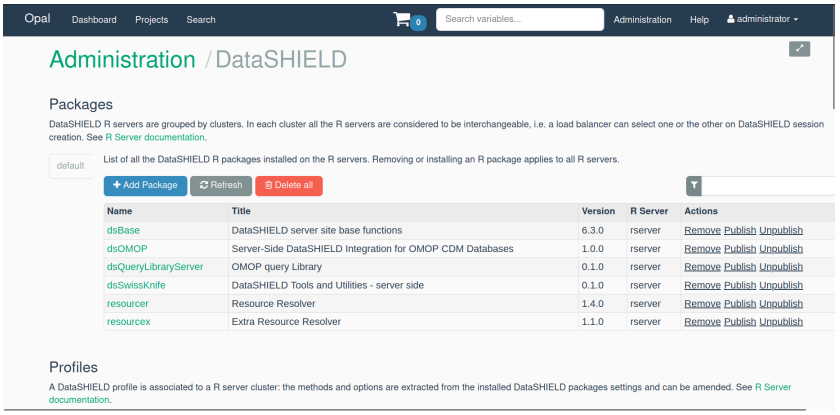


Figure 81. *DsOMOP package added in the DataSHIELD Administration part of Opal*

2. **Add *datoscat_tfm* PostgreSQL as a resource**.
    1. Go to the "Projects" tab and select the project, in this case, "DATOS-CAT".
    2. Go to the "Resources" section and click the "+ Add Resource" button.

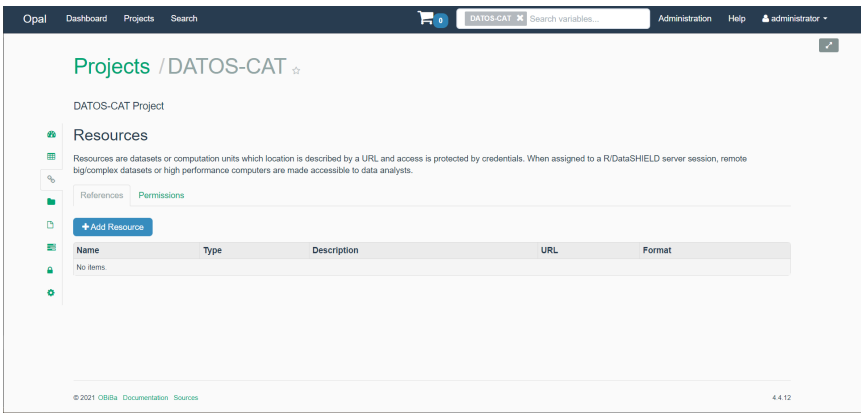


Figure 82. *DATOS-CAT project resource view in Opal*

3. A pop-up window will appear, where all the information related to the PostgreSQL database should be completed. In order to use dsOMOP, the "OMOP CDM" should be chosen as Category. There is a "Credentials" tab, where the database username and password should be specified in order to connect to the database.

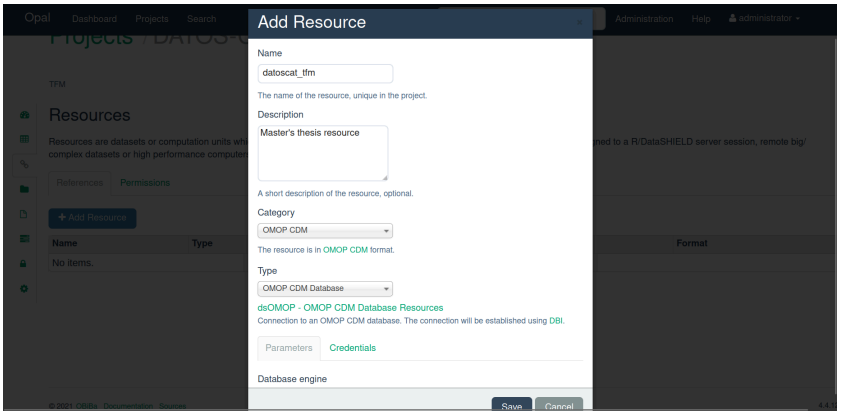


Figure 83. *Adding the datoscat_tfm PostgreSQL resource in Opal*

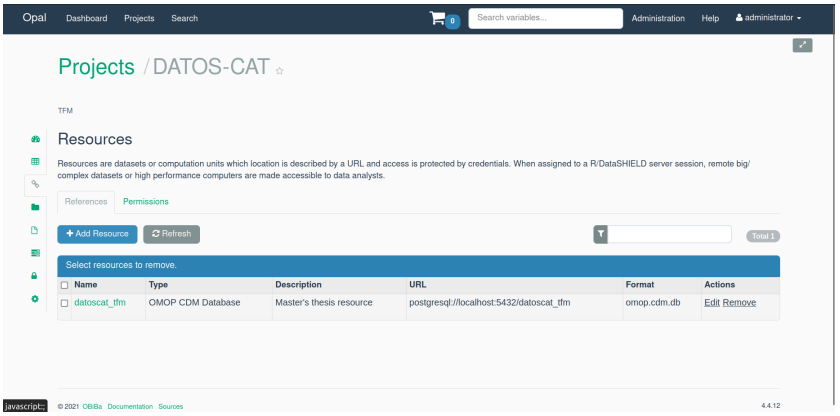


Figure 84. *Added the datoscat_tfm PostgreSQL as a resource in the DATOS-CAT project in Opal*

Appendix E. Communication

This appendix contains papers presented at conferences that complement the main content of this work. These papers represent broader contributions to their respective fields of study and have been presented at various academic conferences and scientific events listed below.

- ***SWAT4HCLS 2024***, February 26-29, 2024.
  (https://www.swat4ls.org/)

  At the SWAT4HCLS conference, a poster of the DATOS-CAT project entitled "*DATOS-CAT: Methodologies for the standardization, integration and analysis of population-based biomedical data using semantic technologies" was presented*.

- ***OHDSI Europe Symposium 2024***, June 1-3, 2024.
  (https://www.ohdsi-europe.org/symposium-event)

  At the OHDSI Europe Symposium conference, a poster of the DATOS-CAT project entitled "*DATOS-CAT: Leveraging OMOP CDM for the standardization, integration and analysis of population-based biomedical data" was presented.*

- ***All Hands ELIXIR 2024***, June 10-12, 2024.
  (https://elixir-events.eventscase.com/EN/ahm2024)

  At the All Hands ELIXIR conference, a poster from the DATOS-CAT project entitled "*DATOS-CAT: Leveraging ELIXIR ecosystem for the standardization, integration and federated analysis of population-based biomedical data"* will be presented.

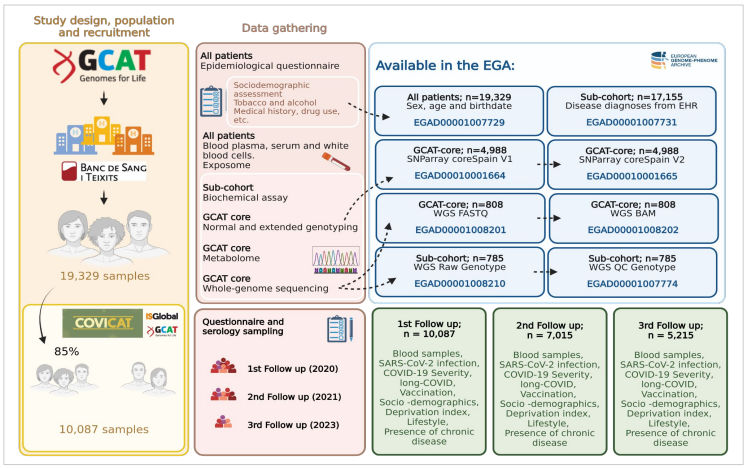
These communications address a variety of relevant topics and reflect a continued commitment to academic outreach and engagement with the scientific community.

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

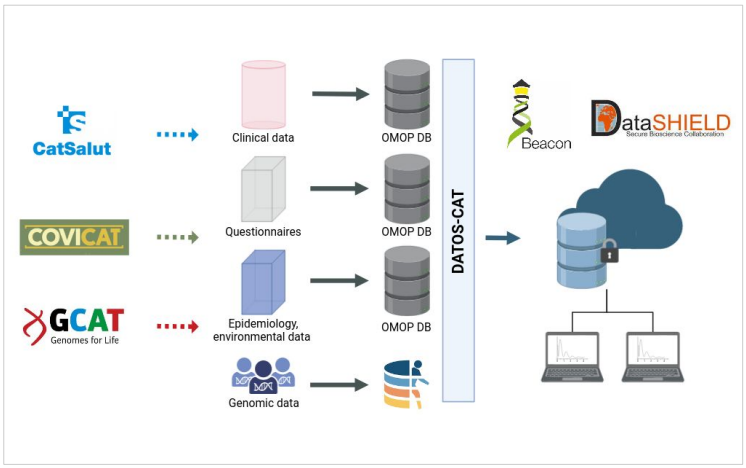
Judith Martinez-Gonzalez [1,2], Guillem Bracons Cucó [1], Aikaterini Lymperidou [1,3], Xavier Escribà-Montagut [4], Marta Huertas [1,5], Ramon Mateo-Navarro [1,4], David Sarrat-González [4], Santiago Frid [6], Rafael de Cid [3], Juan R González [4] , and Alberto Labarga [2]

[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.

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, Reusable) 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.

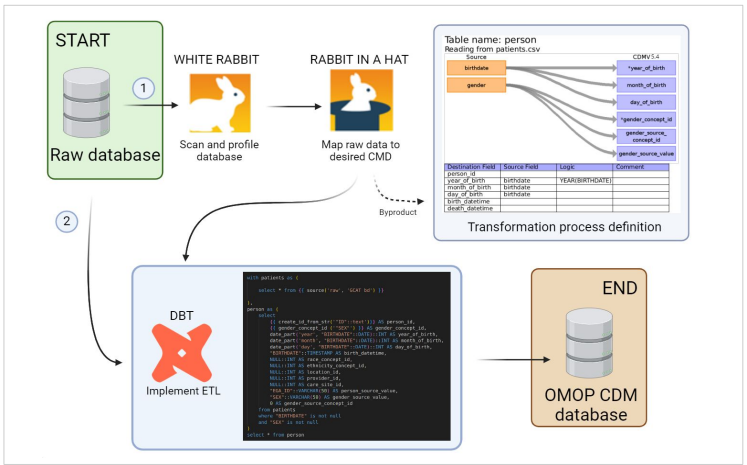


The GCAT Cohort is a prospective cohort study that has been following 20.000 volunteers living in Catalunya since 2014. The GCAT Study is a long-term genomic, environmental and lifestyle cohort project that aims to evaluate and track multiple pathologies, as well as biologically related traits. Therefore, the GCAT Study offers a unique opportunity to integrate diverse data to allow the identification of novel relations among different biomarkers and conditions.



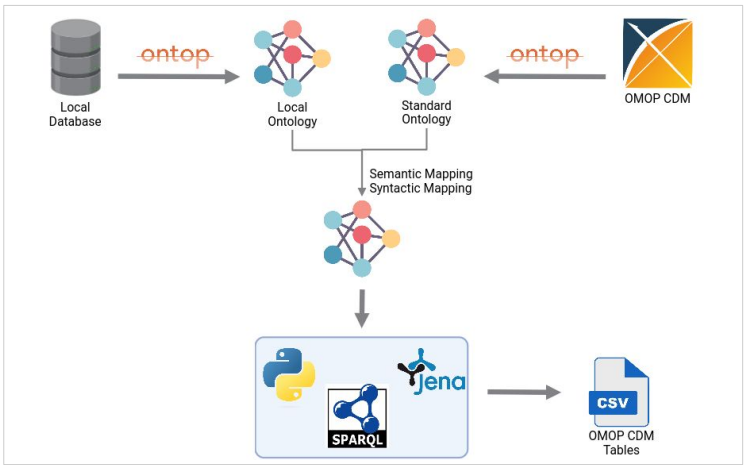
The OMOP Common Data Model is intended to be used as a standardized framework for clinical data representation and analysis. Building an OMOP-CDM database usually requires Extract-Transform-Load (ETL) processes that facilitate data conversion to a standard model and terminology, ensuring data consistency and integrity through the integration process. In the context of the DATOS-CAT project, we have evaluated two different ETL approaches: a traditional ETL and a semantic one.

Traditional ETL



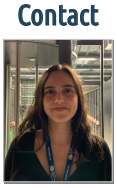
The traditional ETL process involves extracting general information from the database using the White Rabbit software, followed by a phase of mapping variables to the corresponding tables and columns of the OMOP-CDM using Rabbit-in-a-Hat (1). Based on the resulting mapping, together with the corresponding vocabulary mappings, a specific SQL-based transformation process is implemented to transform the raw data into the desired format (2), which will then be inserted into the OMOP-CDM compliant database.

Semantic ETL



OntoBridge is an ontology-based tool created by Hospital Clínic de Barcelona that transforms local databases to CDMs. It uses Ontop to replicate relational data into RDFs through R2RML mappings, and then inserts the triples into an ontology that represents the local data model. Local concepts are mapped to standard ones by means of the owl:sameAs property to allow semantic interoperability, while the syntactic equivalence is performed by defining local properties to be instances of metaclasses that model attributes of 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 .

Contact



Judith Martinez-Gonzalez
(judith.martinez@bsc.es)

Funding

Financiado por la Unión Europea NextGenerationEU
Plan de Recuperación, Transformación y Resiliencia
Next Generation Catalunya
Generalitat de Catalunya

Financiado por el "Plan Complementario de Biotecnología aplicada a la Salud", coordinado por el Institut de Bioenginyeria de Catalunya (IBEC) en el marco del Plan de Recuperación, Transformación y Resiliencia (C17.I1) - Financiado por la Unión Europea - NextGenerationEU

DATOS-CAT: Leveraging ELIXIR ecosystem for the standardization, integration and federated analysis of population-based biomedical data

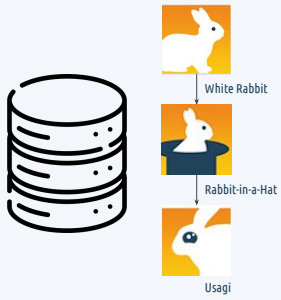
Judith Martinez-Gonzalez [1,2], David Sarrat-González [3], Ramon Mateo-Navarro [1,3], Guillem Bracons Cucó [1], Aikaterini Lymperidou [1,4,8], Marta Huertas [1,5], Sergi Aguiló-Castillo [2,6] Salvador Capella-Gutiérrez [2,6], Rafael de Cid [4,6,8], Santiago Frid [7], Juan R González [3] and Alberto Labarga [2,6]

[1] Institute for Bioengineering of Catalonia (IBEC). [2] Barcelona Supercomputing Center (BSC). [3] Barcelona Institute for Global Health (ISGlobal). [4] Genomes for Life- GCAT lab-Germans Trias i Pujol Research Institute (IGTP). [5] Centre for Genomic Regulation (CRG). [6] Spanish National Bioinformatics Institute (INB/ELIXIR-ES). [7] Hospital Clínic de Barcelona. [8] Grup de Recerca en Impacte de les Malalties Cròniques i les seves Trajectòries (GRIMTra).

Extract

The GCAT Cohort is a prospective cohort study that includes 19,329 volunteers from Catalonia. Participants complete a self-administered computer-based questionnaire that collects data on a large number of lifestyle and health-related factors. Detailed epidemiological and anthropometric measurements, and the electronic health records (EHR) of the Catalan Public Healthcare System are also collected. Blood plasma, blood serum and white blood cells are extracted, and their genomic profile has been characterized by comprehensive genotyping and whole genome sequencing. **Since 2023, GCAT is part of the INB/ELIXIR-ES resources portfolio.**

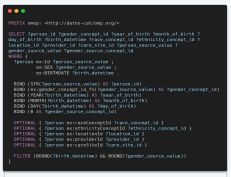
Load



Raw data has been loaded into a Postgres database and characterized using OHDSI open source tools. White Rabbit was used to generate a database profile. Rabbit-in-a-Hat to design the ELT and Usagi for mapping the medical terminologies in the original data to OMOP *concept_ids*.

Transform

For data transformation, two different approaches were evaluated: traditional SQL-based ELT using DBT and a novel approach using a tool called OntoBridge, which models data sources, CDMs like OMOP and their mappings by means of ontologies. Here, we use the outputs of White Rabbit and Usagi to model data sources, and use SPARQL queries to extract data as OMOP-compliant tables.



Ontobridge: SPARQL

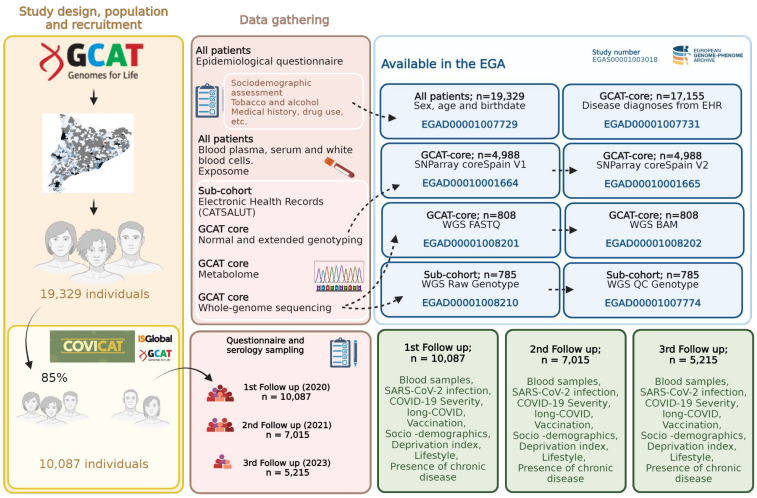
DBT: SQL

https://github.com/InformaticaClinica/OntoBridge/

OHDSI tools

OMOP CDM DB

GCAT Cohort

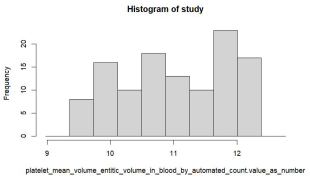


Federated data analysis: dsOMOP

The **dsOMOP package** is designed to facilitate the interaction with remote databases formatted in the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) from within a DataSHIELD environment. DataSHIELD allows statistical analyses to be performed on individual-level data, but without being transferred from their original location.

```
o$auto(tables = c("measurement"),
       concepts = c(3043111),
       columns = c("value_as_number"))

ds.table("omop$platelet_mean_volume_entitic_volume_in_blood_by_automated_count
.value_as_number")
```



https://github.com/isglobal-brge/dsOMOP

Federated data discovery: G4GH Beacon v2

Beacon

The **GA4GH Beacon v2** enables the federated discovery of data sources that meet specific genomic or phenotypic criteria. We developed a **plugin** that allows integrating **OMOP CDM** databases into the Beacon ecosystem.

https://gitlab.bsc.es/impact-data/impd-beacon_omopcdm

```
curl -s http://omop.beacon.bsc.es/api/individuals | jq
'[.response.resultSets[].results[].measures[] | select(.assayCode.id ==
"LOINC:32623-1") | {id: .assayCode.id, label: .assayCode.label, value:
.measurementValue.value, unit:
.measurementValue.unit.id}]'_mean_volume_entitic_volume_in_blood_by_automate
d_count.value_as_number)

[{
    "id": "LOINC:32623-1",
    "label": "Platelet mean volume [Entitic volume] in Blood by Automated
count",
    "value": 10.8,
    "unit": "UCUM:fL"
  },
  {
    "id": "LOINC:32623-1",
    "label": "Platelet mean volume [Entitic volume] in Blood by Automated
count",
    "value": 9.6,
    "unit": "UCUM:fL"
  }...]
```

[1] Obón-Santacana, M. et al. GCAT|Genomes for life: a prospective cohort study of the genomes of Catalonia. BMJ Open 8, 2018

[2] Rambla J. et al. Beacon v2 and Beacon networks: A "lingua franca" for federated data discovery in biomedical genomics, and beyond. Hum. Mutat., 43, 791–799, 2022.

[3] Marcon, Y. et. al Orchestrating privacy-protected big data analyses of data from different resources with R and DataSHIELD. PLoS computational biology, 17(3), 2021.

Code available at http://github.com/DATOS-CAT

Funding

Financiado por la Unión Europea NextGenerationEU

Gobierno de España · Ministerio de Ciencia e Innovación

Plan de Recuperación, Transformación y Resiliencia

Next Generation Catalunya

Generalitat de Catalunya

Financiado por el "Plan Complementario de Biotecnología aplicada a la Salud", coordinado por el Institut de Bioenginyeria de Catalunya (IBEC) en el marco del Plan de Recuperación, Transformación y Resiliencia (C17.I1) - Financiado por la Unión Europea - NextGenerationEU