Universidade de Aveiro Departamento de Eletrónica,
Telecomunicações e Informática
2016

**Tiago Miguel
Vigário Novo**

Arquitetura para Integração e Exploração de Registos Eletrónicos de Saúde

Architecture for Integration and Exploration of Eletronic Health Records

Universidade de Aveiro Departamento de Eletrónica, Telecomunicações e Informática
2016

Tiago Miguel Vigário Novo

Arquitetura para Integração e Exploração de Registos Eletrónicos de Saúde

Architecture for Integration and Exploration of Eletronic Health Records

"*And if anyone happens to ask whether I made any material difference to the welfare of this planet, you can tell them I came and went like a summer cloud.*"

— Fifth Doctor

Universidade de Aveiro Departamento de Eletrónica,
2016 Telecomunicações e Informática

Tiago Miguel
Vigário Novo

Arquitetura para Integração e Exploração de Registos Eletrónicos de Saúde

Architecture for Integration and Exploration of Eletronic Health Records

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Engenharia de Computadores e Telemática, realizada sob a orientação científica do Doutor José Luís Guimarães Oliveira, Professor associado do Departamento de Eletrónica, Telecomunicações e Informática da Universidade de Aveiro.

The final countdown begun
And now its time for one last
bow
Like all your other selves
Elevens hour is over now
The clock is striking twelves

Eric Ritchie Junior

o júri / the jury

presidente / president

Prof. Dr. Joaquim Manuel Henriques de Sousa Pinto
Professor Auxiliar, Departamento de Eletrónica, Telecomunicações e Informática da Universidade de Aveiro

vogais / examiners committee

Prof. Dr Joel Perdiz Arrais
Professor Auxiliar Convidado, Departamento de Engenharia Informática da Fac. de Ciências e Tecnologia da Universidade de Coimbra

Prof. Dr. José Luis Guimarães Oliveira
Professor Associado, Departamento de Eletrónica, Telecomunicações e Informática da Universidade de Aveiro

Palavras Chave

arquitetura de software, estudos biomédicos, bases de dados da saúde.

Resumo

A sucessiva digitalização da informação de saúde dos cidadãos tem potenciado o desenvolvimento de aplicações que permitem estudar e extrair informação, facilitando a produção de conhecimento através de análise dos dados armazenados. A normalização de modelos de dados permite que as mesmas ferramentas possam ser usadas em diferentes bases de dados. O crescimento de comunidades que mantêm repositórios clínicos locais e isolados uns dos outros tem impedido que estudos epidemológicos, por exemplo, passar a ser realizados sobre um conjunto alargado de pessoas. Existe assim uma necessidade de transparentemente estudar múltiplas populações distribuidas globalmente. Esta dissertação propõe soluções para integrar distintos catálogos clínicos e ferramentas de software e permitir que possam ser utilizadas de forma distribuida.

Keywords software architecture, biomedical studies, healthcare databases.

Abstract

The increase of patient-level data available on digital format led to the development of aplications that can study and extract information and produce knowledge by analysing stored data. As data standardization is achieved, tools and studies can be shared in different databases. The growth of communities that maintain clinical repositories local and isolated has prevented epidemiological studies, for example, from being carried out on a wide range of people. There is a need for transparently study multiple globally-distributed populations. This dissertation proposes solutions to integrate software tools on distinct health catalogues, allowing them to be used distributely.

Contents

List of Figures

Acronyms

ACHILLES	Automated Characterization of Health Information at Large-scale Longitudinal Evidence Systems
ATHENA	Process developed to build standardized vocabularies compatible with the Common Data Model
ATLAS	Web tool for researchers to conduct scientific analysis on standardized observational data
CALYPSO	Criteria Assessment Logic for Your Population Study in Observational data
CDM	Common Data Model
CIRCE	Cohort Inclusion and Restriction Criteria Expression
HERMES	Health Entity Relationship and Metadata Exploration System
LAERTES	Largescale Adverse Effects Related to Treatment Evidence Standardization
OHDSI	Observational Health Data Sciences and Informatics
OMOP	Observational Medical Outcomes Partnership
EMIF	European Medical Information Framework
JSON	JavaScript Object Notation
REST	Representational state transfer
APOLLO	Augmenting Patient-level Observational Learning by Large-scaling OHDSI tools

1

Introduction

Et, avec un peu de mélancolie, peut-être, il ajouta :
– Droit devant soi on ne peut pas aller bien loin…

Antoine de Saint-Exupéry

1.1 Motivation

Clinical trials for new treatments and drugs assessment, such as many other health related studies require a test population in order to be validated by medical regulations. That test population is not easily found, in particular for rare diseases and conditions, as clinical data from different, disperse population is not publicly available and access may be object of different regulations and policies, rendering it unwise, unworthy or rather costly to bypass, under the risk that population cannot be used for the trial.

The advance of information systems and the informatization of most public (and private) sectors and services boosted the development of many technologies and data sharing between many, otherwise unlinked, markets. Notably, the health care sector started having almost all their medical and clinical records in a digital format.

Some platforms being developed allow intuitive listing and centralized search across distinct databases. Search is conducted based on fingerprints (descriptive information), which are often subjetive, incomplete and not always automated.

To address this problem, some, more customizable, tools have been developed, to complete fingerprint data with more specific information, allowing easier assessment of feasibility (or praticallity) of certain studies in a certain population. Providing certain conditions, like a similar structure, these feasibility tests may be shared and reused accross different databases, without the need for the test creator to know the data inside the database.

As patient databases are reaching a standardized model, studies can be designed for that model, shared with different potentially interesting database owners, reviewed (for privacy or political concerns), and then executed in a tool, with its results returned to the initial requester.

Although there are already tools that assess this kind of feasibility, there is currently no alternative to automate and share, efficiently, a feasibility test across distinct, related - or not, databases. This dissertation tries to address that necessity, by proposing and implementing solutions, to solve the problem.

1.2 Dissertation Outline

This dissertation is organized in three more chapters.

- chapter 2 describe the state of art mostly related with the EMIF and OHDSI projects;
- chapter 3 describes the requirements, the proposed architecture and the final application
- chapter 4 provides an evaluation of the work made and suggests other implementation and contributions to improve the project.

2

State of The Art

I seriously feel like the best days are ahead, and I like the idea of getting to do everything I did before but with more knowledge, experience, and street smarts. There's a certain love, appreciation, and gratitude . . . you don't have when you're younger, and it makes every accomplishment feel so much better.

J. Lo

Medical research has taken a great leap since the development of information systems, providing easier controlled access and sharing of health data. This chapter describes major contribution to this topic that have been done by the European Medical Information Framework (EMIF) and by Observational Health Data Sciences and Informatics (OHDSI) initiative.

2.1 The EMIF Project

The European Medical Information Framework (EMIF) project aims to develop a common information framework that will link up and facilitate access to diverse medical and research data sources, opening up new avenues of research for scientists. In this consortium there are initially leverage data on around 40 Million European adults and children by means of federation of healthcare databases and cohorts from 7 different countries (DK, DE, IT, NL, UK, ES, EE), designed to be representative of the different types of existing data sources (population-based registries, hospital-based databases, cohorts, national registries, biobanks, etc.).

This ongoing project culminated with the development of a web platform[1](through the colaboriteve github project - Montra) that gathers information of health data repositories. It allows the discovery of potentially suitable databases, by indexing and displaying descriptive information about each database (fingerprint) and organizing similar databases in communities.

The available information and features can be extended by means of plugins, which can be just external apps running parallel and external to the Catalogue, global widgets, or database specific specific plugins (e.g.: Achilles Web Visualization Tool)[1]

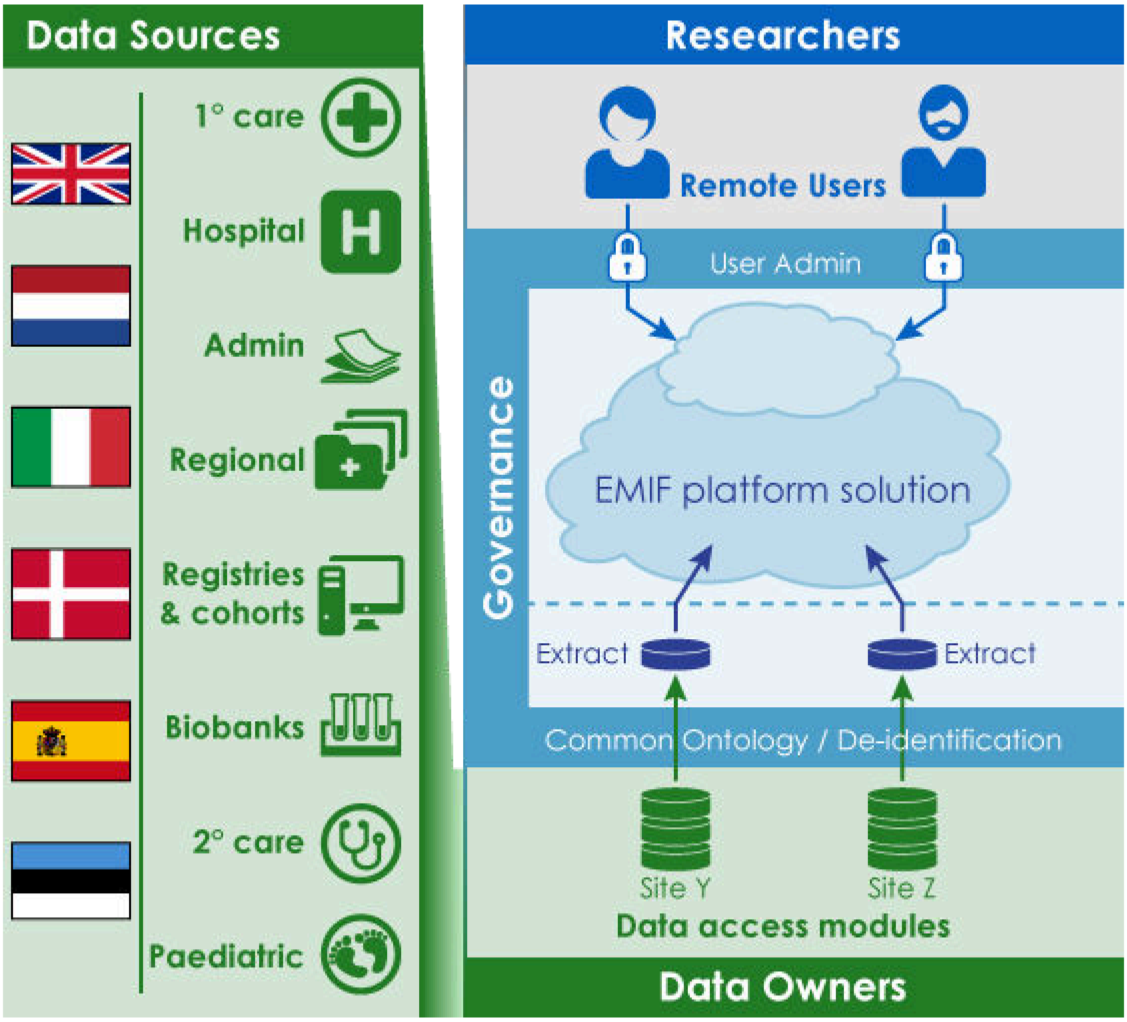


Figure 2.1: The EMIF Platform [2]

The EMIF community databases have been migrating their datasets to store observational patient-level data in a standardized Common Data Model (CDM). This is a big oportunity for researchers as they can now start using tools developed by OMOP and OHDSI across multiple databases.

[1] http://emif-catalogue.eu

2.2 The OHDSI Ecosystem

Observational Health Data Sciences and Informatics (OHDSI) is an international initiative that aims to transform medical decision making by creating reliable scientific evidence about disease natural history, healthcare delivery, and the effects of medical interventions through large-scale analysis of observational health databases for population-level estimation and patient-level predictions[3].

Historically it started in 2007 when the Observational Medical Outcomes Partnership (OMOP) was launched, by the Foundation for the National Institutes of Health (FNIH), in partnership with Pharmaceutical Research and Manufacturers of America (PhRMA) and the Food and Drug Administration (FDA), with the objective to answer a critical challenge: "what can medical researchers learn from assessing these new health databases, could a single approach be applied to multiple diseases, and could their findings be proven?".

By the end of the partnership, in 2013, OMOP had developed, and open sourced, a framework for observational research, consisting of a unified, standardized, Data Model for Health Care Databases, and some tools to extract information from it (like NATHAN, GROUCH and OSCAR). As the objectives envisioned by the founding members was achieved, the partnership was concluded by the FNIH. To continue its mission of developing tools and evidence to support analysis of healthcare observational data, OMOP research investigators have created OHDSI.

Tools and applications developed by OHDSI revolve around databases that follow a Common Data Model (CDM), some were built to help in the construction of a new CDM (WhiteRabbit, Athena, Usagi,. . .), others to retrieve information from the CDM, either by extracting direct information from the database (HERMES, Iris,ACHILLES,Heracles, . . .), or by analysing the database further and deriving information (CALYPSO,LAERTES, (ATLAS), . . .), and even others as resources to be used by other tools (WebAPI, CIRCE).

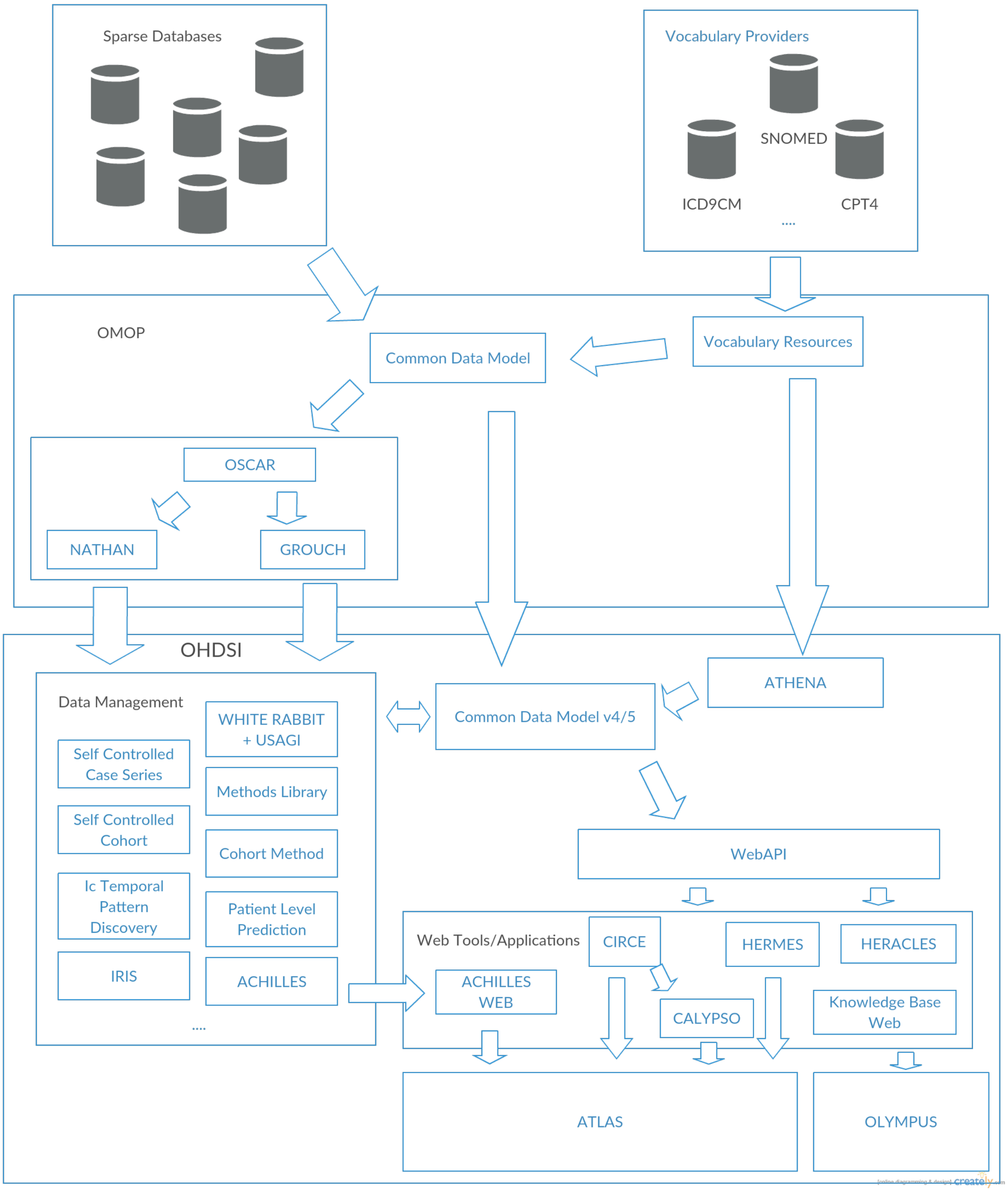


Figure 2.2: OMOP and OHDSI evolution

2.2.1 Common Data Model

The Common Data Model (CDM) was firstly designed and created by OMOP and it is currently being maintained by OHDSI as the core of component of its architecture. This medical-related observational database structure developed with the purpose of standardizing medical-observational datasources, allowing for concurrent optimal analysis of multiple data sources. It is currently on v5, but v4 databases can be easily updated to the newer version. Its design aims at identifying associations between interventions and outcomes (such as the patients exposed to a certain drug develop a certain condition).

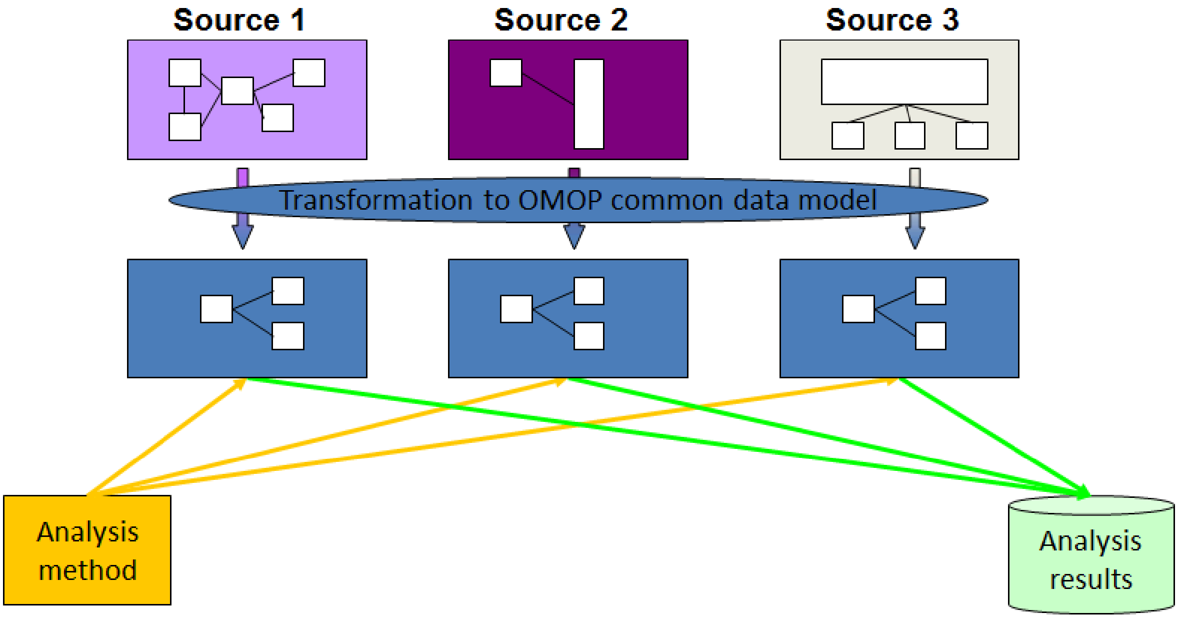


Figure 2.3: The importance of the CDM for data exploration.

With design-level support for standardized content the CDM ensures that research methods can be systematically applied to produce meaningfully comparable and reproducible results.

The CDM groups data in six main standardized, interconnected groups:

- Vocabularies – centralized information about concepts used in fact tables. These tables are mantained centrally as a service to the comunity;
- Metadata – general metadata about the data. It does not intend to be complete as most metadata is derived from data during ETL;
- Clinical – information and relation about clinical events during observation periods and demographic information for each person (eg. death, procedures, . . .);
- Health System – data about the healthcare provider responsible for administering the healthcare of the patient;
- Health Economics – data about costs, affordability, health plan and demography of patients, drugs, procedures;
- Derived Elements – data derived from other tables, such as agreggation in periods (eras) of exposure, dosage, and condition occurrence, as well as other data from similar subjects (patients, providers, visits)

2.2.2 Vocabulary Resources

Initially developed at OMOP, standardized vocabularies are maintained by OHDSI, through the ATHENA application, to enable transparent and consistent content across disparated observational databases.

ATHENA compiles vocabularies and terminology from different sources and converts them to a format that can be easily included in a custom deploy of the CDM.

The HERMES tool provides a interface for search and explore vocabularies in a local deploy of the CDM.

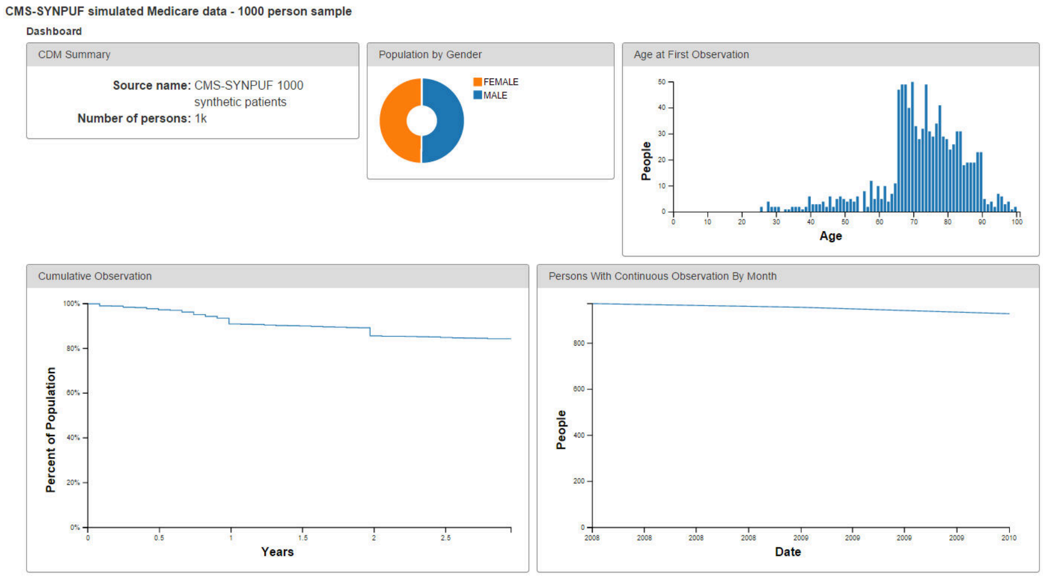
2.2.3 WebAPI

A REST API, developed in Java Spring, serve as support for OHDSI applications by providing resources regarding the CDM. It provides the default layer of communication between all OHDSI Web Applications and a CDM database. Main functionalities include: vocabulary search, defining and listing cohorts and performing feasibility and evidence studies.

2.2.4 ACHILLES and ACHILLES Web

ACHILLES (Automated Characterization of Health Information at Large-scale Longitudinal Evidence Systems) is a data characterization tool for the CDM, written in R, generating JSON reports that can be viewed in Achilles Web (Figure 2.4).

It provides high level description of the database population, drug exposures, medical procedures, observation periods, diseases, deaths and other data extracted from CDM databases.



CMS-SYNPUF simulated Medicare data - 1000 person sample

Dashboard

CDM Summary

Source name: CMS-SYNPUF 1000 synthetic patients

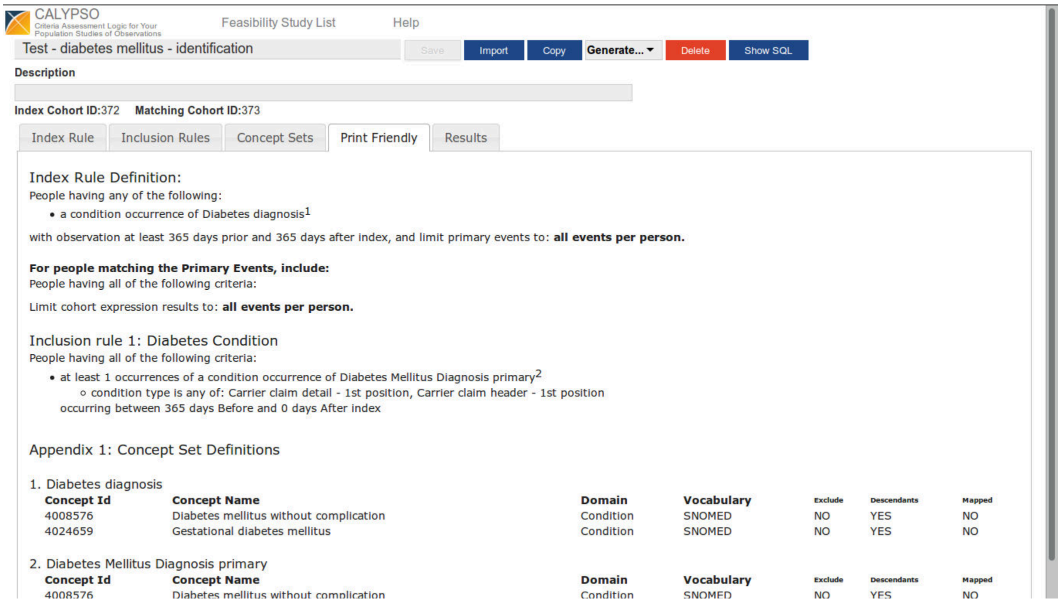
Number of persons: 1k

Figure 2.4: Achilles Web

2.2.5 CALYPSO

CALYPSO (Criteria Assessment Logic for Your Population Study in Observational data) is a web application that utilizes real world data (as provided by a WebAPI on top of a CDM) to simulate the availability of eligible patients for a study (feasibility study). It extends functionality to other OHDSI tools, by using Cohort Inclusion and Restriction Criteria Expression (CIRCE) cohort definitions to define inclusion and index rules.

With this tool, researchers can define an index rule (event they want to study) and inclusion rules (criteria to be analysed further). After the study is created, it can be exported, to a JSON format, or ran against the CDM database, showing, for each of the inclusion criteria, the impact on the availability of patients from the index population.



Test - diabetes mellitus - identification

Description

Index Cohort ID:372 **Matching Cohort ID**:373

Index Rule Definition:
People having any of the following:

- a condition occurrence of Diabetes diagnosis[1]

with observation at least 365 days prior and 365 days after index, and limit primary events to: **all events per person.**

For people matching the Primary Events, include:
People having all of the following criteria:

Limit cohort expression results to: **all events per person.**

Inclusion rule 1: Diabetes Condition
People having all of the following criteria:

- at least 1 occurrences of a condition occurrence of Diabetes Mellitus Diagnosis primary[2]
  - condition type is any of: Carrier claim detail - 1st position, Carrier claim header - 1st position

  occurring between 365 days Before and 0 days After index

Appendix 1: Concept Set Definitions

1. Diabetes diagnosis

Concept Id	Concept Name	Domain	Vocabulary	Exclude	Descendants	Mapped
4008576	Diabetes mellitus without complication	Condition	SNOMED	NO	YES	NO
4024659	Gestational diabetes mellitus	Condition	SNOMED	NO	YES	NO

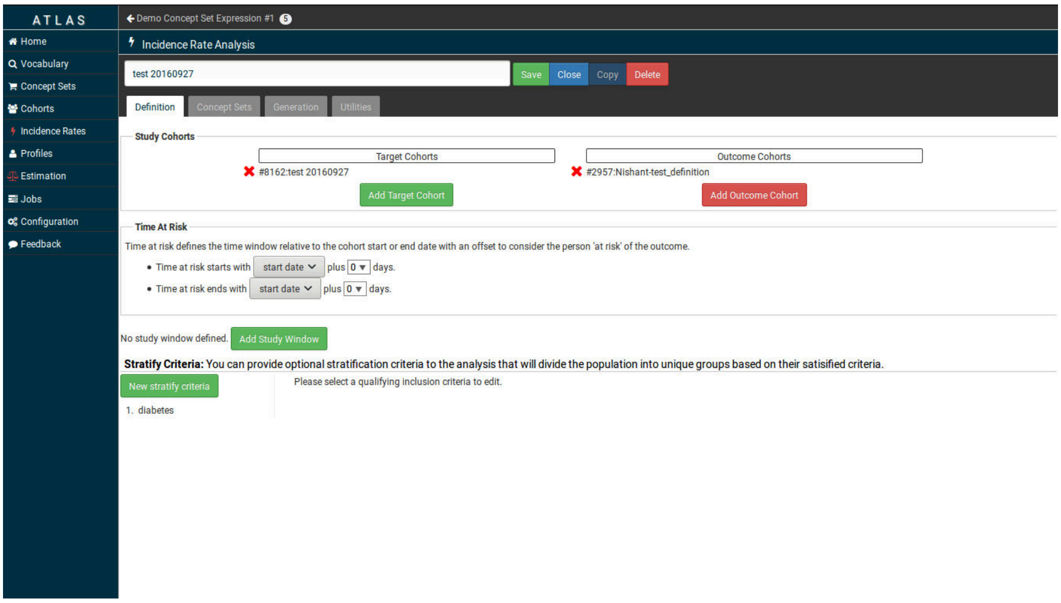
2. Diabetes Mellitus Diagnosis primary

Concept Id	Concept Name	Domain	Vocabulary	Exclude	Descendants	Mapped
4008576	Diabetes mellitus without complication	Condition	SNOMED	NO	YES	NO

Figure 2.5: CALYPSO - example of a cohort definition

2.2.6 ATLAS

ATLAS is the most recent OHDSI application. It is an integrated platform combining the features from Automated Characterization of Health Information at Large-scale Longitudinal Evidence Systems (ACHILLES), Health Entity Relationship and Metadata Exploration System (HERMES) and CIRCE, enhancing them with patient level estimation analyses, comparing different cohorts, and profiling users. As its development continues, older applications are being deprecated (HERMES is already superceded by ATLAS, and CIRCE features are available on both Criteria Assessment Logic for Your Population Study in Observational data (CALYPSO) and ATLAS).



test 20160927

✖ #8162:test 20160927

✖ #2957:Nishant-test_definition

Time at risk defines the time window relative to the cohort start or end date with an offset to consider the person 'at risk' of the outcome.

- Time at risk starts with start date plus 0 days.
- Time at risk ends with start date plus 0 days.

No study window defined.

Stratify Criteria: You can provide optional stratification criteria to the analysis that will divide the population into unique groups based on their satisified criteria.

Please select a qualifying inclusion criteria to edit.

1. diabetes

Figure 2.6: ATLAS main interface

3

System Definition

Sustainability can't be like some sort of a moral sacrifice or political dilemma or a philanthropical cause. It has to be a design challenge.

Bjarke Ingels

The Catalogue provides a centralized dashboard for searching and detailing different healthcare database from different institutions (organized in communities). As many of those databases have been adopting the CDM (and subsequent WebAPI), there's an opportunity for extending its features by providing out-of-the-box access to some OHDSI tools (notably CALYPSO and ATLAS) and also allow these tools to be used on a complete community (instead of singular databases at a time).

In this chapter, the project developed is described, from its requirements, arquitecture, implementation and deployment.

3.1 Main Goal

This project aims to allow extraction of knowledge from patient-level data on databases that have joined the EMIF project.

Using the catalogue as platform to discover health observational databases, and taking advantage of the works produced by OHDSI of study creation, the basic scenario to be explored is to provide WebAPI based tools to each database complying with CDM. The second, main, scenario is to allow studies to be spread across entire communities. These scenarios will be refered, respectively, as Simple and Community.

3.2 Study Organization

Studies created by OHDSI tools like CALYPSO and ATLAS, can be divided in 4 main phases of execution. Each with its own set of requirements for both scenarios.

Phase 1 – Concept Set Creation : In this phase, there is need to group vocabulary concepts that are going to be used in the study definition. This is done through to HERMES and CIRCE tools, which are integrated on CALYPSO and ATLAS.

In community studies, these concept sets, must exist in each of the remote databases and must all have the same definition. Having that in consideration, concepts sets must only use vocabulary available to all databases in the community.

Phase 2 – Cohort Rule Definition : After the concept sets are created for all databases in a community, CIRCE provides an interface and methods to specify index and inclusion rules.

Rules define (and are defined by) cohorts, or population groups, which are used in a study to define population being subject to a study. In feasibility tests (constructed with CALYPSO), an Index Population is the set of all patients being subjected to a study, defined by an index rule (which are defined by cohort intersection and union operations). Analogous, inclusion rules define which of the index population patients, match given criteria.

This process should be synchronized between all databases in the study, and the same rules, applied to the same cohorts and concept sets, must be replicated in each CDM.

Phase 3 – Study Execution : During this step, study definitions are sent to the WebAPI, which creates an assynchronous job to execute it in a pre-defined CDM, matching cohorts defined in the previous phases.

Phase 4 – Results Collection and Display : Results should be displayed individually for each database queried, and depending on the application/study made, can also display aggregated results (after each individual database is completed).

3.3 Architecture Design

The system design was guided by several non-functional requirements, namely:

Transparency for Researchers: All this process should be transparent to researchers: they should be able to create and execute a study in the same basic steps whether it is for a single database or a whole community.

Privacy Concerns: Healthcare databases very often contain sensitive data, and access to it must be regulated. There is, then, a requisite to implement, or, at least, take into consideration, a policy system, where database owners can approve or refuse study execution and/or limit the access to its results.

Integration Requirements: To ensure maintainabilty, eventual changes made to OHDSI tools, must not be specifc to this solution, being generic enough to exist on their own.

The fulfillment of these requirements led to the design and definition of two architectural solutions, with different approaches to the usage of OHDSI tools (namely WebAPI): pull and push models.

3.4 Pull Architecture

The first approach studied was named "pull architecture" due to its focus on running each application with a mediated connection to a remote WebAPI, configured by each database administrator (pulling data when needed), with almost no interaction with the owner, and with minimal data pre-processing.

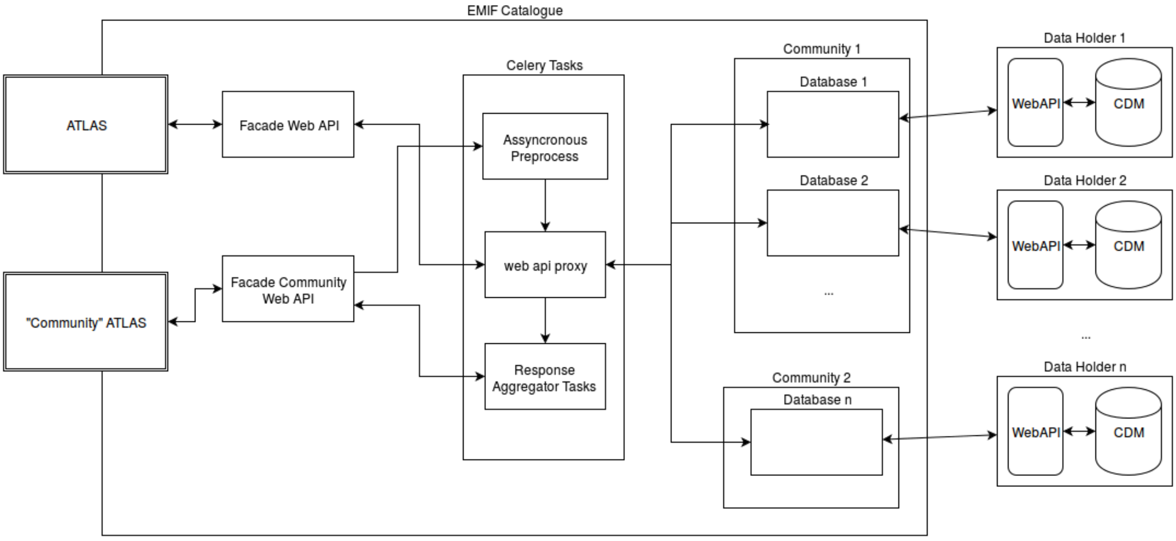


Figure 3.1: Pull Architecture

In this section we discuss how studies can mapped into this pull model.

3.4.1 Database Study

In the simple scenario (Figure 3.2), each OHDSI web application is connected to a proxy WebAPI endpoint (which only redirects requests to each database). Here, pre- and post-processing of requests and responses are minimal because no data aggregation is needed.

By using this proxy, it is possible to monitor requests, adding a layer of security and privacy, and also cache and limit requests to each database.

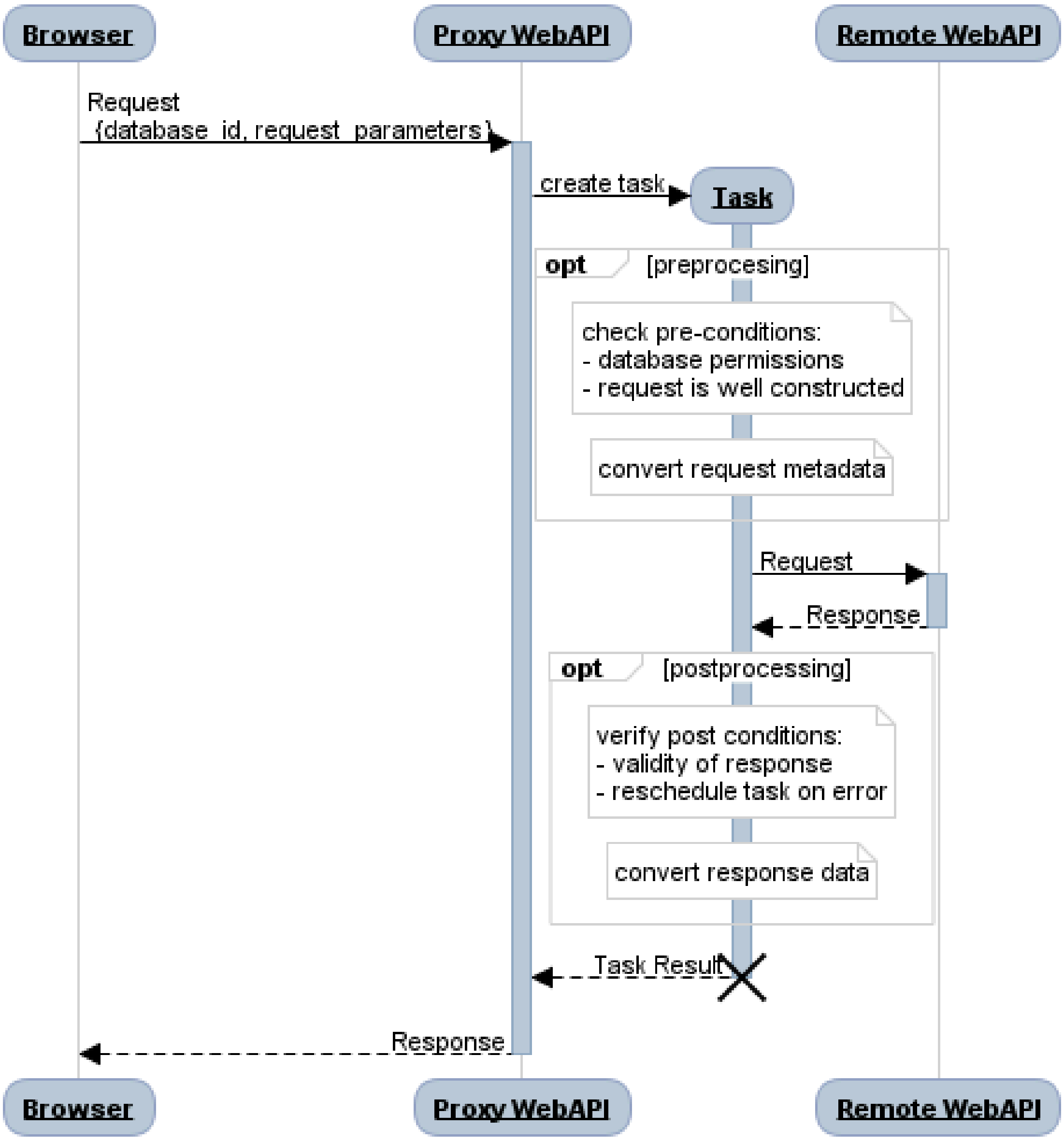


Figure 3.2: Simple Interaction (redirect to only one database)

3.4.2 Community Studies

With this pull-based approach, community wide studies require special handling of requests and a layer of external synchronization between the different, remote, databases. Below, it is described how each of the study phases refered in section 3.2 are handled:

Phase 1 – Concept Set Creation In this phase, vocabulary concepts which are going to be used in the index or inclusion rules are grouped in concept sets, who must be created in each of the remote APIs and must all have the same definition. Having that in consideration, and knowing the WebAPI and the CDM specifications, when searching for a concept, results from each datasource must be filtered and intersected, so as to only display concepts that are shared by all databases in the community. This phase also requires that a mapping of the concept IDs (and concept sets IDs) is mantained between databases, so as to further ease translation and distribution of requests.

Phase 2 – Cohort Rule Definition After the concept sets are created for all databases in a community, rules that specify cohorts can be created from them. These cohort rules, like the concept sets, must be shared with all the databases in the community, and IDs should be mapped for each database.

In this phase and the previous, latency is an important issue, and interaction cannot progress without results from previous requests are processed (Figure 3.3).

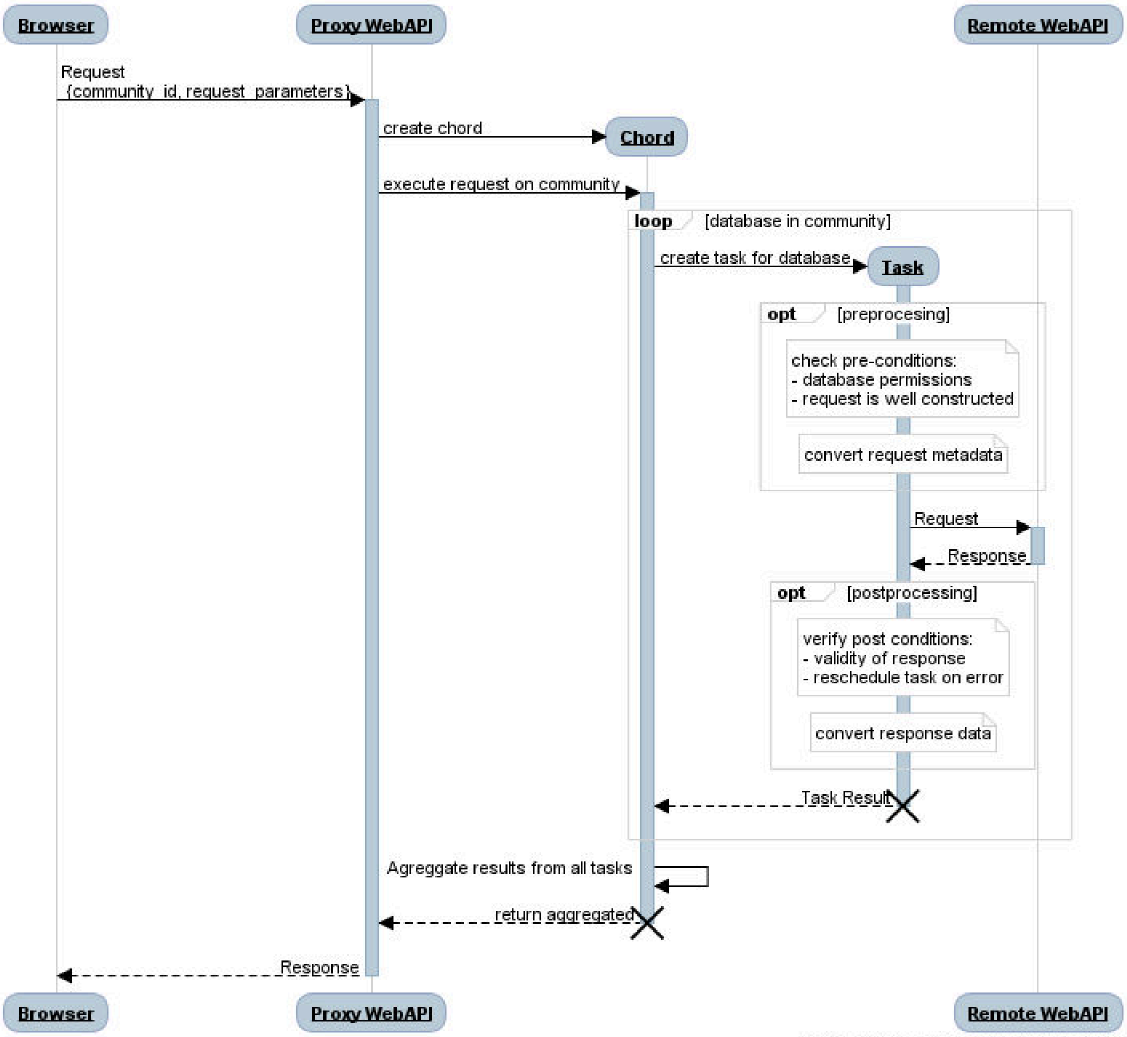


Figure 3.3: Community Interaction, Synchronous Requests

Phase 3 – Study Execution When the database is queried, cohorts are matched against each other in each database. Besides remapping of cohorts and concept set IDs, data holders permission may be requested to run the study. After permission is granted, the study should be run in the remote database and the results should be stored locally until they're needed for display.

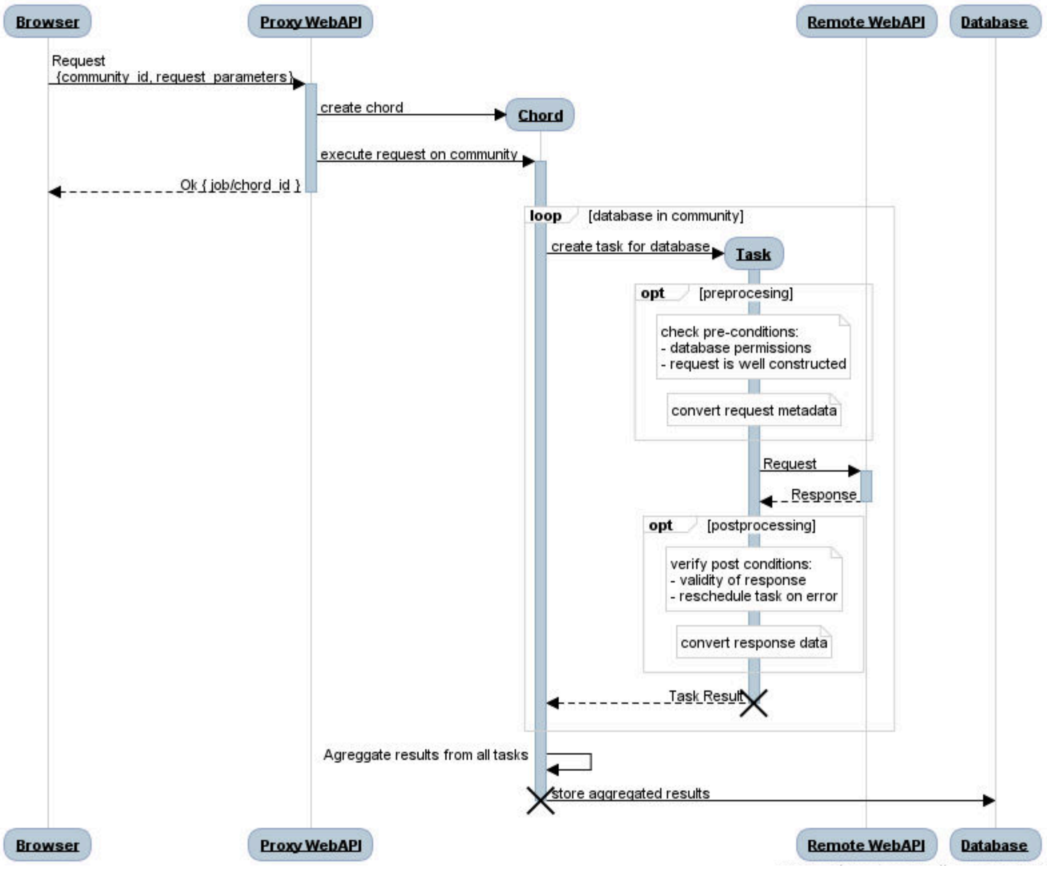


Figure 3.4: Community Interaction, Assynchronous Requests

Phase 4 – Results Collection and Display Results should be displayed individually for each database queried, and depending on the application/study made, can also display aggregated results (after each individual database is completed).

As studies are run assynchronously in the WebAPI, latency is not an issue during the last two phases and follow the behaviour described in Figure 3.4.

3.4.3 Drawbacks

This process involves high latency between the time of the study design and the time the user gets the results. This is observed on simple database studies (which go through a proxy), but it is more noticeable on community studies which may need granted express authorization before being ran.

Synchronization between concept sets and cohorts across remote databases can also be an issue, as there is no control nor garantee changes won't be made between first creation and its use. Alternatives to that involve creating a new cohort/concept set whenever it is needed, which, despite minimizing the problem, also increases the data stored remotely.

This solution also decreases maintainability, as tools from OHDSI are open-source and under constant development. So, keeping a local version and an adaptation to support communities in sync with new releases would be an exhausting work.

3.5 Push Architecture

To address the problems and limitations of the first architectural design, an alternative was developed to simplify the study creation process. Here we remove the direct interaction with remote databases by exploring study importation and exportation features available on most OHDSI webtools. This solution was named “Push” Architecture because studies are sent to each remote database owner, to be run in a remote installation. Results can then be shared optionally by each data owner and displayed to researchers as they become available.

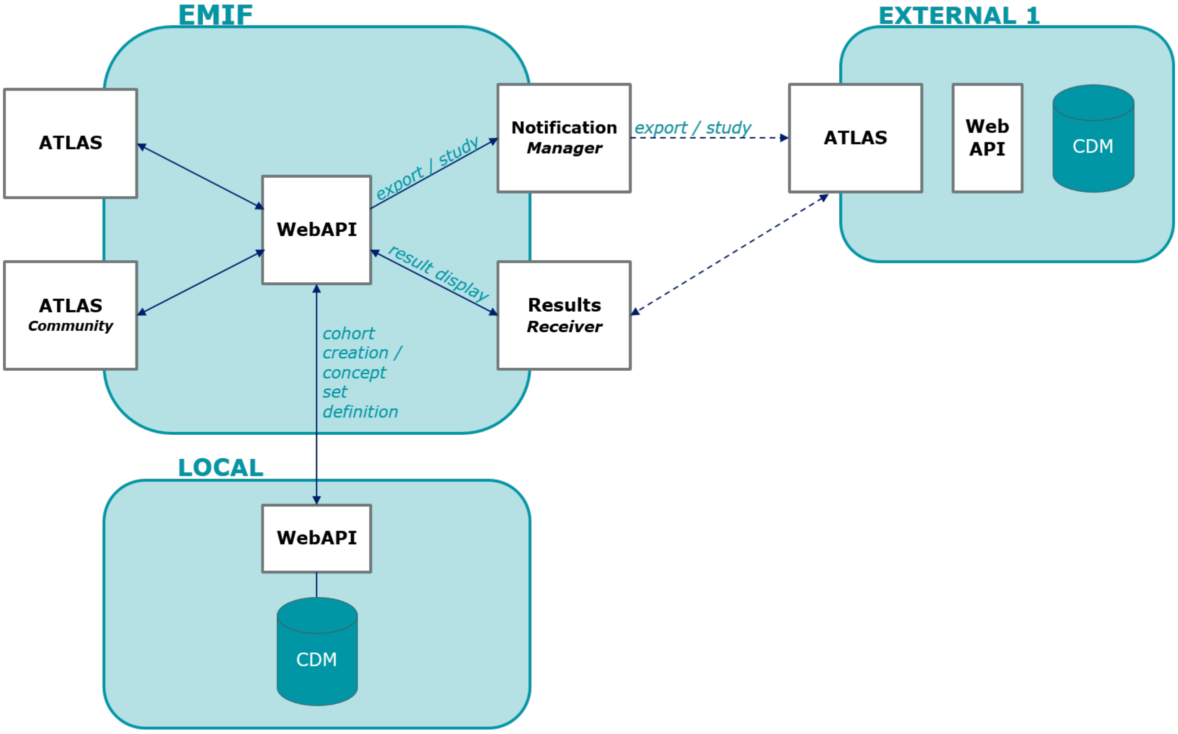


Figure 3.5: Push Architecture

This approach (see Figure 3.5) relies on 2 main components: a local WebAPI, backed by a CDM database with only vocabulary and a result schema for each remote database, and a REST API overriding most WebAPI resources, or redirecting them to a local WebAPI.

3.5.1 Study Creation

With a locally defined vocabulary database, concepts used for concept set creation and cohort definitions can be searched and created locally (Figure 3.6). Using this approach no change to OHDSI tools is needed, for both simple and community scenarios, and created concepts and cohorts may be reused for different studies, without the need to be recreated by researchers.

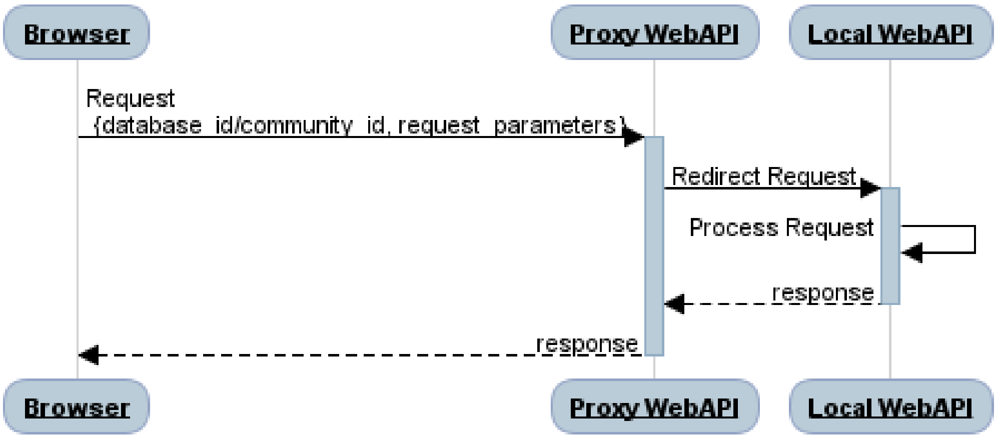


Figure 3.6: Phase 1 – Concept Set Creation and Phase 2 – Cohort Rule Definition execution sequence with the "push" architecture

3.5.2 Getting Results

During Phase 3 – Study Execution, studies must be exported, sent to each data owner, and imported in their side, to be reviewed and executed in the remote databases.

To ensure transparency for researchers, notably those already familiar with OHDSI tools, the resources normally provided/used by the WebAPI are "forged", meaning methods invoked are overridden to handle exportation on server side, without intervention of the researcher (Figure 3.7).

As Data Owners are notified, they can use their own deployment of the respective tools, import and run the study, analyse its results and upload them to the application(Figure 3.8).

As results are uploaded, they are added to a database acessible by the Local WebAPI and researcher who created the study is notified.

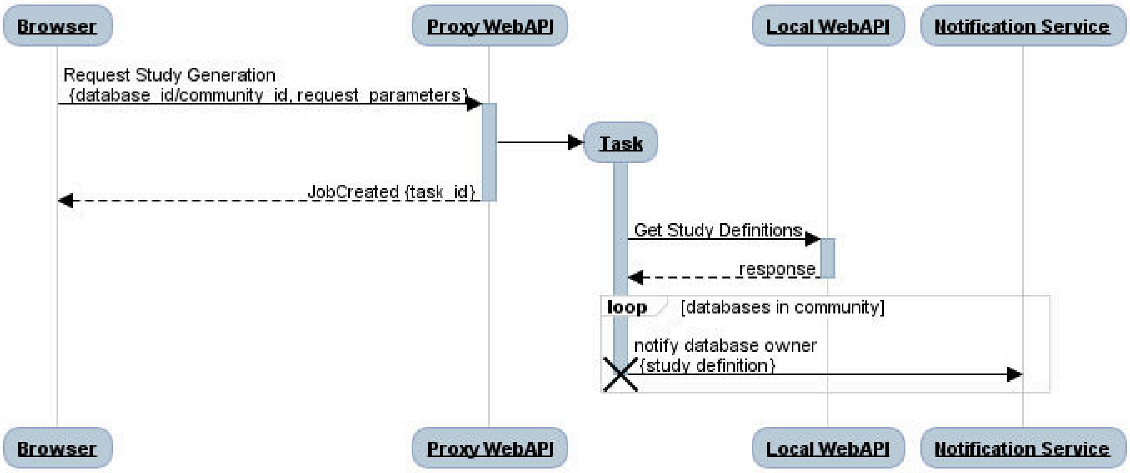


Figure 3.7: Phase 3 – Study Execution handling by the push architecture

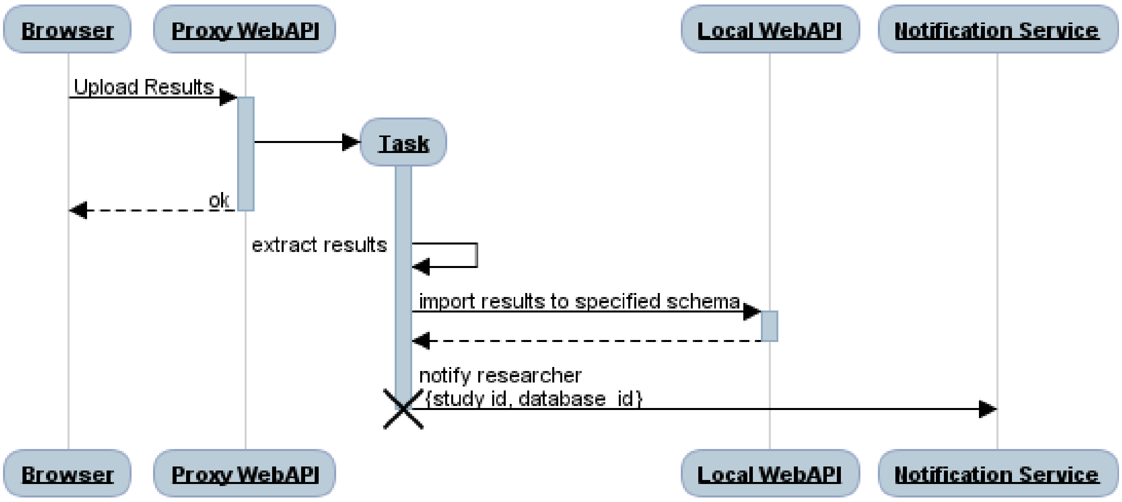


Figure 3.8: Result upload by the data owner

3.5.2.1 Drawbacks

This solution leaves cohort creation to be dealt and supervised by each data holder, thus increasing time between study creation and display of results. Besides that it is not mandatory for data owners to upload and share study results (or even execute it) as they can choose to abort the process in any step, namely because the study and its results may violate privacy policys (either by being too generic or too specific). It is also possible (but unlikely) for results to be forged during the upload results process.

3.6 Final Application – APOLLO

The Push Architecture resulted in a distributed application which we named APOLLO (Augmenting Patient-level Observational Learning by Large-scaling OHDSI tools).

Augmenting Patient-level Observational Learning by Large-scaling OHDSI tools (APOLLO) consists of a web server, developed in Python and Django[1], implementing a Proxy WebAPI and static serving OHDSI applications – ATLAS, CALYPSO – a database to support django functionalities, with contact information of database owners (used for notifications purposes), a Celery[2] worker for assynchronous task execution - used to parallelize requests to remote databases, a RabbitMQ[3] broker for communication between Django and Celery.

The Proxy WebAPI provides, for entire communities and single databases, resources similar to the ones returned by OHDSI WebAPI, as most of them are directly forwarded to our local WebAPI(for instance, vocabulary and results services, see Figure 3.6), by using python requests package[4]. Others, like study generation services, use asynchronous celery tasks (or chords) for background exporting and notifying data owners (Figure 3.7).

To support this application, we also have an OHDSI Stack, consisting of a WebAPI deployed in a Tomcat server and a PostgreSQL database, containing schemas and tables necessary for the well functioning of an OHDSI application and its WebAPI:

CDM schema with only vocabulary tables filled with relevant concepts for a study creation;

WebAPI schema with tables regarding webapi functionality, such as job execution status, cohort definitions, study definitions;

Result schema for every remote database automatically created every time a new remote database is added, and is used to store imported results.

3.6.1 Deployment

As each component of the application is designed to be independent with well defined communication interfaces (eg: to interact with a database engine http messages following a specific protocol should be sent to a certain port; WebAPI only accepts http requests following REST norms; etc.), the deployment, orchestration and installation was made using lightweight Docker[5] containers.

As seen in Figure 3.9, the application is composed of 6 containers from 5 images (postgresql image is shared by both OHDSI database and CDM tools). All images come from the official repositories, with the exception of APOLLO's which is built from the Django.

[1] http://www.djangoproject.com

[2] http://www.celeryproject.org/

[3] http://www.celeryproject.org/

[4] http://docs.python-requests.org

[5] www.docker.com

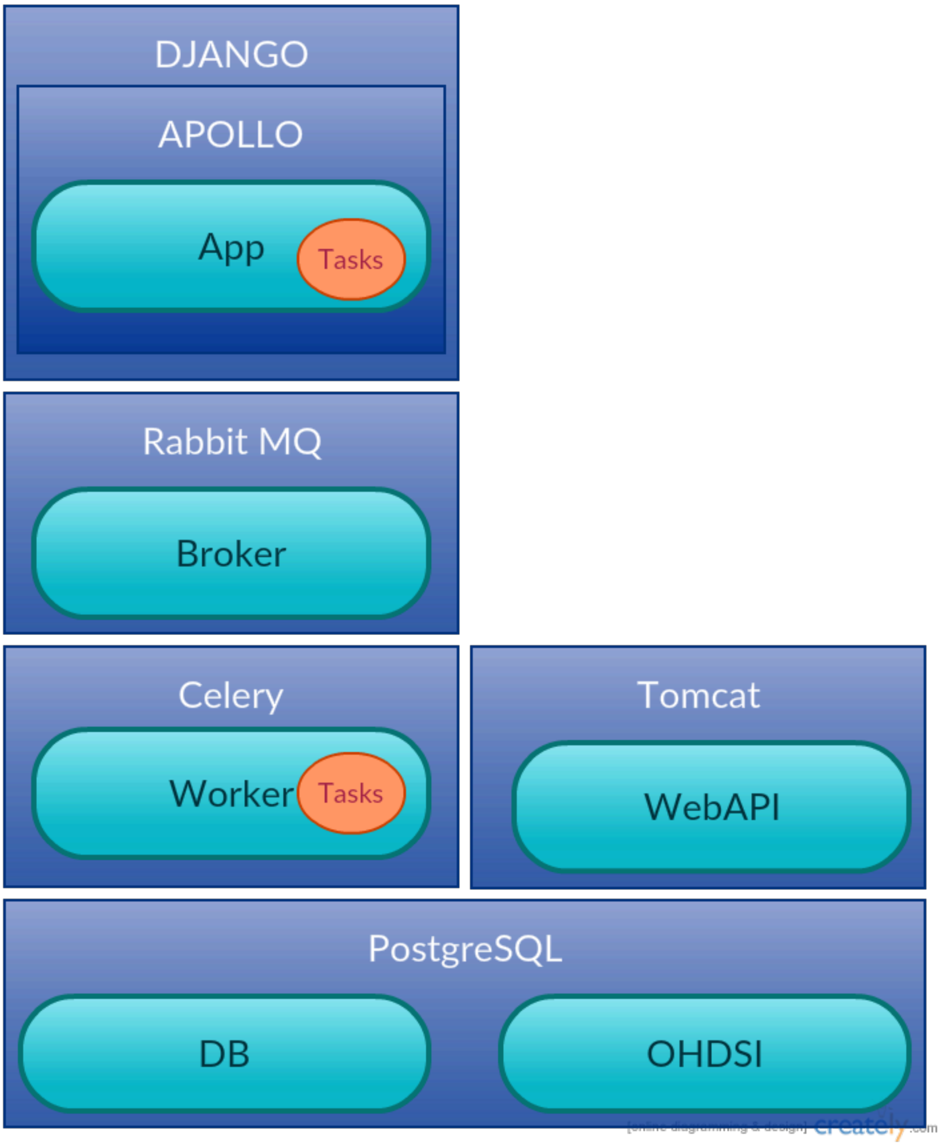


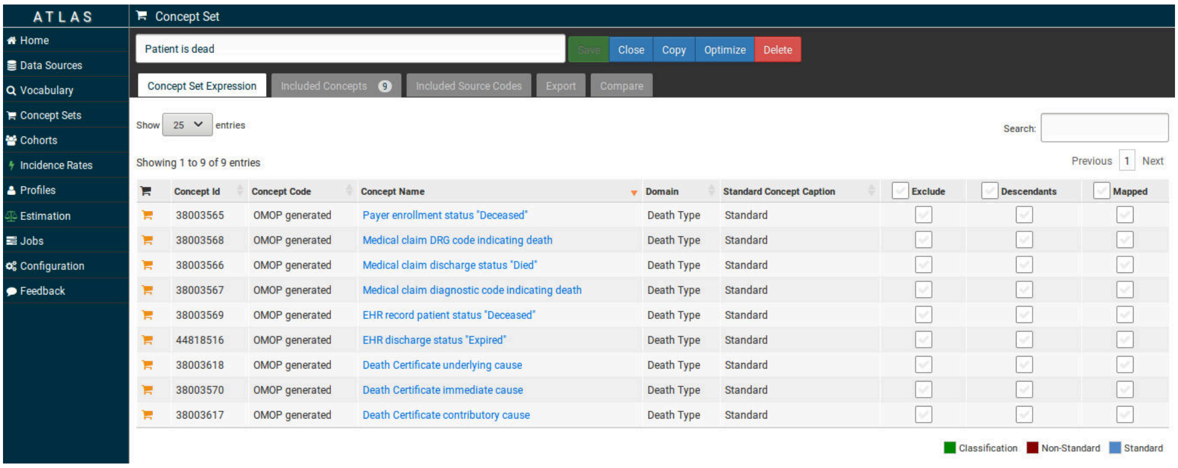
Figure 3.9: APOLLO Deployment: Docker Containers (rounded rectangles), Images (normal rectangles) and Shared Volumes (circles)

3.6.2 Usage

In this section we describe basic screenshots from the user interface, relating to the 4 phases described in 3.2.

During Phase 1 – Concept Set Creation (Figure 3.10) users can create and define concept sets, which are used during Phase 2 – Cohort Rule Definition (Figure 3.11) to define initial events rules and inclusion criteria.

When a researcher completes all cohort definitions, he can start Phase 3 – Study Execution (Figure 3.12). In this phase, APOLLO aggregates data from the Local WebAPI and sends them to each database owner participating in the study. The generated JSON can be imported and reviewed by each data owner to remote installation



Patient is dead

Show 25 entries

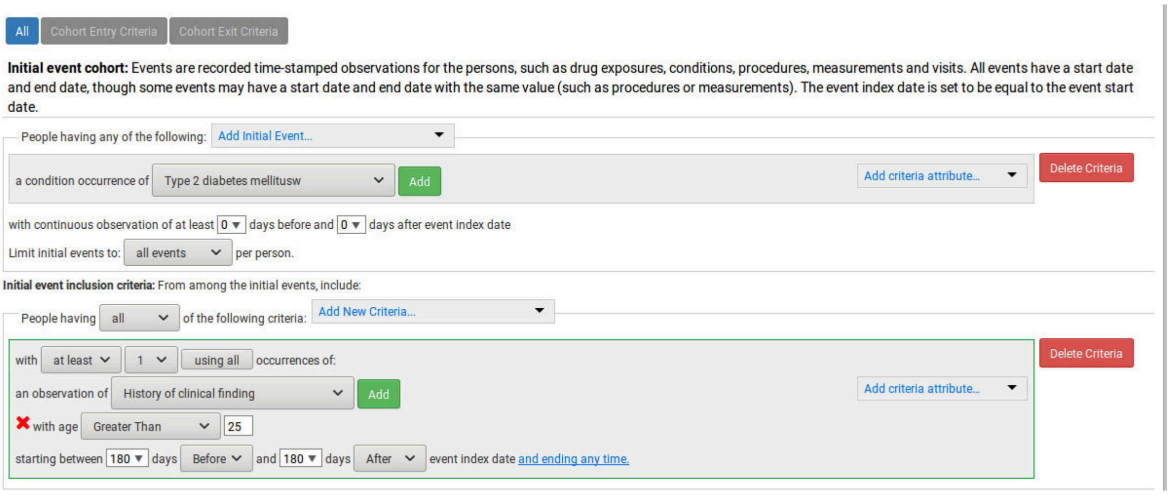
Search:

Showing 1 to 9 of 9 entries

	Concept Id	Concept Code	Concept Name	Domain	Standard Concept Caption	Exclude	Descendants	Mapped
	38003565	OMOP generated	Payer enrollment status "Deceased"	Death Type	Standard	□	□	□
	38003568	OMOP generated	Medical claim DRG code indicating death	Death Type	Standard	□	□	□
	38003566	OMOP generated	Medical claim discharge status "Died"	Death Type	Standard	□	□	□
	38003567	OMOP generated	Medical claim diagnostic code indicating death	Death Type	Standard	□	□	□
	38003569	OMOP generated	EHR record patient status "Deceased"	Death Type	Standard	□	□	□
	44818516	OMOP generated	EHR discharge status "Expired"	Death Type	Standard	□	□	□
	38003618	OMOP generated	Death Certificate underlying cause	Death Type	Standard	□	□	□
	38003570	OMOP generated	Death Certificate immediate cause	Death Type	Standard	□	□	□
	38003617	OMOP generated	Death Certificate contributory cause	Death Type	Standard	□	□	□

🟩 Classification 🟥 Non-Standard 🟦 Standard

Figure 3.10: Concept set creation screen



Initial event cohort: Events are recorded time-stamped observations for the persons, such as drug exposures, conditions, procedures, measurements and visits. All events have a start date and end date, though some events may have a start date and end date with the same value (such as procedures or measurements). The event index date is set to be equal to the event start date.

People having any of the following:

a condition occurrence of Type 2 diabetes mellitusw

with continuous observation of at least 0 days before and 0 days after event index date

Limit initial events to: all events per person.

Initial event inclusion criteria: From among the initial events, include:

People having all of the following criteria:

with at least 1 using all occurrences of:

an observation of History of clinical finding

with age Greater Than 25

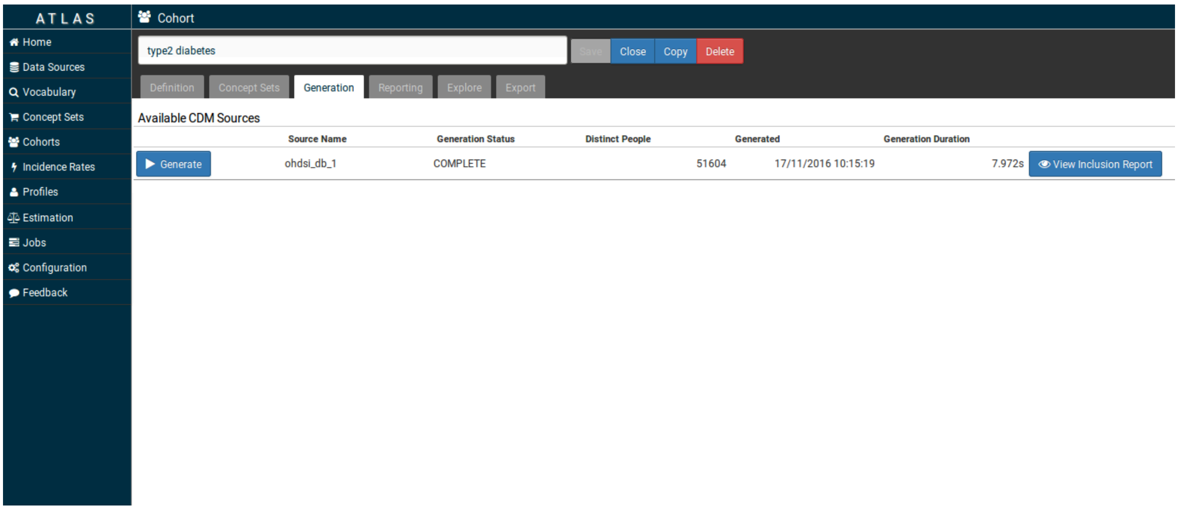
starting between 180 days Before and 180 days After event index date and ending any time.

Figure 3.11: Cohort Definition Screen

of ATLAS and generated as usual (Figure 3.13).

Ending this step, and as results become available, data owners can then begin Phase 4 – Results Collection and Display by exporting its results (Figure 3.14) and uploading them to APOLLO.

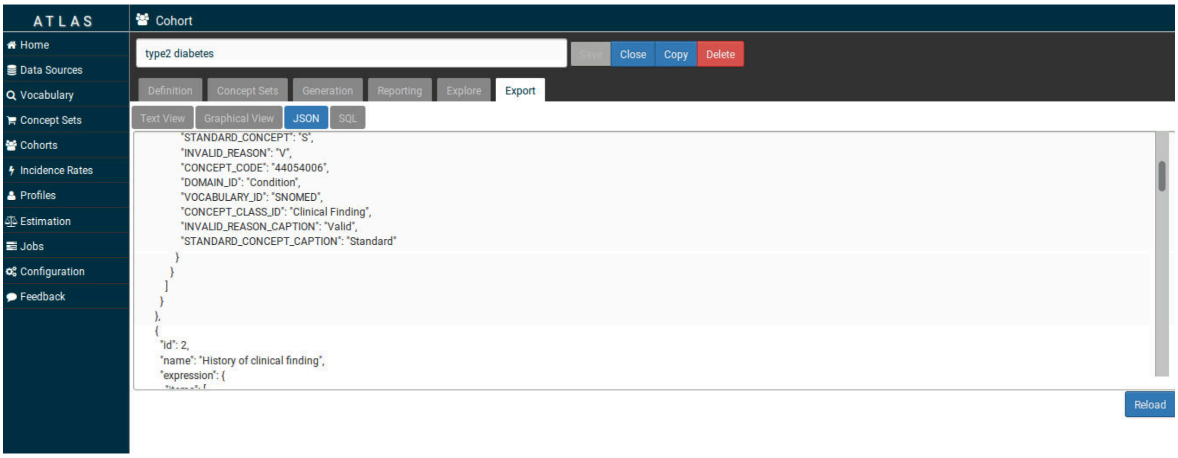
After the upload, researchers are notified and can check results in ATLAS (Figure 3.15).



Available CDM Sources

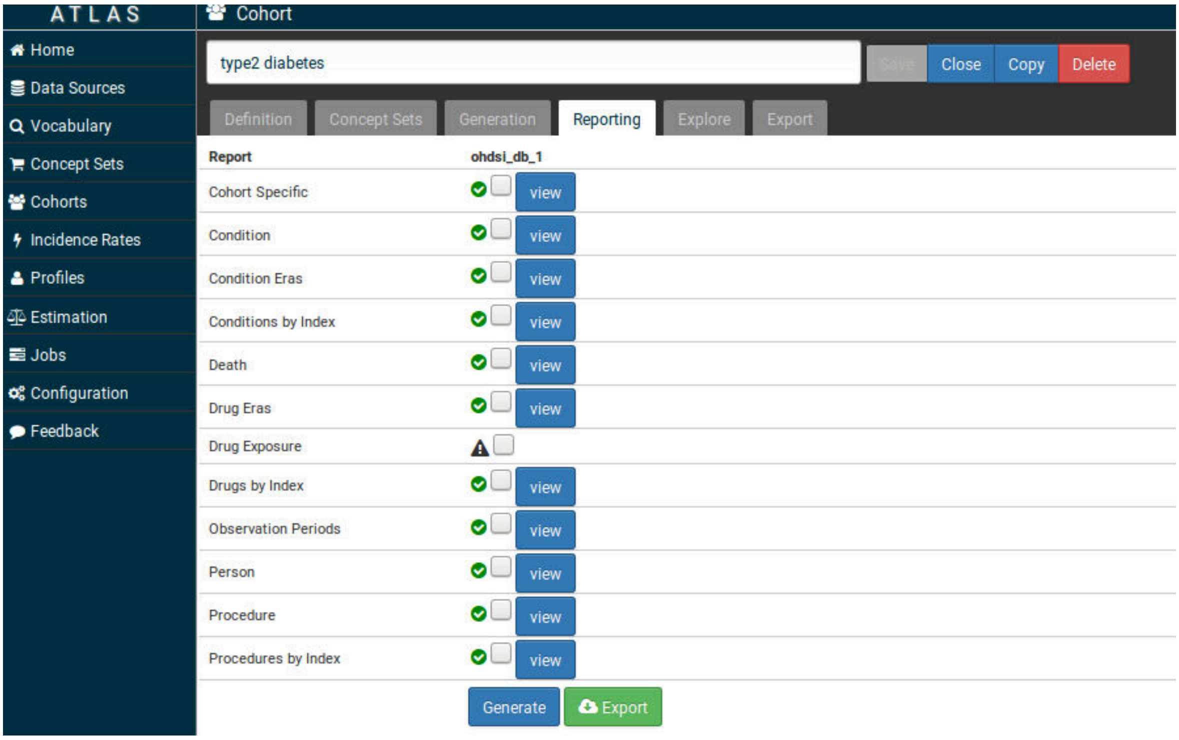
Source Name	Generation Status	Distinct People	Generated	Generation Duration
ohdsi_db_1	COMPLETE	51604	17/11/2016 10:15:19	7.972s

Figure 3.12: Cohort Generation Screen



```
            "STANDARD_CONCEPT": "S",
            "INVALID_REASON": "V",
            "CONCEPT_CODE": "44054006",
            "DOMAIN_ID": "Condition",
            "VOCABULARY_ID": "SNOMED",
            "CONCEPT_CLASS_ID": "Clinical Finding",
            "INVALID_REASON_CAPTION": "Valid",
            "STANDARD_CONCEPT_CAPTION": "Standard"
          }
        }
      ]
    }
  },
  {
   "id": 2,
   "name": "History of clinical finding",
   "expression": {
```

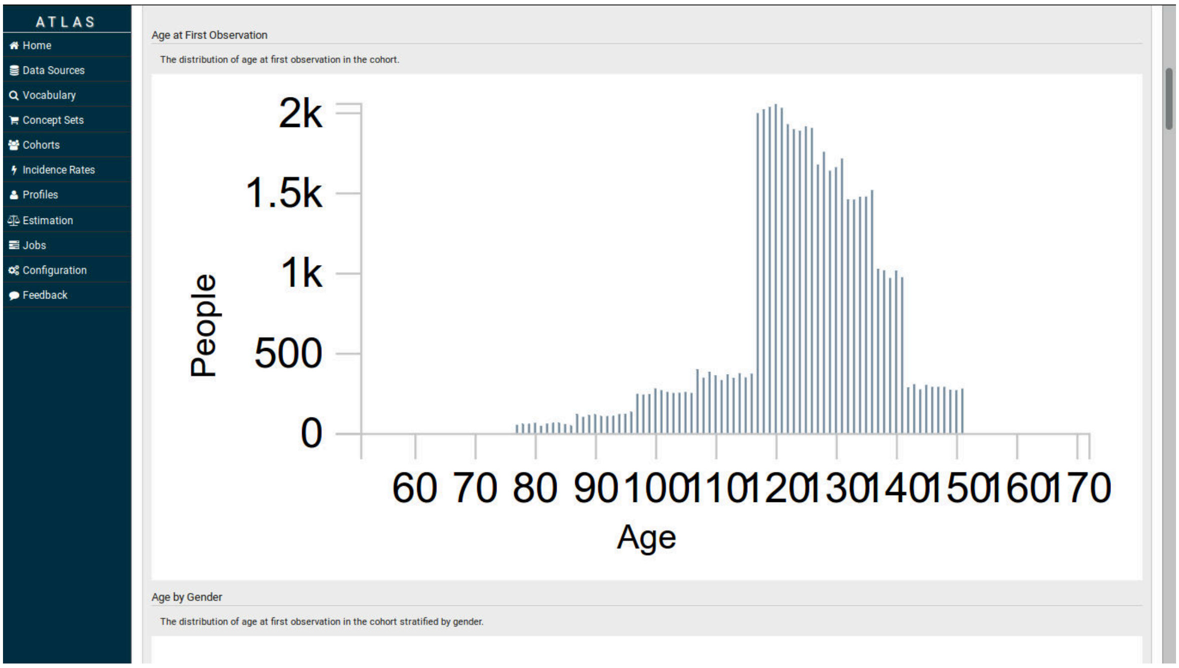
Figure 3.13: Cohort Importation Screen



type2 diabetes

Report	ohdsi_db_1	
Cohort Specific	✓ ☐	view
Condition	✓ ☐	view
Condition Eras	✓ ☐	view
Conditions by Index	✓ ☐	view
Death	✓ ☐	view
Drug Eras	✓ ☐	view
Drug Exposure	⚠ ☐	
Drugs by Index	✓ ☐	view
Observation Periods	✓ ☐	view
Person	✓ ☐	view
Procedure	✓ ☐	view
Procedures by Index	✓ ☐	view

Figure 3.14: Results overview and exportation screen



Age at First Observation

The distribution of age at first observation in the cohort.

Age by Gender

The distribution of age at first observation in the cohort stratified by gender.

Figure 3.15: Results detailed screen

3.7 Summary

The developed application addresses most requirements and goals of the project. For researchers (notably those familiar with OHDSI tools) which usability issues could arise, distributed studies can be created and exported with no changes in the ATLAS interaction. The responsability of ensuring patient privacy and enforcing policies, as studies are exported, is passed down to data owners who can review them and decide not to share its results. From a development standpoint, OHDSI tools aren't changed from the online-available versions and integration with new releases should not compromise the well functioning of the application, or involve a redesign of its architecture.

4

Conclusions

May I never be complete. May I never be content. May I never be perfect.

Chuck Palahniuk

In this chapter we evaluate the work done, and provide sugestions to the continuity of the project.

4.1 Evaluation of the work

This dissertation was developed with the intention of finding solutions to allow easy extraction of knowledge from patient-level data in a large number of databases. Since the very beginning, the adoption of OHDSI technologies and their integration with the catalogue was always the main concern. Having that said, requirements refered in chapter 3 were addressed during the design of the system, despite implementation limitations.

Also, when we try to combine different technologies and tools, most of them still under-development or community-maintained, with no clear, structured, documentation and with heterogeneous development pratices, we often have to spend a lot of time in learning and reinterpreting recently learned concepts or pratices. We had to deal with that since the beginning of this project, as tools like Calypso and Atlas, despite being quite intuitive to use by medical researchers and specialists, are rather complex in their interaction with the WebAPI and the study creation process. Designing, therefore, an infrastructure capable of, almost seamlessly, solve the problem without increasing technical complexity was the first and main issue found and addressed during the production of this dissertation, and was one of the reasons for the delay and deficiencies in the implementation.

4.2 Future Work

Permissions and policy implementation, despite being defined on the requirements, were not taken into consideration or exhaustively defined, and should be covered before the system is moved and integrated with the production environment.

Solutions presented were projected for software versions that are constantly on-release, namely, the SHIRO branch of WebAPI, introducing role-based permissions, was not considered during this writing, however both solutions should be adaptable enough to fit with it, once it is released.

Finally, this work is not completely finished until its implementation is done, considering that, this dissertation provides a good standpoint to go from theory to pratice.

References

[1] L. A. Bastião Silva, "Federated architecture for biomedical data integration", PhD thesis, Universidade de Aveiro, Portugal, 2015. [Online]. Available: http://hdl.handle.net/10773/15759.

[2] *The emif platform*, Nov. 2016. [Online]. Available: http://www.emif.eu/about/emif-platform.

[3] G. Hripcsak, J. D. Duke, N. H. Shah, C. G. Reich, V. Huser, M. J. Schuemie, M. A. Suchard, R. W. Park, I. C. K. Wong, P. R. Rijnbeek, J. van der Lei, N. Pratt, G. N. Norén, Y.-C. Li, P. E. Stang, D. Madigan, and P. B. Ryan, "Observational health data sciences and informatics (ohdsi): Opportunities for observational researchers", *Stud Health Technol Inform*, vol. 216, pp. 574–578, 2015.

[4] R. Murray, P. Ryan, and S. Reisinger, "Design and validation of a data simulation model for longitudinal healthcare data", *AMIA Annu Symp Proc., USA*, pp. 1176–1185, 2011.

[5] J. Overhage, P. Ryan, C. Reich, A. Hartzema, and P. Stang, "Validation of a common data model for active safety surveillance research", *J Am Med Inform Assoc*, vol. 19, no. 1, pp. 54–60, Jan. 2012.

[6] *Ohdsi website*, Oct. 2016. [Online]. Available: http://www.ohdsi.com.

[7] M. Mayer, L. Furlong, P. Torre, I. Planas, F. Cots, E. Izquierdo, J. Portabella, J. Rovira, A. Gutierrez-Sacristan, and F. Sanz, "Reuse of ehrs to support clinical research in a hospital of reference", *Studies in Health Technology and Informatics*, vol. 210, pp. 224–6, 2015.