

Development of a scalable architecture to extract metadata from distributed medical databases

André Silva Pedrosa

Master's Dissertation in Informatics Engineering

Supervisor: José Luis Guimarães Oliveira



Context



Researchers
want to perform
studies



EHR Databases



No direct
access to data

Context

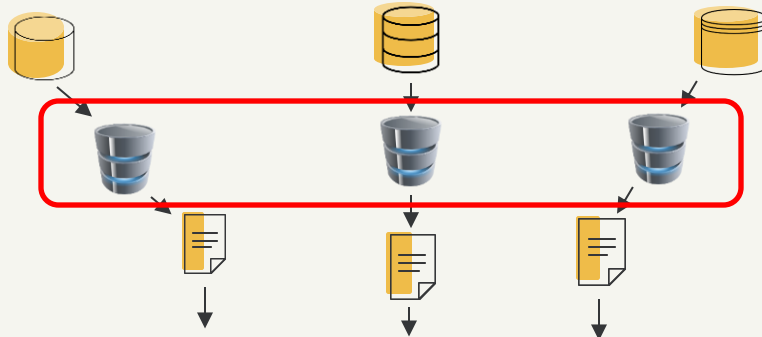


Metadata



Researchers can
find databases

Context



EHDEN PORTAL

Free text search EHDEN

Subscribe Manage

AP-HM

Fingerprint Literature Database Dashboard

Hits: 391 Unique Views: 176 Filled: 100 %

CONTACT DETAILS

Database Acronym
AP-HM

Database Name
[Health Data Warehouse of Assistance Publique - Hôpitaux de Marseille](#)

Institution name
[Assistance Publique - Hôpitaux de Marseille](#)

Department name
[Provence Alpes Cotes d'Azur](#)

1. Contact Details (14/19)	100%	✓
2. Database Description (11/11)	100%	✓
3. Technical Details CDM (5/5)	100%	✓
4. Data Governance and Ethics (9/9)	100%	✓
5. Publications (2/2)	100%	✓



Objectives



Provide a platform capable of holding and displaying metadata in an intuitive and user-friendly way



Develop or find a tool that extracts metadata from a database



Design a system capable of sending data to the platform, to keep it up-to-date

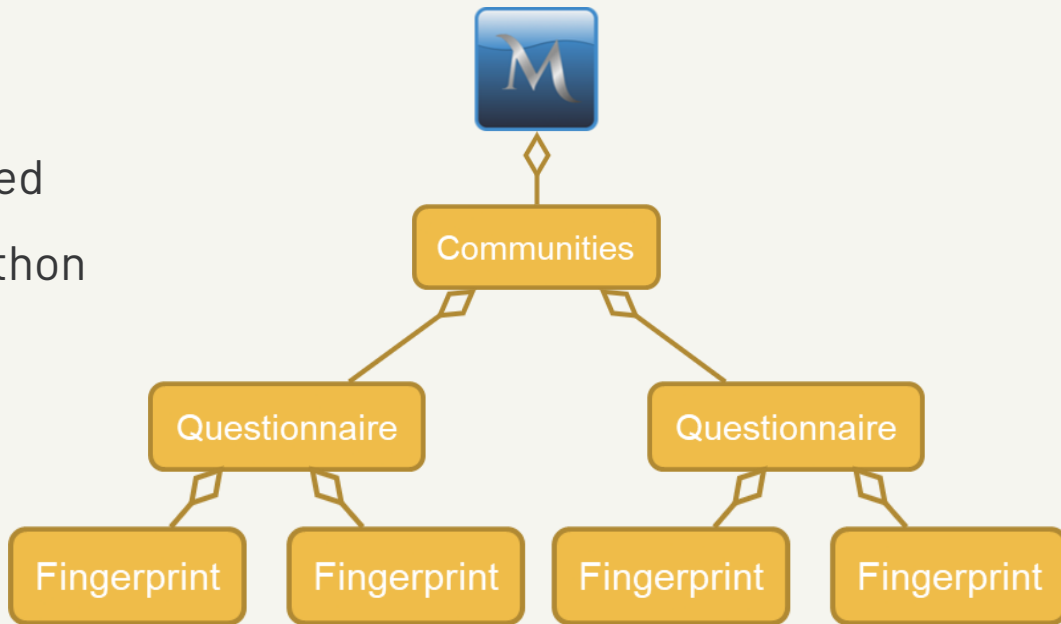
Background

Tool Name	Open Source	Visualization/Interaction		Extraction	Network
		Data protection	FAIR		
eGenVar [1]	✓ ⁴	✓ (Users + Permissions)	✓	✗	✗
MONTRA [2]	✓ ⁵	✓ (Role based)	✓	✗	✗
REDCap [10]	✗	✓ (Role based)	✓	✗	✗
Data Sphere [13]	✗	✓ (Authorized Users Only)	✗	✗	✗
MOLGENIS [15]	✓ ⁶	✓ (Role based)	✓	✗	✗
Cafe Variome [18]	✗	✓ (Role based)	✓	✗	✓
Mica & Opal [19]	✓ ⁷	✗	✓	✗	✗
BioSharing [20]	✓ ⁸	✗	✓	✗	✗
Dataverse [22]	✓ ⁹	✓ (Role Based)	✓	✗	✗
NADA [23]	✓ ¹⁰	✓ (Access Request)	✗	✗	✗
ACHILLES [25]	✓ ¹¹	✗		✓	✗
DataMed [27]	✓ ¹²	✗		✓	✗
Xtract [30]	✗	✗		✓	✓
Skluma [31]	✓ ¹³	✗		✓	✗
GAAIN [32]	✗	✗		✗	✓
PopMedNet [33]	✗	✗		✗	✓
EHR4CR [34]	✗	✗		✗	✓
NextGen Connect [35]	✓ ¹⁴	✗		✗	✓



Metadata Visualization: MONTRA

- Database-Centric design
- FAIR access to data provided
- Implemented in Django/Python
- Key Concepts
 - Communities
 - Questionnaires
 - Fingerprints



Fingerprint User Interface

The screenshot displays the EHDEN PORTAL Fingerprint user interface. The interface is divided into several sections:

- Left Sidebar:** Contains navigation links: HOME, CATALOGUE, DASHBOARD, ACADEMY, ARACHNE (DEMO), ATLAS (DEMO), EHDEN, PUBLICATIONS, SERVICE DESK (DEMO), STATUS, MANAGE (10), PORTAL, ABOUT, and SIGN OUT.
- Top Header:** Displays "EHDEN PORTAL" and a search bar labeled "Free text search EHDEN".
- Main Content Area:**
 - Displays "EHDEN PORTAL" and a search bar labeled "Free text search EHDEN".
 - Includes a "Subscribe" button and a "Manage" dropdown.
 - Shows the "AP-HM" section with tabs for "Fingerprint", "Literature", and "Database Dashboard".
 - Displays search statistics: "Hits: 391 Unique Views: 176 Filled: 100 %".
 - Features a "CONTACT DETAILS" section with input fields for:
 - Database Acronym: AP-HM
 - Database Name: Health Data Warehouse of Assistance Publique - Hopitaux de Marseille
 - Institution name: Assistance Publique - Hôpitaux de Marseille
 - Department name: Provence Alpes Cotes d'Azur
 - Includes a "Summary" panel on the right with a table of progress:

Item	Progress	Status
1. Contact Details (19/19)	100%	✓
2. Database Description (11/11)	100%	✓
3. Technical Details CDM (5/5)	100%	✓
4. Data Governance and Ethics (9/9)	100%	✓
5. Publications (2/2)	100%	✓

At the bottom, there are navigation buttons: "Previous", "Next", "Search", and "Advanced Query".

Fingerprint User Interface

Active Pathways

HOME

STRATEGIC

INSTRUMENTS

ACADEMY

ARCHIVE

ATLAS (2025)

NEWS

PUBLICATIONS

SERVICE DESK

STATES

BAROQUE

ABOUT

FAQ

HELP

FEEDBACK

PEOPLE

SIGN OUT

DEEN PORTAL

Post last search-DEEN

Contact Details

Store < Collapse all > Permissions >

Database Acronym

Database Name

Institution name

Department name

Street Address

City

Postal code

Country

Database Query Results Database

Control Sample (2/11)

0%

2. Statistical Outcomes (2/11)

0%

3. Technical Details (2/11)

0%

4. Data Governance and Ethics (2/11)

0%

5. Publications (2/11)

0%

< Previous

Next >

Cancel

Save

Admin Panel

HOME

STATISTICS

DASHBOARD

ACADEMY

ARCHIVE (2020)

FILED SERIES

NEWS

PUBLICATIONS

SERVICES DATA (2020)

STATUS

HISTORY

ABOUT

RAG

HELP

FEEDBACK

PROFILE

LOGOUT

OPEN PORTAL

Please login search-ENEN

Contact Details

Details > College id > Permissions >

Database Acronym
AP-HM

Database Name
Health Data Warehouse of Assistance Publique - Hôpitaux de Marseille

Institution name
Assistance Publique - Hôpitaux de Marseille

Department name
Provence Alpes Cotes d'Azur

Street Address
80 rue Brochier

City
Marseille

Postal code
13005

Country
France

2. Database Description (2017-18) 100% S

2. Service Details (2018-19) 100% S

4. Data Governance and Ethics (N/A) 100%

5. Publications (2/2) 100% S

The screenshot displays the 'EDEN PORTAL' interface. At the top, a navigation bar includes the user profile 'Andee Pedrosa' and a search bar with the text 'Find text search EDEN'. A left sidebar contains a list of navigation items: HOME, DATABASE, PERSONNEL, ACCOUNT, ARCHIVE (DEMO), ATLAS (DEMO), EVENT, PUBLICATIONS, SERVICE DATA (DEMO), STATISTICS, MESSAGE, NEWS, ABOUT, FAQ, HELP, FEEDBACK, PEOPLE, and SIGN OUT. The main content area is titled 'Contact Details' and features a form with the following fields: 1.01 Database Name, 1.02 Database Name, 1.03 Institution name, 1.04 Department name, 1.04.01 Street Address, 1.04.02 City, 1.04.02 Level, and 1.04.02 Postal code. A 'Collapse all' button is located to the right of the form. A right sidebar contains a 'Contact details' section with a list of links: 1. Database description, 2. Technical Details, CDM, 3. Data Governance and Ethics, and 4. Policy/Privacy. At the bottom of the form, there is a 'Cancel Query' button and a 'Load City' button with a dropdown menu. The bottom of the page shows a 'Previous' button, a 'Next' button, a 'Search' button, and an 'Advanced Query' button.

The screenshot displays the EHDEN Portal interface. On the left is a dark sidebar navigation menu with options like HOME, CATALOGUE, DATASOURCE, ACADEMY, BIBLEONE (DEMO), ATLAS (DEMO), EH-DEN, PUBLICATIONS, SERVICE DESK (DEMO), STATUS, MANAGE, PORTAL, and ABOUT. The main header area includes the "EHDEN PORTAL" logo, a search bar containing "Find test search EHDEN", and user controls for "Subscribe" and "Manage". Below the header, the breadcrumb trail shows "Pageprint", "Literature", and "Database Dashboard". The main content area features a red banner with statistics: "Hits: 301 Unique Views: 176 FPM: 100%". A blue box highlights the "CONTACT DETAILS" section, which contains fields for Database Acronym (AP-HM), Database Name (Health Data Warehouse of Assistance Publique - Hôpitaux de Marseille), Institution name (Assistance Publique - Hôpitaux de Marseille), and Department name (Provence Alpes Côte d'Azur). To the right of this section is a table titled "Current Details (1/4)" listing five items, each with a status of "100%": 1. Database Description (1/1), 2. Technical Details (2/4), 3. Data Governance and Ethics (3/4), and 5. Publications (2/2).

Fingerprint Schema

Type ▼	Text/Question ▼	Level/Number ▼
Category	Administrative Contact	h1
Question	Title	h2
Question	First Name	h2



Administrative Contact

Title

First Name

Fingerprint Schema

Type ▼	Text/Question ▼	Data type ▼	Value list ▼
Question	If you have a documented data dictionary, is the data dictionary a document (paper or electronic) or structured (spread sheet, database, XML, ISO 11179 etc.)	choice-multiple	Paper {...} Word (unstructured electronic){...} Spread sheet, Database {...} XML {...} ISO 11179 {...}



8.02. If you have a documented data dictionary

☐ Paper

☐ Word (unstructured electronic)

☐ Spread sheet, Database

☐ XML

☐ ISO 11179

Fingerprint Schema

Type	Text/Question	Data type	Value list
Question	If your database contains vaccine data, please indicate the completeness of recording in the target population (in the database) with respect to each vaccine	choice-tabular	None Partially Complete Complete Don't Know BCG Diphtheria Haemophilus influenzae Hepatitis A Hepatitis B HPV Influenza Measles Meningococcal Mumps Pertussis Pneumococcal Polio myelitis Rabies Rotavirus Rubella Shingles Tetanus Tick born encephalitis Typhoid Varicella choice



5.04. If your database contains vaccine data, please indicate the completeness of recording in the target population (in the database) with respect to each vaccine

	None	Partially Complete	Complete	Don't Know	More
BCG	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="text"/>
Diphtheria	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="text"/>

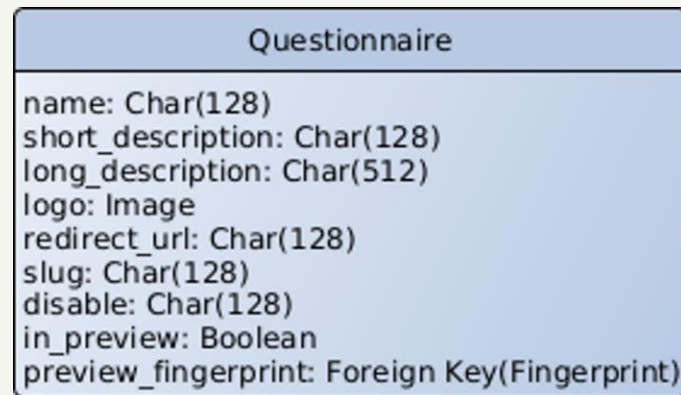
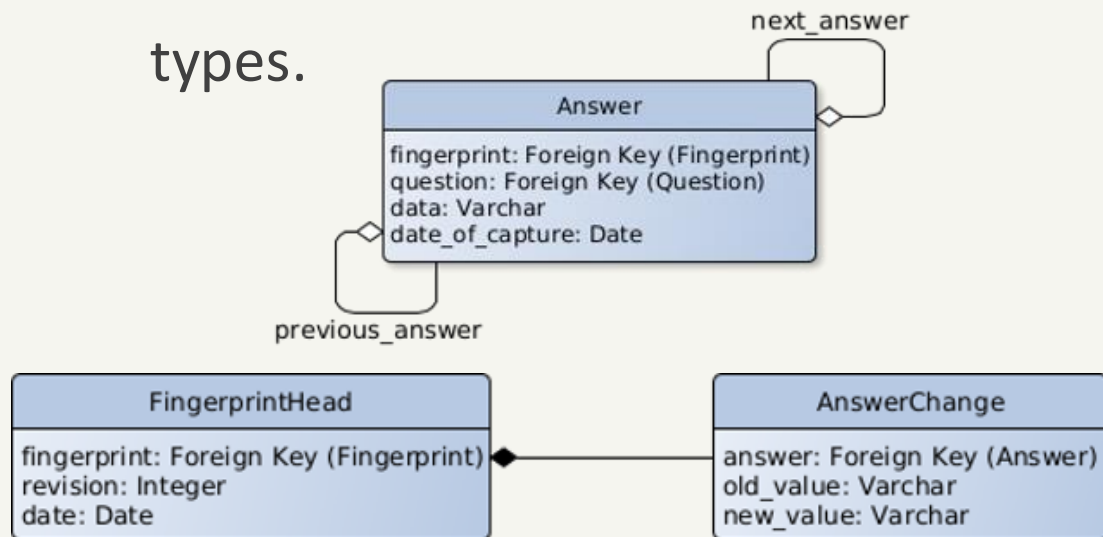
Data Models

- All answers' data is stored as text;
- Only 4 models to store information about a questionnaire:
 - Questionnaire, QuestionSet, Question and Choice;
 - Question model with a high number of fields to store information associated with different question types;

Question
questionset: Foreign Key (QuestionSet)
number: Char(255)
text: Varchar
type: Char(32)
extra: Char(128)
checks: Char(128)
footer: Varchar
slug: Char(128)
slug_fk: Foreign Key (Slugs)
help_text: Char(2255)
stats: Boolean
category: Boolean
tooltip: Boolean
visible_default: Boolean
mlt_ignore: Boolean
disposition: Integer
metadata: Varchar
show_advanced: Boolean

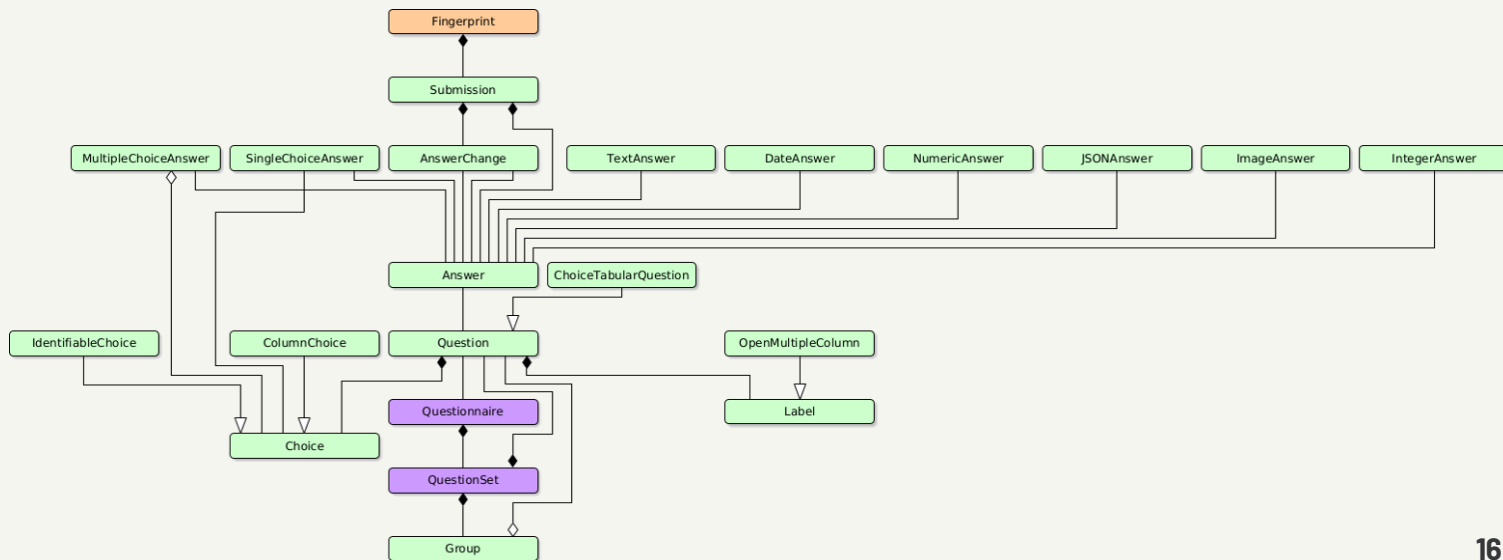
Data Models

- Other models, related to fingerprints, also had duplicated information and used incorrect data types.



Refactoring: Data Models

- Other models, related to fingerprints, also had duplicated information and used incorrect data types.



Refactoring: Fingerprint Views

Active Pathways

HOME

STRATEGIC

INSTRUMENTS

ACADEMY

ARCHIVE

ATLAS (2025)

NEWS

PUBLICATIONS

SERVICE DESK

STATES

BAROQUE

ABOUT

FAQ

HELP

FEEDBACK

PEOPLE

SIGN OUT

DEEN PORTAL

Post last search-DEEN

Contact Details

Store < Collapse all > Permissions >

Database Acronym

Database Name

Institution name

Department name

Street Address

City

Postal code

Country

Database Quantitative Database

Control Sample (2/11)

0%

2. Statistical Outcomes (2/11)

0%

3. Technical Details (2/11)

0%

4. Data Governance and Ethics (2/11)

0%

5. Publications (2/11)

0%

EDEN PORTAL

Free text search EDEN

Contact Details

Database Acronym
ADHM

Database Name
Health Data Warehouse of Assistance Publique – Hôpital de Marseille

Institution name
Assistance Publique – Hôpital de Marseille

Department name
Provence Alpes Côte d'Azur

Street Address
85 rue Brocheir

City
Marseille

Postal code
13005

Country
France

CONTACT CHANNELS (%)

Channel	Percentage
1 Database Description (11/11)	100%
2 Technical Guide (08/08)	100%
3 Data Governance and Ethics (N/A)	100%
4 Publications (02/02)	100%

ADMIN PORTAL

Free text search ADMIN

Avatar Placeholder

Admin Profile

HOME

CONTACTUS

HOMEPAGE

ACADEMY

AWARDS (EVENTS)

ATLAS (EVENTS)

DASH

PUBLICATIONS

SERVICE DESK (EVENTS)

STATISTICS

MANAGE

NEWS & ABOUT

FAQ

HELP

FEEDBACK

PROFILE

SIGN OUT

Contact Details

Collapse all

1.01 Database Acronym

1.02 Database Name

1.03 Institution name

1.04 Department name

1.04.01 Street Address

1.04.02 City

1.04.03 Postal code

1.05 Organization

2. Database Description

3. Technical Details CDM

4. Data Governance and Ethics

5. Publications

André Pedrosa

HOME

CATALOGUE

DASHBOARD

ACADEMY

BILLBOARD (DEMO)

ATLAS (DEMO)

ENDEN

PUBLICATIONS

SERVICE DESK (DEMO)

STATUS

MANAGE

PORTAL

ABOUT

ENDEN PORTAL

Free text search ENDEN

Subscribe Manage

AP-HM

Pageview

Literature

Databases Dashboard

Hit: 381

Unique Views: 170

FPM: 100%

Summary Colaps Show

CONTACT DETAILS

Database Acronym

AP-HM

Database Name

Health Data Warehouse of Assistance Publique - Hôpitaux de Marseille

Institution name

Assistance Publique - Hôpitaux de Marseille

Department name

Provence Alpes Côte d'Azur

1. General Details (3/3)

100%

2. Database Description (1/1)

100%

3. Technical Details (24/9)

100%

4. Data Governance and Ethics (9/9)

100%

5. Publications (2/2)

100%

Refactoring: Fingerprint Data Manipulation

- Move the data validation to the backend;
- Make use of Django's built-in data validation system;

Refactoring: Fingerprint Schema

- Fixed clutter problems
- Removed unused columns

Type	Text/Question	Data type	Value list
Question	If your database contains vaccine data, please indicate the completeness of recording in the target population (in the database) with respect to each vaccine	choice-tabular	None Partially Complete Complete Don't Know BCG Diphtheria Haemophilus influenzae Hepatitis A Hepatitis B HPV Influenza Measles Meningococcal Mumps Pertussis Pneumococcal Polio Poliovirus Rubella Shingles Tetanus Tick born encephalitis Typhoid Varicella choice

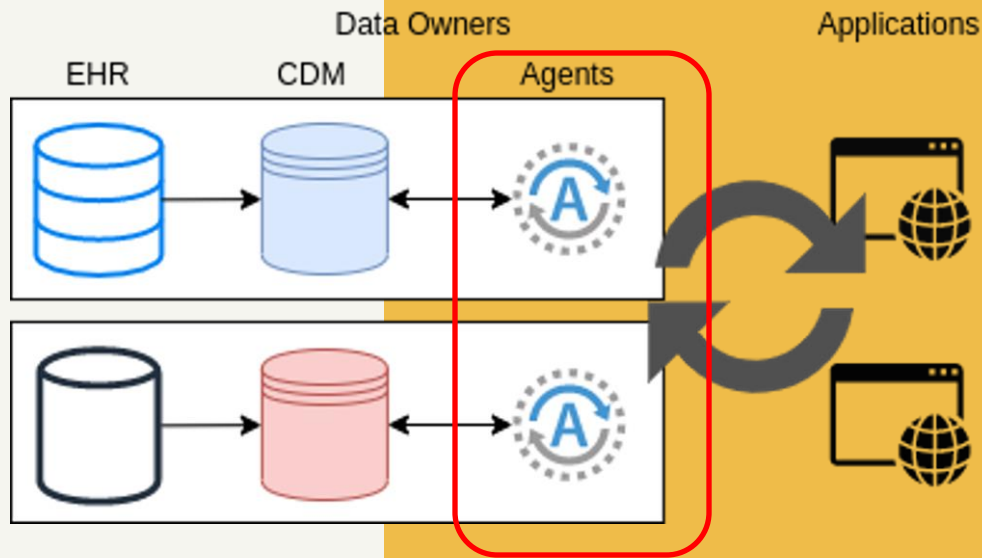


Type	Text/Question	Data type
Question	If your database contains vaccine data, please indicate the completeness of recording in the target population (in the database) with respect to each vaccine	choice tabular single
TabularChoice	None	
TabularChoice	Partially Complete	
TabularChoice	Complete	
TabularChoice	Don't Know	
TabularRow	BCG	
TabularRow	Diphtheria	
TabularRow	Haemophilus influenzae	
TabularRow	Hepatitis A	
TabularRow	Hepatitis B	
TabularRow	HPV	



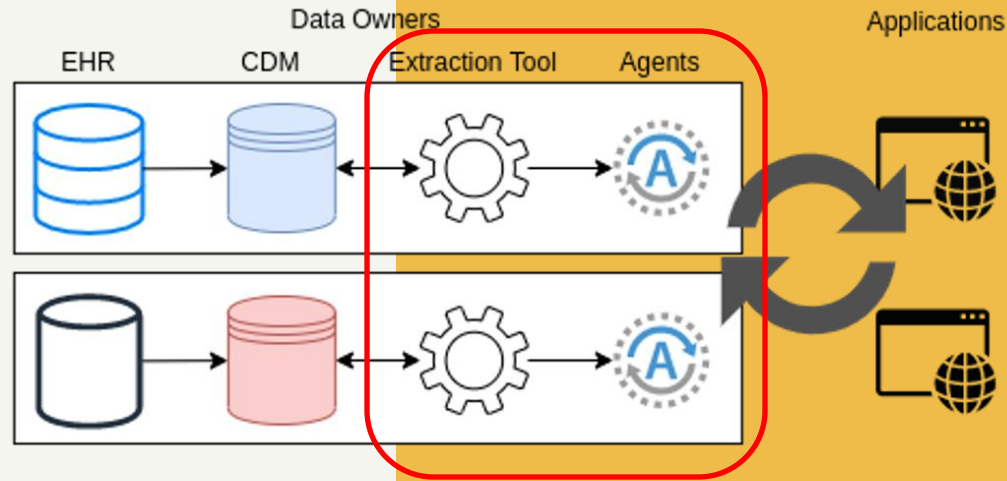
Metadata Extraction & Update

- Agents:
 - Run on data owner's deployment environment;
 - Extracts and sends metadata to the applications;
- Data owners might not want/be able to provide direct access to data;



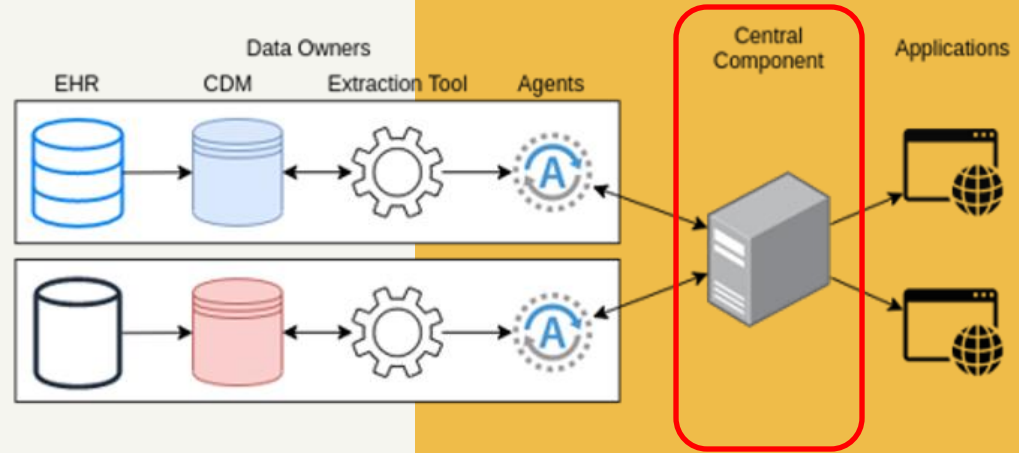
Metadata Extraction & Update

- Agent is now only in charge of sending metadata to applications;
- Data owners take care of the extraction process;
- How data gets to the applications?
 - Peer-to-peer architecture?



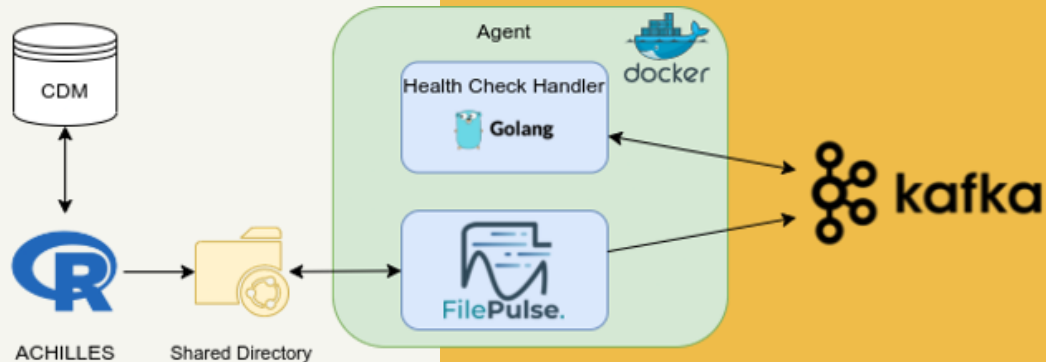
Metadata Extraction & Update

- A central component receives the data from the agents and sends it to the applications;



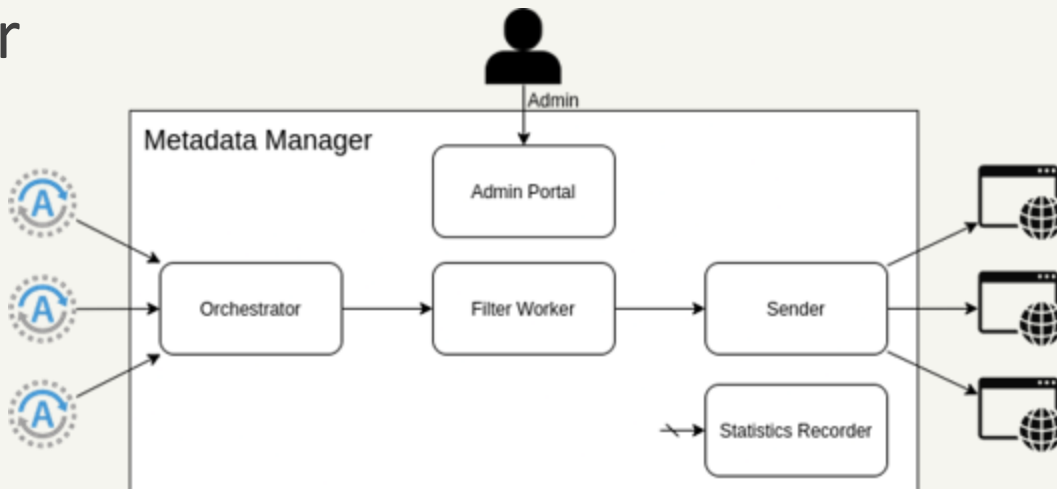
Metadata Extraction

- ACHILLES
- Asynchronous messaging systems:
 - Kafka
- FilePulse Source connector



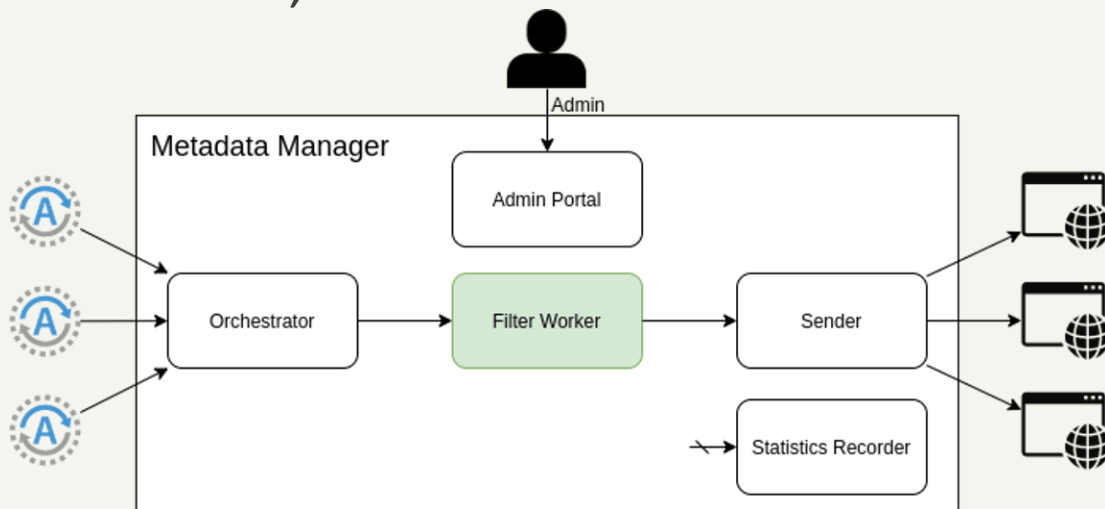
Metadata Update

- Data is in Kafka, now what?
- Kafka Sink connectors do not allow customization;
- Kafka deals with data as unbounded data flow;
- Metadata Manager
 - 5 components



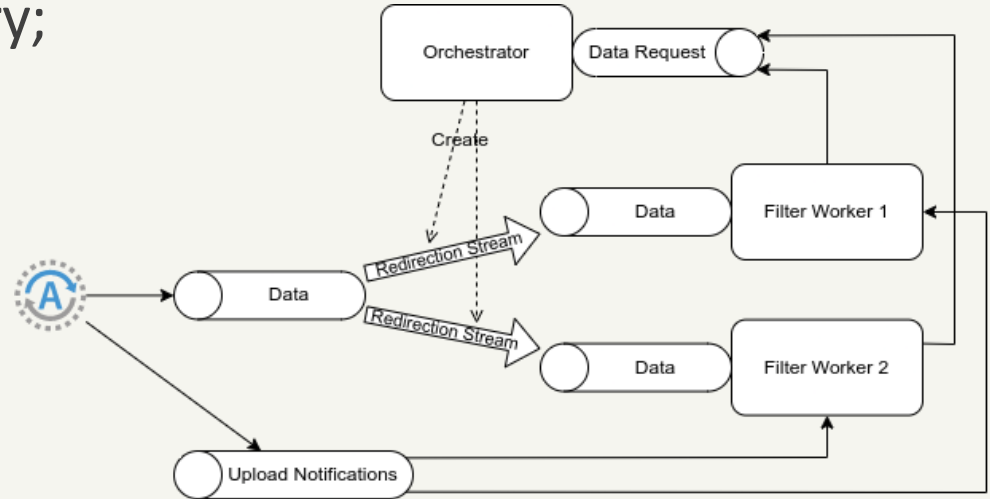
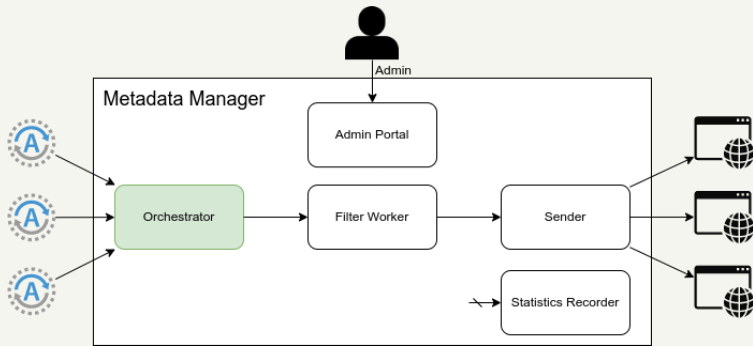
Filter Worker

- Application might not require the entire data received from the databases;
- Several filters at the same time;
- Implemented in Go;
- Can be scaled out.



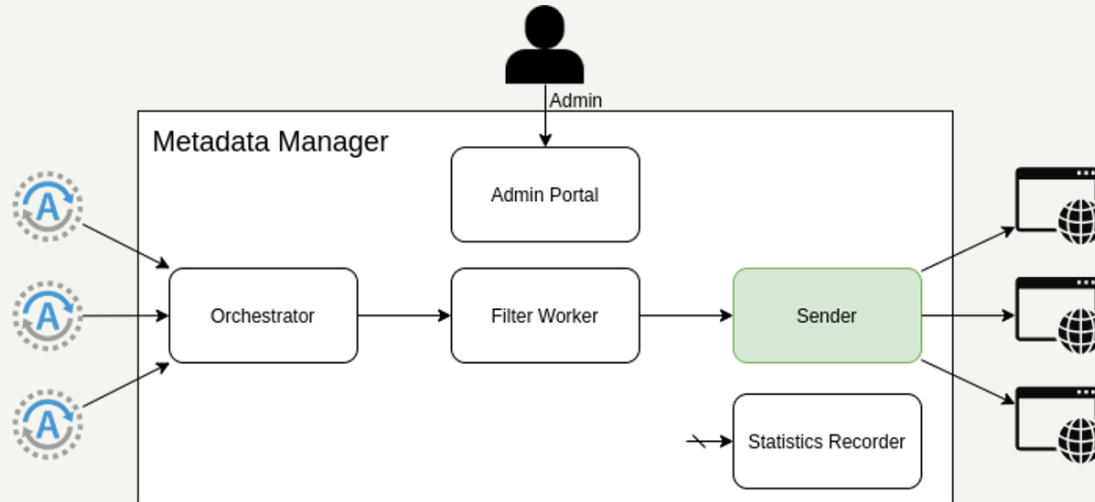
Orchestrator

- Distributes the data received from the databases across the existing Filter Worker instances;
- Uses Kafka Streams library;
- Implemented in Java;
- Can be scaled out;



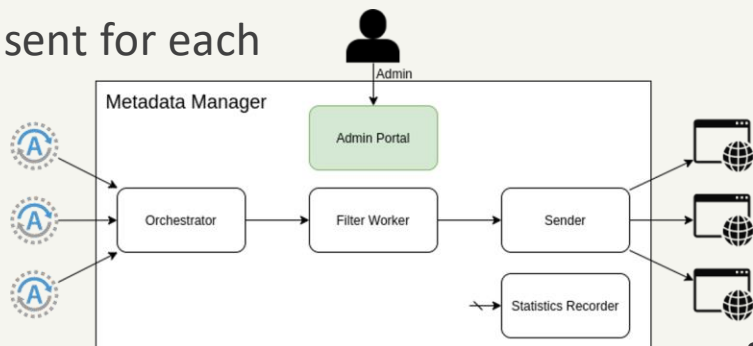
Sender

- Sends data to the applications in form of a HTTP request;
- Allows customization of the several request properties (method, URL, body, ...);



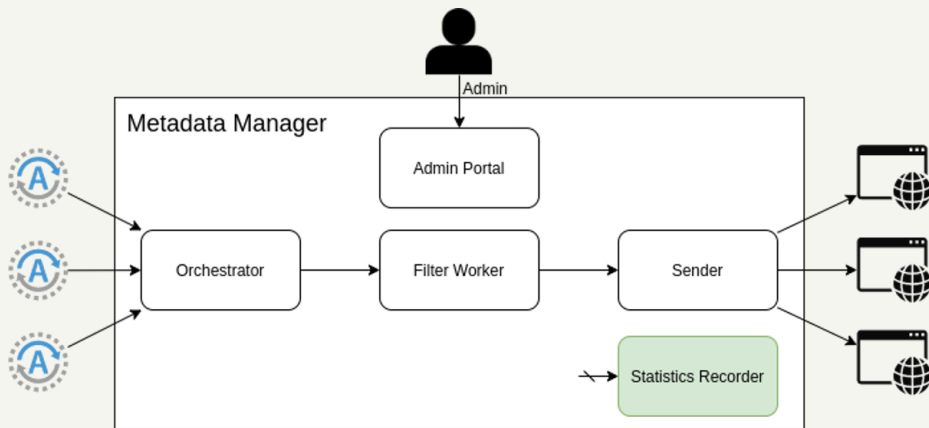
Admin Portal

- Offers an interface to manage the whole Metadata Manager;
- Used to:
 - register new databases;
 - group databases in communities;
 - check statistics to get feedback on data flowing on the system;
 - specify the format template of data to be sent for each application.
- Composed by two components:
 - Frontend - built using React;
 - Backend API - built using Django.



Statistics Recorder

- Stores statistics of the system;
- Implemented as a Kafka Sink Connectors.





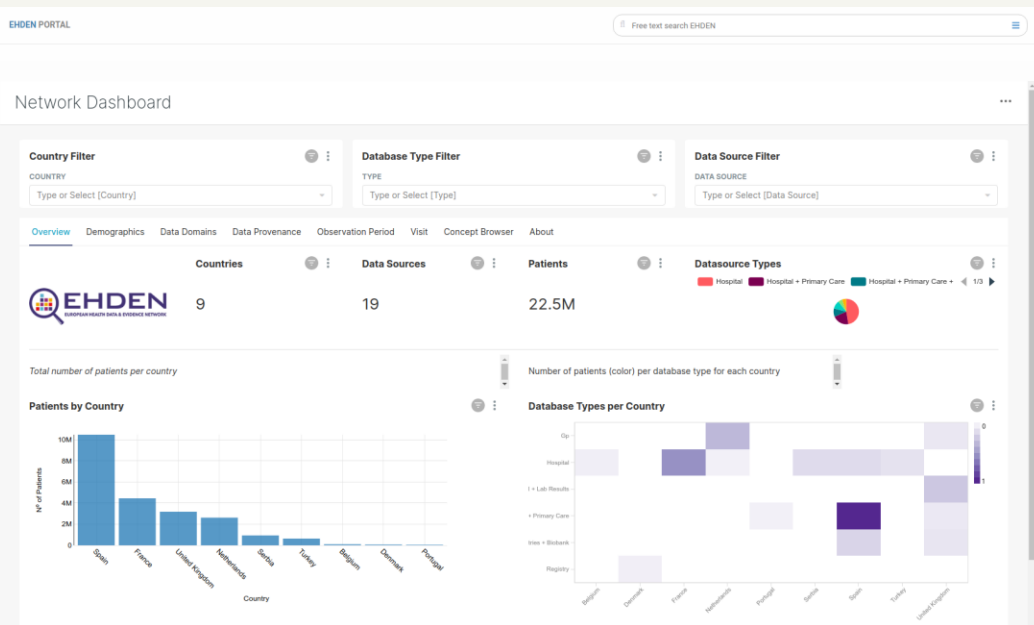
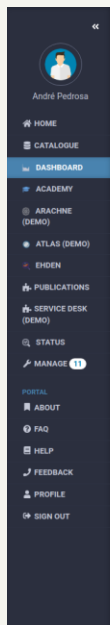
Results

- **Montra Refactoring** - Improvements were done to a fully-fledged tool for metadata visualization;
 - Keep a compatible was a challenge during the development process;
 - Pull request waiting for review with over 150 commits with change on about 200 files.
- **Data Profiling** - A tool able to extract data from databases conforming to the OMOP CDM was proposed:
 - Data owner's data privacy concerns were considered;
 - A System capable of gathering and sending extracted data to the applications was developed
 - Simple components with well-defined objectives;



Conclusions

- EHDEN WP4
- GitHub
- Dashboard



fix: avoid fetching favorite status for anonymous user • size/M

#15590 by aspedrosa was merged on Jul 9 • Approved 1 of 8 tasks

4

fix: no roles being returned for anonymous user • size/S

#15585 by aspedrosa was merged on Jul 27 • Approved 1 of 8 tasks

5

Avoid fetch fav dashboard stat not logged in • size/M

#5827 by aspedrosa was merged on Jan 29, 2020 • Approved 3 of 12 tasks

5

Remove gevent installation as a separate docker layer • size/XS

#8078 by aspedrosa was merged on Aug 21, 2019 • Review required 2 of 12 tasks

1





**Any
Questions?**

