



EHDEN

EUROPEAN HEALTH DATA & EVIDENCE NETWORK

806968 – EHDEN

European Health Data & Evidence Network

WP4 – Technical Infrastructure

D4.1 Technical Framework Design and Architecture

Lead contributor	7- Odysseus
Lead contributor email	gregory.klebanov@odysseusinc.com
Other contributors	1 - EMC 4 - UTARTU 5 - UAVR 6 - The Hyve
Due date	30/06/2019
Delivery date	12/09/2019
Deliverable type	R
Dissemination level	PU
DoA - Version	V1
Date	12/11/2018





	D4.1 - Technical framework design and architecture		
	WP4 - Technical implementations	Version: v1.3 – Final	
	Author(s): Antje Hottgenroth, Gregory Klebanov, Kees Van Bochoven, Nigel Hughes, Maxim Moinat, Michel Van Speybroeck, Sebastiaan van Sandijk, Peter Rijnbeek	Security: PU	2/33


TABLE OF CONTENTS

Table of contents	2
Document History	3
Definitions	4
Publishable Summary	8
1. Introduction	9
Principles underlying the EHDEN framework.....	9
2. EHDEN Business Processes	10
3. High Level Architecture Overview	11
4. Core Functionality	12
4.1 Database Catalogue.....	12
4.2 Study Designer.....	14
4.4 Analytical tools	15
4.5 EHDEN Network Study Workflow	16
4.6 EHDEN Portal	18
4.7 Study Result Dissemination	19
4.8 Shared Artefacts Repository.....	20
5. Extraction Transform Load (ETL)	21
5.1 ETL Design.....	22
5.2 Vocabulary Mapping.....	22
5.3 OMOP Standardized Vocabulary Explorer and Distribution.....	23
5.4 CDM Validation Tools	23
6. Training and Education	24
6.1 EHDEN Academy.....	24
6.2 Virtual training environment.....	24
6.3 Query Library.....	24
6.4 EHDEN Environments	25
1. (Build) Integration	26
2. Test/QC	26
3. Production.....	26
4. Sandbox.....	26
7. Information Architecture	27
7.1 Information Models.....	27
7.1.1 OMOP CDM	27
7.1.2 Study Information and Execution (ARACHNE)	28
7.1.3 Analysis Design (ATLAS)	29
7.2 Information Flow	30
8. Integration Architecture	31
9. Security	32
Next steps	32
Conclusion	33

	D4.1 - Technical framework design and architecture		
	WP4 - Technical implementations		Version: v1.3 – Final
	Author(s): Antje Hottgenroth, Gregory Klebanov, Kees Van Bochoven, Nigel Hughes, Maxim Moinat, Michel Van Speybroeck, Sebastiaan van Sandijk, Peter Rijnbeek		Security: PU

DOCUMENT HISTORY

Version	Date	Description
V0.1	03 May 2019	Blueprint architecture_version_0.1, including comments
V0.3	21 June 2019	Blueprint architecture_comments_Rev_003_archive, with comments
V1.0	15 July 2019	Draft D4.1-Technical Framework Design and Architecture. Submitted for
V1.1	28 August 2019	Update of deliverable after review by Matthew Wiener and Daniel
V1.2	5 September 2019	Small updates, extended summary after consortium review
V1.3	12 September 2019	Small updates after additional consortium review

	D4.1 - Technical framework design and architecture		
	WP4 - Technical implementations	Version: v1.3 – Final	
	Author(s): Antje Hottgenroth, Gregory Klebanov, Kees Van Bochoven, Nigel Hughes, Maxim Moinat, Michel Van Speybroeck, Sebastiaan van Sandijk, Peter Rijnbeek	Security: PU	4/33




DEFINITIONS

Participants of the EHDEN Consortium are referred to herein according to the following codes:


EMC	Erasmus Universitair Medisch Centrum Rotterdam- The Netherlands (Project Coordinator)
Synapse	Synapse Research Management Partners S.L. - Spain
UOXF	The Chancellor, Masters and Scholars of the University of Oxford - United Kingdom
UTARTU	Tartu Ulikool - Estonia
UAVR	Universidade de Aveiro – Portugal
The Hyve	The Hyve BV – the Netherlands
Odysseus	Odysseus Data Services SRO – Czech Republic
EPF	Forum Europeen des Patients (FPE) - Luxembourg
NICE	National Institute for Health and Care Excellence – United Kingdom
UMC	Stiftelsen WHO Collaborating Centre for International Drug Monitoring - Sweden
ICHOM	International Consortium for Health Outcomes measurement LTD - United Kingdom
Janssen	Janssen Pharmaceutica NV - Belgium (Project Lead)
Pfizer	Pfizer Limited – United Kingdom
Abbvie	AbbVie Inc - United States
IRIS	Institut De Recherches Internationales Servier - France
SARD	Sanofi Aventis Recherche & Developpement - France
Bayer	Bayer Aktiengesellschaft - Germany
Lilly	Eli Lilly and Company Limited – United Kingdom
AZ	AstraZeneca AB - Sweden
Novartis	Novartis Pharma AG - Switzerland
UCB	UCB Biopharma SPRL - Belgium
Celgene	Celgene Management SARL - Switzerland

Grant agreement	The agreement signed between the beneficiaries and the IMI JU for the undertaking of the EHDEN project (806968).
Project	The sum of all activities carried out in the framework of the Grant Agreement.
Consortium	The EHDEN Consortium, comprising the above-mentioned legal entities.
Consortium agreement	Agreement concluded amongst EHDEN participants for the implementation of the Grant Agreement. Such an agreement shall not affect the parties' obligations to the Community and/or to one another arising from the Grant Agreement.
Business capabilities	This refers to those applications or part of applications that are an integral part of the operational EHDEN platform and supporting the business process that is at the heart of the EHDEN platform: the process of preparing and performing a study. Examples include Database Catalogue, ATLAS and ARACHNE.
Business process	

	D4.1 - Technical framework design and architecture		
	WP4 - Technical implementations		Version: v1.3 – Final
	Author(s): Antje Hottgenroth, Gregory Klebanov, Kees Van Bochoven, Nigel Hughes, Maxim Moinat, Michel Van Speybroeck, Sebastiaan van Sandijk, Peter Rijnbeek		Security: PU




Data Source Preparation Applications	<p>A business process is a collection of related, structured activities or tasks carried out by people or equipment in a specific sequence to produce a service or product for a particular customer or customers. A business process may be visible or invisible to customers, by definition it serves a particular business goal.</p> <p>Those components that can be used to build, deploy or initialize a particular data source in the network. Examples are ETL code or White Rabbit / Rabbit in a Hat.</p>
EHDEN framework	<p>An EHDEN specific assembly of different applications or components that can be used to support an end to end study flow i.e. from data discovery to publishing of results. This EHDEN framework will consist of different loosely coupled applications. These applications will largely be derived from the vast amount of applications that already have been developed as part of the OHDSI-initiative. The so-called ‘supportive applications’ are not part of this EHDEN framework.</p>
EHDEN portal	<p>The EHDEN portal will be the main tool through which the EHDEN framework will be accessed. It will be based on a content management system that allows linkage to the different underlying applications. It will also contain all content that users should have to be able to work with EHDEN: an informational website, fora, links to training (EHDEN Academy), procedural documents and links to the tools used in the SME/data source call process.</p>
EHDEN Ecosystem	<p>Under ‘ecosystem’, we refer to the broader context in which the EHDEN framework is being deployed. While the EHDEN framework is a specific technical concept, the EHDEN ecosystem includes also the organizational, procedural / process aspects and the social component of the network including SME’s, data sources as well as other stakeholders.</p>
Data Source	<p>Refers to a single logical source of data – typically associated with a single owner or custodian. A ‘logical’ data source can consist of one or more physical data sources. In that case, when the data is being mapped, the various data sources will be mapped and integrated together in a single OMOP Common Data Model instance. Typically, there will be one organizational unit responsible for a given data source. This entity can be a data custodian – acting on behalf of one or more data owners- or a data owner itself. On the reverse side, it’s equally possible that a data custodian or data owner manages multiple data sources e.g. an oncology registry and a GP database.</p>
Database	<p>A single physical database. A single entity e.g. a hospital might have databases e.g. a patient database, a radiology database, a lab results database, but it can also be a clinical research database where data is already integrated.</p>
Data Custodian	<p>The team (typically a legal entity) that is responsible for the management and governance of a data source and/or one or more databases. A data custodian might be the original data controller / owner of a database or they might perform this activity on behalf of other parties. The data custodian will typically be the main partner in the mapping of a data source to the OMOP CDM.</p>

	D4.1 - Technical framework design and architecture		
	WP4 - Technical implementations		Version: v1.3 – Final
	Author(s): Antje Hottgenroth, Gregory Klebanov, Kees Van Bochoven, Nigel Hughes, Maxim Moinat, Michel Van Speybroeck, Sebastiaan van Sandijk, Peter Rijnbeek		Security: PU




Data Partner	A Data Partner is anyone (individual or legal entity) that has agreed to participate in EHDEN with their respective data source(s). This might be a data custodian but in certain cases, it might be the assembly of original data providers that acts as partner(s)
Data Mapping	Data mapping is the process by which a local data source is converted to the OMOP common data model. This is typically done using an ETL (Extract – Transform – Load) process. The ETL process contains a structural component (converting from one particular table / field structure to the OMOP CDM structure) and a semantic component (ensuring that the meaning of the variables is preserved when mapping from local standards to OMOP vocabulary standards)
Data Discovery	Data discovery is often the first step in a research question. It's that activity where an initial selection of likely data sources is made that have a high likelihood of containing the right variables and the right subject population to address a particular question. The data that are used to address the discovery step are on the one hand descriptive data of each individual data source (e.g. who is the owner or custodian, what is the scope of the data source, what local governance process is applicable...). On the other hand, summary level data that are obtained from the (mapped) data source can be used as well to support the data discovery process. These summary level data are typically univariate data relating to each of the distinct entities of the OMOP CDM (e.g. drug exposure, persons, condition_occurrence,...) and processed in such a way not to expose a privacy disclosure risk.
Feasibility Assessment	A feasibility assessment is the next logical step in a research question. It will allow fine grained definition of the population of interest through the definition of inclusion / exclusion criteria as well as an identification that for that population the variables of interest are present.
Data Profiling	Data profiling is a technical operation whereby descriptive statistics on a particular dataset are collected. This data profiling can occur on the original source data or on the dataset mapped to OMOP. White Rabbit is an example of a tool that can be used for data profiling
SSO, Single Sign-On	Single Sign-On is a property of access control of multiple related, but independent, software systems through a single authentication mechanism and standard.
A&A, Authentication and Authorization	Authentication and Authorization is two steps process of identifying a user/account trying to access a system and allowing that user to gain access to the authorized resources.
OpenID	OpenID is an open standard sponsored by Facebook, Microsoft, Google, PayPal, Ping Identity, Symantec, and Yahoo. OpenID allows user to be authenticated using a third-party service called "identity providers". Users can choose to use their preferred OpenID providers to log in to websites that accept the OpenID authentication scheme.
SAML, Security Assertion Markup Language	Security Assertion Markup Language is an open standard for exchanging authentication and authorization data between parties, in particular between an identity provider and a service provider. It was developed by OASIS and is an XML-based open standard for exchanging authentication and authorization data between parties. Here is an explicit trust between Service Provide and Identity Provider.

	D4.1 - Technical framework design and architecture		
	WP4 - Technical implementations		Version: v1.3 – Final
	Author(s): Antje Hottgenroth, Gregory Klebanov, Kees Van Bochoven, Nigel Hughes, Maxim Moinat, Michel Van Speybroeck, Sebastiaan van Sandijk, Peter Rijnbeek		Security: PU

Identity Propagation

Identity Propagation is an ability of an application to propagate verified user's identity through all application layers - UI, API and down to database. While being considered the ideal state, this is a very difficult thing to achieve in federated case and implementation with diverse environments where each tier and component must support the same IP mechanism.

	D4.1 - Technical framework design and architecture		
	WP4 - Technical implementations	Version: v1.3 – Final	
	Author(s): Antje Hottgenroth, Gregory Klebanov, Kees Van Bochoven, Nigel Hughes, Maxim Moinat, Michel Van Speybroeck, Sebastiaan van Sandijk, Peter Rijnbeek	Security: PU	8/33

PUBLISHABLE SUMMARY

The core objective of the EHDEN consortium is to provide all the necessary services that enable a sustainable distributed European data network to perform fast, scalable, and highly reproducible research, while respecting privacy regulations, local data provenance and governance. This includes services and tools to perform data standardization, analytical pipelines, tools to share study results and tools for stakeholder engagement and training. WP4 in EHDEN will implement a technical framework and also all necessary processes for a high-quality and fully reproducible data workflow. Core to EHDEN is open science, which entails the use of open source tools, an open federated health data network, and an engaged community that shares the EHDEN vision and mission.

This deliverable provides a high-level overview of the technical framework and describes the functionality of all the components. Furthermore, we explain which components will need to interact and, to some extent, how they will be integrated.

The underlying principles of the architecture are maximal usage of available and open source tools, a federated data network approach enabling data profiling, data assessments and full studies, data quality assurance and interactive dashboards. The implementation will include a modular framework fit for up-scaling as well as access security measures.

Building the EHDEN ecosystem is clearly a socio-technical challenge. EHDEN must serve a wide range of stakeholders, including industry, regulators, academia, health care system stakeholders, patient organisations, etc. The needs of various stakeholders are considered during the build of the technical framework.

The central components, which will be accessible via the EHDEN portal, are:

- 1) Database Catalogue (Database characteristics, partly automated meta-data generation)
- 2) EHDEN Network Study Workflow Platform (ARACHNE), including Data Node
- 3) Study Designer (ATLAS)
- 4) Dashboards
- 5) EHDEN Academy (Training)


The central platform will interact with local components at Data Partners and Researchers (e.g. OMOP mapping, ETL tools).

The core hosting infrastructure is hosted on the AWS platform and co-located in the same region and data centre, with an exception of the ELIXR A&A system, with detailed provider later in this document.

The EHDEN ecosystem will implement a Single Sign On (SSO) approach to security, including authentication and authorization, with ELIXR being used as an identity provider.

The systems will be integrated e.g. each system will master manage and share with other system its core data sets, as defined in the common data model (to be defined).

This document provides an architecture blueprint for the EHDEN framework. It lays a foundation for further work which will continue to evolve during the project lifetime.

	D4.1 - Technical framework design and architecture		
	WP4 - Technical implementations		Version: v1.3 – Final
	Author(s): Antje Hottgenroth, Gregory Klebanov, Kees Van Bochoven, Nigel Hughes, Maxim Moinat, Michel Van Speybroeck, Sebastiaan van Sandijk, Peter Rijnbeek		Security: PU



1. INTRODUCTION

WP4 is responsible for the technical coordination and implementation of the EHDEN framework and all its services. This WP is at the core of EHDEN. WP4 collects requirements from the other work packages and translates these to specific technical implementations. The aim of WP4 is a **fully integrated, highly efficient framework** to enable re-use of standardized health data across Europe. Our consortium envisions a socio-technical framework that enables researchers to **find data, re-use data, and share analytical methods and study results**.

Principles underlying the EHDEN framework


WP4 is based on the following major design choices:

1. We will profit from the extensive experience acquired by participants in related projects and will **maximally use available tools** (specifically the OHDSI open source tools). This allows us to obtain a first implementation fast, based on these existing tools, and further optimize to a fully integrated solution (EHDEN portal) using a cyclic agile approach.
2. All tools developed in the project will be **open source** to enable a community-wide development process and full transparency for all stakeholders. All developed analytical methods will be made available to the community to enable replication in each step of the process.
3. The framework is based on the **federation of databases** mapped to the OMOP-CDM in which each data custodian retains control over its data. We will support different levels of data sharing to enable re-use, e.g., for analytical purposes, in different privacy and governance settings. The General Data Protection Regulation (GDPR) will be enforced in the design.
4. Mechanisms will be implemented to support different phases of the study process in the federated data network, including **data profiling, study feasibility assessments and full studies**.
5. Throughout the framework, benchmarking will be applied to ensure **high-quality data** in data sources mapped to the OMOP-CDM. This includes methods to ensure there is no loss of information in the database mapping step, validation of analytical methods, quality assessment of results, etc.
6. We will adopt and/or develop a framework for **interactive dashboards** to facilitate dissemination of data source characteristics, study results and benchmarking results.
7. The implementation is based on an **extensible modular framework**. Although the selected use cases and application domains will drive the initial development cycles, the aim is to produce a solution that can scale up to many more data sources, analytical methods etc.
8. **Security measures** will be put in place in the integrated environment following common approved standards for authentication and authorisation.

Further, the EHDEN framework is implemented to make the data Findable, Accessible, Interoperable and Reusable (FAIR) across the network. This with a view to support transparency and reproducibility of research in the network.

1. **Findable**. Data is made findable by further extending the data catalogue developed in EMIF that holds metadata. We will complement this with profiles generated directly from the OMOP-CDM. Different levels of data access are implemented to accommodate data sharing preferences of the data custodians.
2. **Accessible**. Federation is operationalized through the ARACHNE platform. This advanced tool puts the data custodian in control of data interrogation and sharing of study results in a OMOP-CDM data network. In short, the user will receive study requests in a web tool, has full access to review the



	D4.1 - Technical framework design and architecture		
	WP4 - Technical implementations		Version: v1.3 – Final
	Author(s): Antje Hottgenroth, Gregory Klebanov, Kees Van Bochoven, Nigel Hughes, Maxim Moinat, Michel Van Speybroeck, Sebastiaan van Sandijk, Peter Rijnbeek		Security: PU

analytical code, can accept or declined the request, run the study, review the study results, and approve results sharing. All can be done in a very efficient and transparent workflow.

3. **Interoperable.** The OMOP-CDM and standard vocabularies enable a high-level of interoperability. By adopting the ATLAS tool (<http://www.ohdsi.org/web/atlas/>), we fully integrate all the powerful functionality developed in the OHDSI community, i.e. vocabulary browsing, cohort definitions, study generation etc.
4. **Reusable.** Mechanisms will be put in place to promote re-use of all study components, e.g. cohort definitions, study specifications, analytical code etc. We believe this will have a high impact on the efficiency of future study execution. Moreover, it is crucial to create a ‘research memory’ in the order to maximally benefit from obtained knowledge and experience.


2. EHDEN BUSINESS PROCESSES

For the development of the technical infrastructure it is important we understand who the users of the system will be and what their business requirements are. Sustainable solutions cannot be built without proper stakeholder mapping and an intense interaction with those stakeholders. Understanding the requirements and constraints of the stakeholders will therefore be a continuous effort in the next steps following up on this document, in close collaboration with WP6 “Outreach and Sustainability”.

Building the EHDEN ecosystem is clearly a socio-technical challenge. EHDEN must serve a wide range of stakeholders, including industry, regulators, academia, health care system stakeholders, patient organisations, etc. Each BD4BO project, for example, that will use the EHDEN framework will have a different stakeholder composition. A data partner may want to use tools to obtain high exposure to future users, tools for quality assessment of their mapping process, and/or opportunities for participation in a large data network. Data providers will, at the same time, need to comply with local governance and ethical requirements, which can be challenging. Conversely, data users may require a single access point to data and rich analytical tools for a wide range of study designs that can efficiently be distributed in the data network. This user will likely seek quick and easy access to data for remote analysis and/or study execution to generate high-quality evidence. This requires relevant governance procedures and a rich user-friendly toolbox.

Standardized tools for dissemination of study results may be of interest to a wider community, e.g., sharing the results of a drug utilization study, or prediction tool, in the data network using an interactive dashboard as supplementary data to a more traditional, static research paper. Moreover, the EHDEN ecosystem will contain stakeholders that will provide services, like the Small Medium Enterprises (SMEs) that will be involved in data source mapping and local tool installation. All these aspects and interests will be considered while engaging with our stakeholders both in and outside of the BD4BO projects as this will serve to design and test engagement policies and tool adoption.

Figure 1 describes the Conceptual Framework the users are interacting with.

	D4.1 - Technical framework design and architecture	
	WP4 - Technical implementations	Version: v1.3 – Final
	Author(s): Antje Hottgenroth, Gregory Klebanov, Kees Van Bochoven, Nigel Hughes, Maxim Moinat, Michel Van Speybroeck, Sebastiaan van Sandijk, Peter Rijnbeek	Security: PU

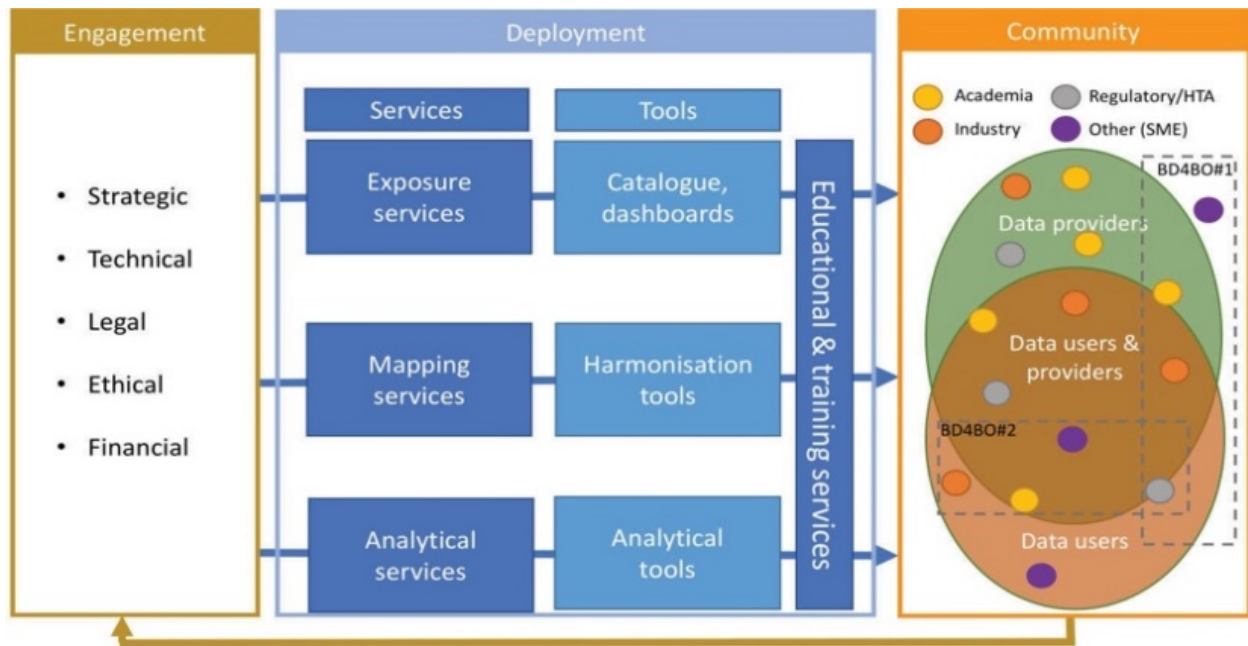


Figure 1: The EHDEN Conceptual Framework.


In what follows, this document discusses user requirements both for the framework as a whole and for each of its components. Main functional capabilities include finding relevant data through a database catalogue, enabling high-quality data standardization using harmonization tools, enabling analytical studies in the federated network, and training tools.

3. HIGH LEVEL ARCHITECTURE OVERVIEW

The **Core** EHDEN components as shown in Figure 2 include:

- 1) EHDEN Portal
- 2) EHDEN Network Study Workflow Platform (ARACHNE), including Data Node
- 3) Study Designer (ATLAS)
- 4) Database Catalogue
- 5) Dashboards
- 6) EHDEN Academy (Training)

as well as various **Supporting** ETL and Mappings tools.

	D4.1 - Technical framework design and architecture	
	WP4 - Technical implementations	Version: v1.3 – Final
	Author(s): Antje Hottgenroth, Gregory Klebanov, Kees Van Bochoven, Nigel Hughes, Maxim Moinat, Michel Van Speybroeck, Sebastiaan van Sandijk, Peter Rijnbeek	Security: PU 12/33

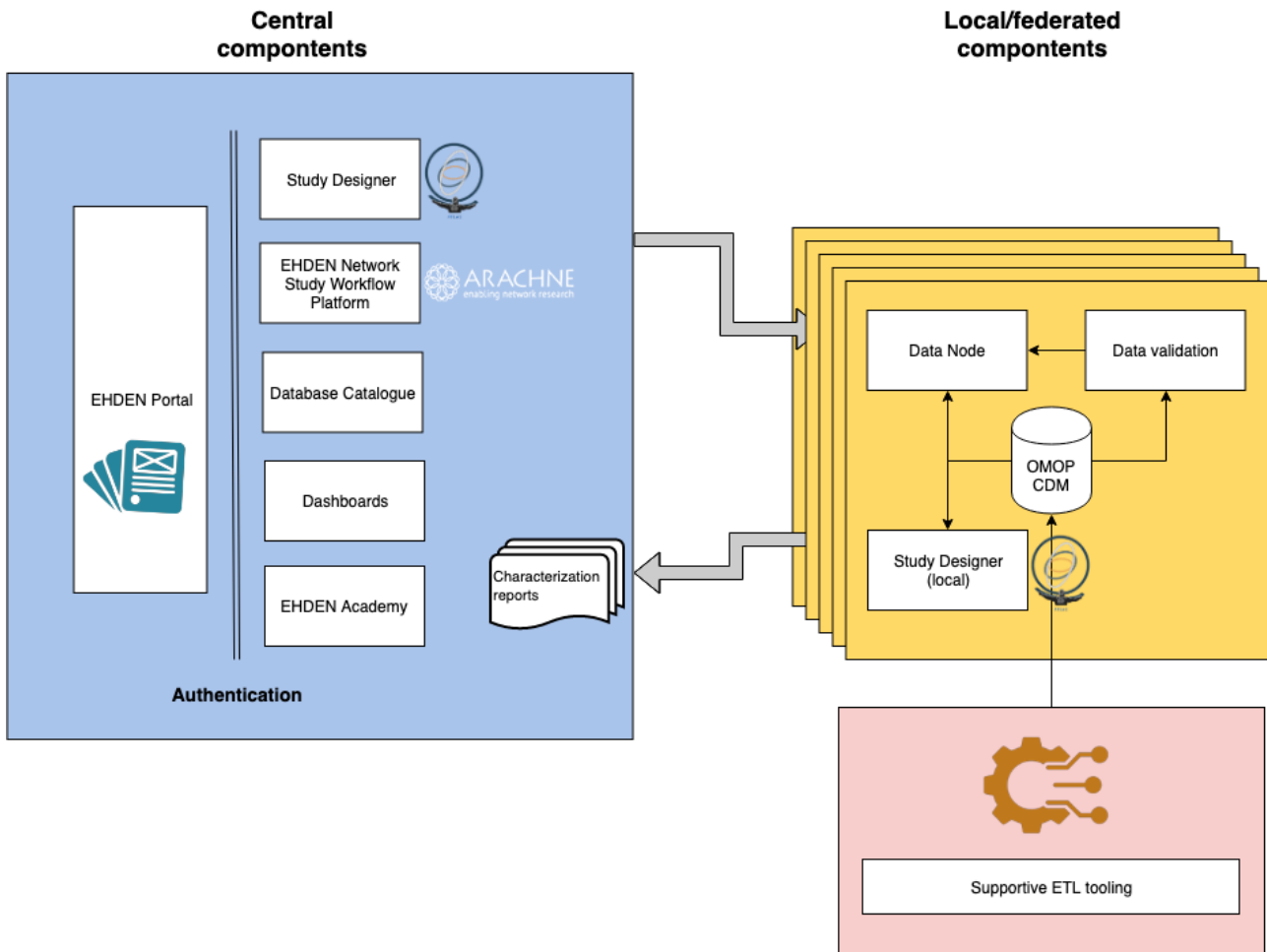


Figure 2: The EHDEN Components.

The core tools will be user-facing and will enable the execution of research utilizing EHDEN platform, while supporting tools will be used to enable (a.k.a. prepare) various sites to be a part of the EHDEN Research Network. This would include the conversion of the internal data into OMOP CDM, mapping internal vocabularies into the OMOP Standardized Vocab and enabling OHDSI tools and EHDEN platforms on the participating site. Via the EHDEN platform, participating sites get connected to the EHDEN Network and community - and vice versa.


4. CORE FUNCTIONALITY

In the sections below a high-level overview is provided for each of the components of the technical framework. We describe the functional requirements based on input from the users and will provide insights in the implementation of these requirements.

4.1 Database Catalogue

Functional Requirements:

The purpose of the Database Catalogue is to make data “**Findable**”, so it contains metadata describing a data source. The Database Catalogue is intended to be the primary point to start the data discovery process. The Database Catalogue will host a searchable questionnaire with extrinsic metadata, e.g. contact details and governance procedures. Furthermore, it will host intrinsic metadata, i.e. information we can extract

	D4.1 - Technical framework design and architecture		
	WP4 - Technical implementations	Version: v1.3 – Final	
	Author(s): Antje Hottgenroth, Gregory Klebanov, Kees Van Bochoven, Nigel Hughes, Maxim Moinat, Michel Van Speybroeck, Sebastiaan van Sandijk, Peter Rijnbeek	Security: PU	13/33

automatically from the CDM. The aim of the intrinsic information in the Database Catalogue is to provide high level database characteristics, e.g. CDM version, vocabulary version, the number of patients, etc. This intrinsic metadata will be made available through visualizations or tables whatever is more appropriate. Note that the aim is not to replace feasibility studies in which, for example, cohort definitions are distributed to the data partners using federation. This is out-of-scope for the Database Catalogue and will be handled by other tools such as Atlas and Arachne described in this deliverable.

The Database Catalogue will be mainly used by researchers that want to initiate a study and data partners that want to promote their data sources.

The following user requirements need to be addressed:


1. The Database Catalogue needs to contain up-to-date information. This implies the need of procedures to engage the Data Custodian – e.g., into keeping Catalogue information up-to-date - but also technical functionality to make this a lightweight process.
2. It needs to contain all relevant data for the user to make a first high-level assessment on the suitability of the data source for a specific research question. The tool therefore needs to be very flexible to allow for further improvement during the project based on user input.
3. The database catalogue needs to have powerful and user-friendly search capabilities
4. The tool needs to be web-based and needs to have a set of robust API-endpoints to allow other tools to interact with data in the Database Catalogue.
5. At least a portion (or subset of metadata) of the Database Catalogue will be publicly accessible. Full visibility will be configurable by data stewards and allow multiple levels of access, including to be restricted to EHDEN (registered research network users) or study members only.
6. Information in the Database Catalogue needs to very accessible and easy to understand, for example by providing visualizations of the data.
7. The user needs easy access to publications produced on the data source.
8. There needs to be one single database catalogue for the EHDEN framework.
9. The Catalogue should allow several user roles, which may be associated to distinct access rights.
10. The sign-in process must allow several federated identity providers, such as ELIXIR and ORCID.

Implementation:

The existing EMIF catalogue will act as a powerful basis for the EHDEN database catalogue. However, the current EMIF catalogue is built around the concept of communities, i.e. it hosts several projects. For EHDEN, a tailored solution will be created that will only host the data sources in the EHDEN data network (and eventually the OHDSI Network). It will be possible to define new data sources, enter the descriptive metadata. Furthermore, the OHDSI ACHILLES tool (see www.github.com/OHDSI/Achilles) will be extended to automatically extract the intrinsic metadata from the CDM. The output of this tool can be uploaded to the Database Catalogue to populate the questionnaire but will also enable visualizations on the database and data network level. User requirements for the visualizations are currently being defined with multiple stakeholders and will be reported on in the yearly progress report (D4.3). Finally, the database catalogue will be integrated with the rest of the technical framework, including ARACHNE (described later in more detail).

Roles:

Administrators can administer data catalogue, including creating and managing common meta data and govern requests to register new data sets

	D4.1 - Technical framework design and architecture		
	WP4 - Technical implementations		Version: v1.3 – Final
	Author(s): Antje Hottgenroth, Gregory Klebanov, Kees Van Bochoven, Nigel Hughes, Maxim Moinat, Michel Van Speybroeck, Sebastiaan van Sandijk, Peter Rijnbeek		Security: PU

Data Stewards - besides administrators who could e.g. define additional attributes and/or configure the search capabilities, data stewards will be able to create new data sources, manage access, approve user requests, manage ACHILLES access and provide the annotation.

Users - catalogue users who can perform search, view results and request access to use data in studies (Study Leads and other roles)

4.2 Study Designer

Functional Requirements:

We need a powerful tool that enables the specification of the network studies. This requires search functionality for the vocabulary, code set definitions, cohort definitions, a rich set of study designs and the functionality to fully specify the analytical pipeline. Furthermore, the tools should support the Data Partner in leveraging their own data by providing database dashboards and an engine to execute studies directly to their owned assets.

The users of this functionality are: Data Partner, Study Coordinator, Data Analyst.

The following user requirements need to be addressed:

1. Study designs must encapsulate the full study specification (cohort definitions, study protocol, etc.) and must be exchangeable amongst study team members and/or members of the EHDEN network. The tool must support the import and export of study designs.
2. Designs must follow the semantic versioning scheme (<https://semver.org/>) to establish the provenance of the tools used in the design.
3. The tool needs to be web-based and have a set of robust API-endpoints that allow for the retrieval of designs and executed results in a local environment.
4. The tool must use the same security mechanisms for authentication and authorization as other components of the EHDEN architecture
5. The process for updating the tool must be clear, documented and support the requirements of the EHDEN network.


Implementation:

For this we will fully use the very powerful ATLAS tool developed by the OHDSI community (www.github.com/ATLAS/OHDSI). ATLAS is a free, publicly available web-based, open-source software application developed to support the design and execution of observational analyses to generate real world evidence from patient level observational data. ATLAS is an open science analytics platform that can be installed locally within an institution to perform analyses across one or more observational databases which have been standardized to the OMOP Common Data Model and can facilitate exchange of analysis designs with any other organizations across the OHDSI community who have adopted the same open science community standards and tools.

Dataset characterization: The current 'Data Sources' functionality provides a dashboard for the data partner.

Concept Set definition: Atlas allows to define concept sets, i.e. a collection of concepts to define for example a condition. These concept sets are building blocks in the cohort definitions.



	D4.1 - Technical framework design and architecture		
	WP4 - Technical implementations		Version: v1.3 – Final
	Author(s): Antje Hottgenroth, Gregory Klebanov, Kees Van Bochoven, Nigel Hughes, Maxim Moinat, Michel Van Speybroeck, Sebastiaan van Sandijk, Peter Rijnbeek		Security: PU

Cohort definition: this is the basis of the feasibility analysis and other more advanced analytical analyses and will be used extensively in EHDEN. Future work might focus on usability aspects of the UI and include additional elements such as version control, provenance tracking etc.

Analysis Design: This is currently partially covered. It's expected that the existing library of methods accessible through Atlas gets extended, under guidance from WP1 through 3.

There will be one central Atlas instance, accessible through the EHDEN portal and individual Atlas instances associated with the local data sources. The central instance will not be connected to the data sources, but can be used to collaborate on cohort definitions, analysis design. In EHDEN we will review the current functionality and may propose and developed additional features in close collaboration with the OHDSI community.

Roles:

Administrators can administer tool installation, including creating and managing data sets, database level characterizations and control RBAC (user access management).

Users can create, edit and execute (local data) analysis designs. The Users role will be split into more fine-grained roles that can be used to manage access to features as well as data sets - e.g., see under functional requirements.


4.4 Analytical tools

Functional Requirements

OHDSI has developed a large set of analytical methods, e.g. patient-level prediction, risk effect estimation including propensity score matching, and study designs such as self-controlled case series. The work in EHDEN will be heavily driven by the analytical framework pioneered in OHDSI and will further extend this together with OHDSI. It is critical that we build tools that can be used across the world. The task of EHDEN is **not** to implement and support all possible questions a researcher could ask the data network, nor to build specific queries for the other BD4BO project. EHDEN's role is to provide a solution that allows others to easily add new analytical tools. For this it is important we implement software quality criteria. To drive this development each application domain in WP1-3 has a task to re-use and further develop the necessary analytical code following predefined quality standards such as developed in the OHDSI community. For example, the current activities in WP1 on Drug Utilization will require the development of a new analytical pipeline that will extract the necessary data from the CDM, will perform analyses on this data, and will generated dashboards with results that can be further disseminated.

Furthermore, the analytical pipelines need to be executed across the data network following a federated approach. The process of conducting a study in OHDSI has so far been partly manual, requiring a significant effort and a number of different, unlinked communication channels, including email exchanges, forums, and chats. Once in the execution phase, adapting code to different environments often requires tweaking and these changes and activities are typically not tracked. The statistical code and related results are not linked, and often stored in personal folders on the scientists' file systems.

Software validation needs to be a high priority for the analytical pipeline development. This includes architecture standards, including standard models, interfaces, components and blueprints, best coding practices, error handling, versioning and configuration management, and extensive tool documentation. We will follow established software development principles, including continuous integration, test-driven development, test automation, continuous testing, code reviews, and code profiling. Moreover, we will

	D4.1 - Technical framework design and architecture		
	WP4 - Technical implementations		Version: v1.3 – Final
	Author(s): Antje Hottgenroth, Gregory Klebanov, Kees Van Bochoven, Nigel Hughes, Maxim Moinat, Michel Van Speybroeck, Sebastiaan van Sandijk, Peter Rijnbeek		Security: PU

instantiate separate environments for development, testing, pre-production, and production. Finally, we will define a framework for functional testing, non-functional testing, security testing and access controls, and compliance testing.

We can identify the following users of the analytical pipeline: the researcher that wants to develop a new method, the study coordinator that wants to execute an analytical pipeline, a user that wants to have access to the results of the analytical pipeline.

The following user requirements were identified:

1. The user that applies an existing method needs to trust the analytical pipeline. This requires proper documentation and full transparency of the applied methods and follows the [software validity guidelines](#) (such as unit tests). Furthermore, validation methods are needed for both the results of a specific analysis as well as the method itself. Measures should be put in place to assure that the planned analysis is valid for the available data and systematic errors need to be flagged.
2. There needs to be clear specification available for those that want to add a specific analytical pipeline. For example, what are the minimal requirements for the R packages that will be distributed through the federation tool?
3. For sharing results a common solution is needed that the analytical pipeline developers can adopt. This solution needs to be very user-friendly for those that generate the results and those that consume the results.

Implementation:

For this we will build on the analytical R pipeline developed by OHDSI and the quality control mechanisms build in that process, e.g. unit testing, documentation requirements etc. We do need to develop a guidance document for external tool developers on how to extend the methods library and need to specify a standard approach to sharing of results. The decision to include study design in Atlas for a new method will be made by the team in close collaboration with OHDSI.

There will be support for an improved network study process orchestration by ARACHNE Network platform, including the ability to import and share design and code, execute and capture the results for various types of analyses, including descriptive and predictive types.

4.5 EHDEN Network Study Workflow

Functional Requirements:


The EHDEN Network Study process is a workflow that involves multiple study collaborators and participating organizations.

EHDEN platforms and tooling will support a complete study lifecycle, which includes three distinct stages:

- Study feasibility
- Study execution
- Study results dissemination

Study Feasibility

The study feasibility stage (pre-study stage) is focused on supporting a definition of a study and the creation of a study protocol, e.g. making sure the study is feasible to be executed as described in the formal protocol.

	D4.1 - Technical framework design and architecture		
	WP4 - Technical implementations	Version: v1.3 – Final	
	Author(s): Antje Hottgenroth, Gregory Klebanov, Kees Van Bochoven, Nigel Hughes, Maxim Moinat, Michel Van Speybroeck, Sebastiaan van Sandijk, Peter Rijnbeek	Security: PU	17/33

The feasibility stage does not have a well-defined process but rather is driven by various supporting activities, including identification and enrolment of relevant databases that contain patient population with required drug exposure, procedure information, condition or demographics information through data characterization, validating and agreeing on target analytical methods and algorithms.

The outcome of the feasibility stage is a formal protocol as well as a list of target collaborators.

Study Execution

The following are the key activities in executing an EHDEN network study:


1. The *Study Lead* initiates a new study.
2. The *Study Lead* publishes study protocol.
3. The *Study Lead* invites participating centres and collaborators, including request ability to use databases within a study.
4. *Study Participating Organizations* assemble teams within each site, assign study roles.
5. The *Data Scientists* (statisticians) use a study protocol to design study analyses and generate study code.
6. The *Data Scientists / Statisticians* submit analysis code for execution to participating sites and generate results in the standardized format following OHDSI guidelines. Each participating site will follow internal institutional processes.
7. The *Data scientists / statisticians* and *Study Lead* collect and review the analysis execution results.
8. Iterate steps 5-7, if reasonable adjustments required.
9. Collaboratively finalize and study lead disseminates study results within team and/or network.
10. *The Study Lead* closes the study.

Study Results Dissemination

Study Results Dissemination is a separate process that continues after the immediate execution, which includes (approval for) dissemination of the initial study results (step 9 above). This will be worked out further in later deliverables – see paragraph 4.7 of this document.

Implementation:

ARACHNE will act as the study collaboration and workflow platform. ARACHNE will manage the end to end study execution, provide traceability between different steps, and implement a coherent security and compliance mechanism. ARACHNE will be integrated with EHDEN Database Catalogue to allow discovery of studies and requests for access to the federated data sets.

	D4.1 - Technical framework design and architecture	
	WP4 - Technical implementations	Version: v1.3 – Final
	Author(s): Antje Hottgenroth, Gregory Klebanov, Kees Van Bochoven, Nigel Hughes, Maxim Moinat, Michel Van Speybroeck, Sebastiaan van Sandijk, Peter Rijnbeek	Security: PU

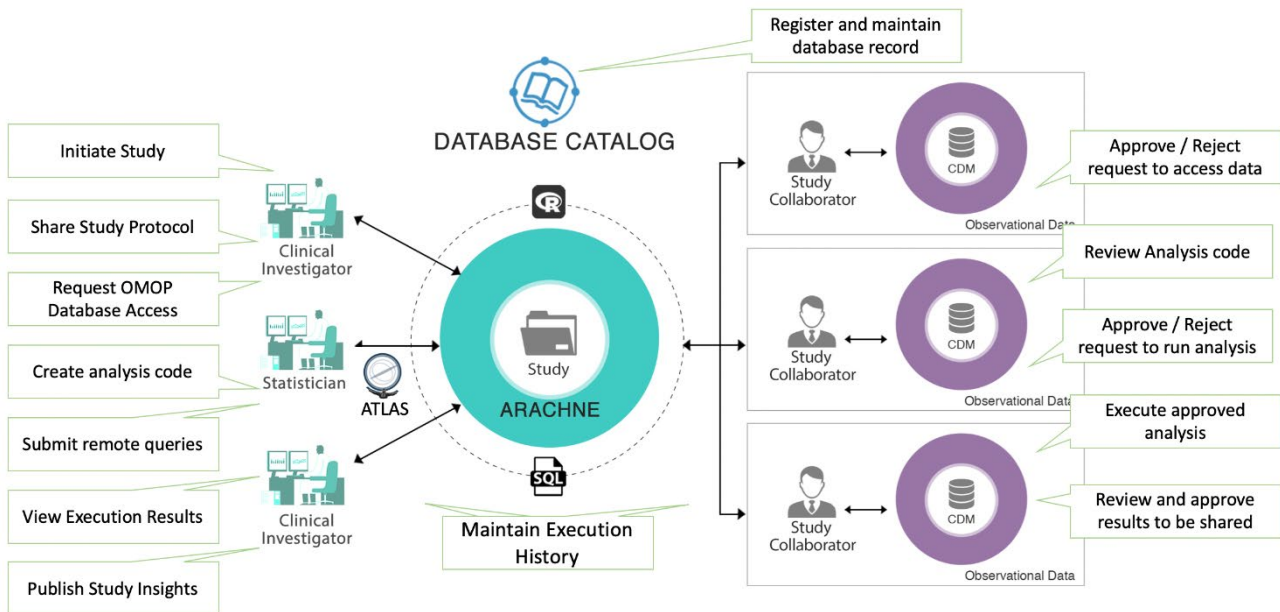


Figure 3: The EHDEN Network Study Workflow

There will be one central EHDEN ARACHNE Central, one EHDEN Sandbox data node and individual data nodes at all participating centres.

Prior to conducting a Study, each participating data organization must register and maintain the database record in the EHDEN Database Catalogue, after which studies can explicitly request using that database for their research.

ARACHNE will feed the tools to be used for Study Results dissemination but will not enable that workflow directly.


Roles:

- **Administrators** have an ability to manage the Research Network (ARACHNE), including managing nodes, A&A mechanisms and integration to other tools and environments.
- **Study Leads** create new studies and manage study properties, invite study participants and assign study roles, request access to data sets to be used in studies, create and execute analyses and finalize study results
- **Study Collaborators** have an ability to create and execute study analyses, annotate results
- **Data Stewards** manage request and approval to use organizational data sets in studies as well as requests for analysis execution

4.6 EHDEN Portal

Functional Requirements:

The EHDEN portal will be a website that provides access to the individual applications using a Single Sign-On (SSO) mechanism. It will also act as documentation repository, gateway to the SME certification process, EHDEN platforms (ARACHNE, ATLAS and Database Catalogue) and the EHDEN academy content.

	D4.1 - Technical framework design and architecture		
	WP4 - Technical implementations	Version: v1.3 – Final	
	Author(s): Antje Hottgenroth, Gregory Klebanov, Kees Van Bochoven, Nigel Hughes, Maxim Moinat, Michel Van Speybroeck, Sebastiaan van Sandijk, Peter Rijnbeek	Security: PU	19/33

Conversely, applications are expected to be setup in such a way that they can be made accessible through the portal and can receive credentials from the EHDEN portal for authentication of the respective users.

Implementation:

The EMIF Catalogue contains all the necessary functionality to be used as the EHDEN portal and will be further improved. The same back-office architecture, MONTRA, will be employed with redesigned user interfaces. The portal will link to several EHDEN related services and functionalities, such as the Database Catalogue, ARACHNE, ATLAS and the EHDEN Academy. The process of linking to different functionalities will be supported by a modular plugin-based architecture, i.e., each portal core component will be treated as a plugin that can be added or removed according to user profile and preferences.

Roles:

- **Administrators** will manage the portal, add or remove plugin functionalities and handle role-based user access
- **EHDEN participants** will be able to browse the Database Catalogue and interact with the remaining EHDEN components according to their profile and respective permissions

4.7 Study Result Dissemination

Functional requirements:

Currently, the de facto standard for sharing results of a study is to publish a paper with static tables and figures. We will improve this by adopting and further expanding the existing frameworks available for sharing study results in interactive dashboards. This will create a new dynamic source of information augmenting the more static scientific paper, therefore enhancing the visibility and impact of the research conducted using the EHDEN data and tools. In EHDEN we will produce fully transparent and reproducible pipelines for study execution. This implies that each study will share - within study team and/or EHDEN community – all the code to enable other data sources to rerun the study and upload their results to the Study Result Dashboard. EHDEN is expected to adopt and further expand a dashboard framework that makes it easy to create additional dashboards but also allows proper access management if required by the study team.


The users of this tool are: the data scientist or statistician that builds a new analytical pipeline, the analytical team that wants to share the results, and the consumer of the results on the interactive website.

The following user requirements were defined:

1. The solution should be able to provide a highly user-friendly experience for viewing study results.
2. Minimal requirements need to be available to assure consistency and quality control on the created dashboards.
3. Training material is needed to make the development of a new dashboard easier.
4. A technical solution is needed to post the dashboards on. This could be a central EHDEN server that hosts all the study results.
5. The solution should be responsive even when there is a high load.
6. Access control is needed to allow sharing of results to a subset of users, e.g. during study execution.
7. Statistics should be provided on the visitors of a specific dashboard.

Implementation:

The current solution used in OHDSI is R-Shiny Dashboards that are hosted on <http://data.ohdsi.org> We will further explore the use of RShiny and possibilities to add functionalities for access control. Moreover, we like to improve the user-interface of the central website that hosts all the studies. This will include a standardized

	D4.1 - Technical framework design and architecture		
	WP4 - Technical implementations	Version: v1.3 – Final	
	Author(s): Antje Hottgenroth, Gregory Klebanov, Kees Van Bochoven, Nigel Hughes, Maxim Moinat, Michel Van Speybroeck, Sebastiaan van Sandijk, Peter Rijnbeek	Security: PU	20/33

interface describing the purpose of the study, links to the study code, list of contributors, references to publications etc.

Roles:

- **Administrators** will manage the website hosting the results dashboards and the supervise the access control
- **Study lead** will be able to push results and control the access.
- **Study collaborators** will be able to share and discuss study results
- **Study leads and collaborators in other teams** will be able to browse, search, examine study results of previous and current studies, in accordance with access settings.

4.8 Shared Artefacts Repository

To enable consistency and re-use across studies, we should create a single repository of versioned shared artefacts. Such shared artefacts may include:


- Cohort definitions (Phenotyping)
- Analyses Definitions
- (Pre-defined) Outcome Measures for benchmarking

This type of functionality will improve the usability of the platform. It makes it easier for users to re-use cohort definitions etc. from previous studies and/or analyses. Moreover, a repository of shared artefacts, especially versioned artefacts, will make it possible to reproduce studies and harmonize between related studies in a controlled manner. In general, this will help to improve future quality assurance and quality control practices in the research community.

There is currently no existing component that can be re-used as-is. However, multiple options exist. Among these:

1. Host a central instance of ATLAS to host and distribute shared artefacts and examples
2. Extend ATHENA to allow publishing (including from ATLAS), governance and distribution of shared artefacts

Due to the aggressive timelines, the initial implementation of the shared artefacts repository will be implemented through a central ATLAS instance - option 1, with the implementation details to be clarified in the integration design phase of EHDEN.

	D4.1 - Technical framework design and architecture		
	WP4 - Technical implementations		Version: v1.3 – Final
	Author(s): Antje Hottgenroth, Gregory Klebanov, Kees Van Bochoven, Nigel Hughes, Maxim Moinat, Michel Van Speybroeck, Sebastiaan van Sandijk, Peter Rijnbeek		Security: PU

5. EXTRACTION TRANSFORM LOAD (ETL)

Functional Requirements

EHDEN will focus on the development of efficient procedures and tools to perform high-quality translation of the local database structure and terminology systems. The EHDEN project is planning to onboard many data sources in the EHDEN network. This implies that for each of these data sources custom Extract Transform and Load (ETL) software must be developed that: extracts the data from the source database; transforms this to the right table structure and Standardized Vocabularies; and loads the data in the OMOP-CDM. The development of ETL software requires a complex set of skills and expertise and will be challenging to build for the data source and even trained SMEs. The ETL process needs high involvement of the data custodian who knows the data and its provenance best, deep expertise in OMOP-CDM and its Standardized Vocabularies, programming and engineering skills, and advanced expertise in processing and hosting Big Data with various database management systems, etc. We need to avoid that the ETLs projects become very large effort, in part because the ETL expertise still needs to be built. This would have high impact on the scalability of the data source transformations in EHDEN.

ETL software that converts data to the CDM also needs maintenance after this initial mapping for several reasons:

- **The source data can change.** If for example a new data domain is added in the source database these need to be included in the ETL. This does not happen frequently in most data sources.
- **The CDM structure is updated based on requirements from the community.** These changes have normally little impact on the CDM because backwards compatibility is in principle enforced. If there is impact these changes are normally very small. These changes are well documented on the CDM wiki page (<https://github.com/OHDSI/CommonDataModel/wiki>) and are supported by the OHDSI CDM Working Group.
- **The Standardized Vocabularies change over time.** This change will happen frequently simply because the source vocabularies themselves change often. Tools are being developed in OHDSI to support the comparison of vocabulary versions. The impact of these changes is relatively small and not much different from the impact of a local terminology change on the source database.


The maintenance procedure that has proven to work best is to build in several ‘sprints’ of the database based on the refresh rate of the source data. For example, if the data source gets updated twice a year then in between the maintenance is scheduled and comparisons are run between the previous version and the new version using the latest data cut.

In EHDEN we will automate the implementation of the ETL as much as possible by developing advanced ETL software that can easily be adapted for a specific local data source. This includes simplifying the CDM refresh procedure and its quality control.

The users of these Harmonization tools are the data custodians that would like to have their data mapped and the SMEs which support the process. Furthermore, the results of the mapping process should be transparent to potential future users of the data.

We have therefore defined the following user requirements:

1. Functional transparency is very important for the acceptance of these tools. This should create trust for all involved in the ETL Process that the tools will not impact the source data and will not share any patient data.

	D4.1 - Technical framework design and architecture		
	WP4 - Technical implementations		Version: v1.3 – Final
	Author(s): Antje Hottgenroth, Gregory Klebanov, Kees Van Bochoven, Nigel Hughes, Maxim Moinat, Michel Van Speybroeck, Sebastiaan van Sandijk, Peter Rijnbeek		Security: PU

2. The tools should enable a high-quality and fully reproducible ETL process. This implies that the tools should help in creating proper documentation and should produce analytical output to assess the quality to improve syntactic and semantic interoperability.
3. Users will require training and tool documentation.

Implementation

In the ETL process we will use several OHDSI tools that are described in the next sections.

5.1 ETL Design

White Rabbit (WR) is a data profiling tool used to characterize the source data. It aggregates the source data per table and field to produce a comprehensive report of data types, completeness and value frequency. This ‘Scan Report’ is exported in the form of an excel sheet, which makes the contents inspectable to anyone and easy to share. This scan report can be used stand-alone to inform the SME about the contents of the data source, but also loaded in Rabbit in a Hat (RiaH).

RiaH provides a graphical user interface to do an initial documentation of the mapping rules from source data to the OMOP CDM. The output is a Word document that forms the basis of the ‘Mapping Document’ in which all transformations are transparently recorded. This document contains the initial set of mapping rules that are subsequently refined before mapping implementation. The final ETL Specifications are documented in the document based on the standard .

Another powerful feature of WR is to generate fake data. This is useful in instances where the SME cannot access the real data, which is often the case with sensitive medical records. The development team can run the transformation scripts on the fake data to test the correctness of the procedure. Validation has still to be done against the real data.

RiaH also provides support for the transformation development by creating a testing framework based on the scan report. The framework helps the SME write end-to-end tests for the transformations.


Both WR and RiaH are existing, open-source, stand-alone java applications from the OHDSI community. Both can be run at the data source site. In EHDEN, we will continue to use and enhance these OHDSI tools. Depending on the needs, WR might be expanded to connect to other types of data sources such as SAS data files. Rabbit in a Hat can possibly be extended as well in particular to ingest a data dictionary, add support for OMOP concepts and a method for generating executable code from the defined mappings.

5.2 Vocabulary Mapping

Similar to WR and RiaH, Usagi is an existing, open-source, stand-alone java application. It is used during the transformation process to map source terms to the standard OMOP vocabulary. The core functionality is rank standard concepts based on string similarity with the source term. The graphical user interface allows for manual review of the automatically suggested mappings.

It is expected that Usagi will continue to be used. Minor functionalities might be added, for instance to aid drug mapping.

Usagi might be replaced by a new tool that also supports vocabulary maintenance (see section 6.8). This tool might accommodate use cases like central review of vocabulary mappings by domain experts, inclusion of the mapping quality, addition of new concepts and sharing of vocabulary mappings within and outside the OHDSI community.

	D4.1 - Technical framework design and architecture		
	WP4 - Technical implementations	Version: v1.3 – Final	
	Author(s): Antje Hottgenroth, Gregory Klebanov, Kees Van Bochoven, Nigel Hughes, Maxim Moinat, Michel Van Speybroeck, Sebastiaan van Sandijk, Peter Rijnbeek	Security: PU	23/33

5.3 OMOP Standardized Vocabulary Explorer and Distribution

During the project, new vocabularies will be added, and existing vocabularies will be extended. To allow reuse of these concepts, a central vocabulary management system (+governance) needs to be in place with transparent procedures for updating and releasing new versions aligned and part of the OHDSI processes.

The OHDSI ATHENA platform will be used to distribute EHDEN vocabularies. It is possible that it will need to be extended to allow additional functionality, including:

- Public vocabulary bundling and versioning
- Location and organization specific categories
- Governance and stewardship processes
- Distribution of other shared artefacts

5.4 CDM Validation Tools

The validation tools are in support of the validation framework. It will be one or more tools that are used specifically to assess the accuracy, completeness and repeatability of the mapping of the data source to the OMOP CDM.


The current validation framework in OHDSI (Achilles) consists of two parts; Achilles Heel for OMOP CDM conformation checks and data characterization. The latter can be used to roughly assess the amount of data loss. However, additional tooling might be needed to quantify the data loss. For this we can use previous individual efforts to:

- Source data profiling, including anomalies and data quality issues
- Compare source code prevalence in source data versus in OMOP CDM data
- Report the percentage of source terms mapped to the OMOP vocabulary
- Report the percentage of records mapped in OMOP CDM dataset

These methods need to be further developed and integrated in the existing data validation tooling (Achilles).

The tool Tantalus is used for vocabulary validation. This enables comparisons between vocabulary releases in order to ensure the quality of a new vocabulary release.

In addition, there is an OHDSI-wide Data Quality (DQ) Initiative in which multiple EHDEN members are involved. This initiative focuses on creating a comprehensive set of rules, tools and best practices across the real-world data-to-evidence life cycle. Multiple EHDEN members are leading the 'ETL software to support data quality review' workstream. Outputs from the other workstreams will also be, where applicable, used within EHDEN.

	D4.1 - Technical framework design and architecture		
	WP4 - Technical implementations	Version: v1.3 – Final	
	Author(s): Antje Hottgenroth, Gregory Klebanov, Kees Van Bochoven, Nigel Hughes, Maxim Moinat, Michel Van Speybroeck, Sebastiaan van Sandijk, Peter Rijnbeek	Security: PU	24/33

6. TRAINING AND EDUCATION

Education and training is a core ambition of EHDEN: only through dissemination and training can we assure that data models, tools and methods will be widely adopted. Training for all stakeholders is one of the critical steps to raise the high quality of the service components, as well as to communicate EHDEN's vision and value propositions.

We recognize that the educational requirements may differ per key stakeholder, but believe the following elements to be essential:

1. Working with OMOP CDM and OHDSI tools within an open science, open source framework (from an ETL and data source perspective, specifically for data sources and SMEs)
2. Performing analysis and research within a federated network, and utilising OHDSI tools and processes
3. Methodological and RWD/RWE study creation, implementation and evaluation
4. Working with Health Outcome Standards in both research and clinical settings
5. Supporting the use of RWE in clinical, regulatory, HTA and reimbursement domains

EHDEN is developing a curriculum to cover the key components above, in a number of tiers dependent on user need. EHDEN will coordinate the development of the educational requirements and training materials in close collaboration with OHDSI. We will re-use educational material from OHDSI and tailor this to the European environment.

In the sections below we describe all the components currently being worked on.

6.1 EHDEN Academy


EHDEN is developing an e-learning environment for the training of all its stakeholders (<https://academy.ehdn.eu/>). We will use Moodle (www.moodle.org) as the framework. All WPs in EHDEN will contribute to the training material development. Initially the EHDEN Academy will be for members and SMEs, but we will eventually open it to all stakeholders, including the full OHDSI community.

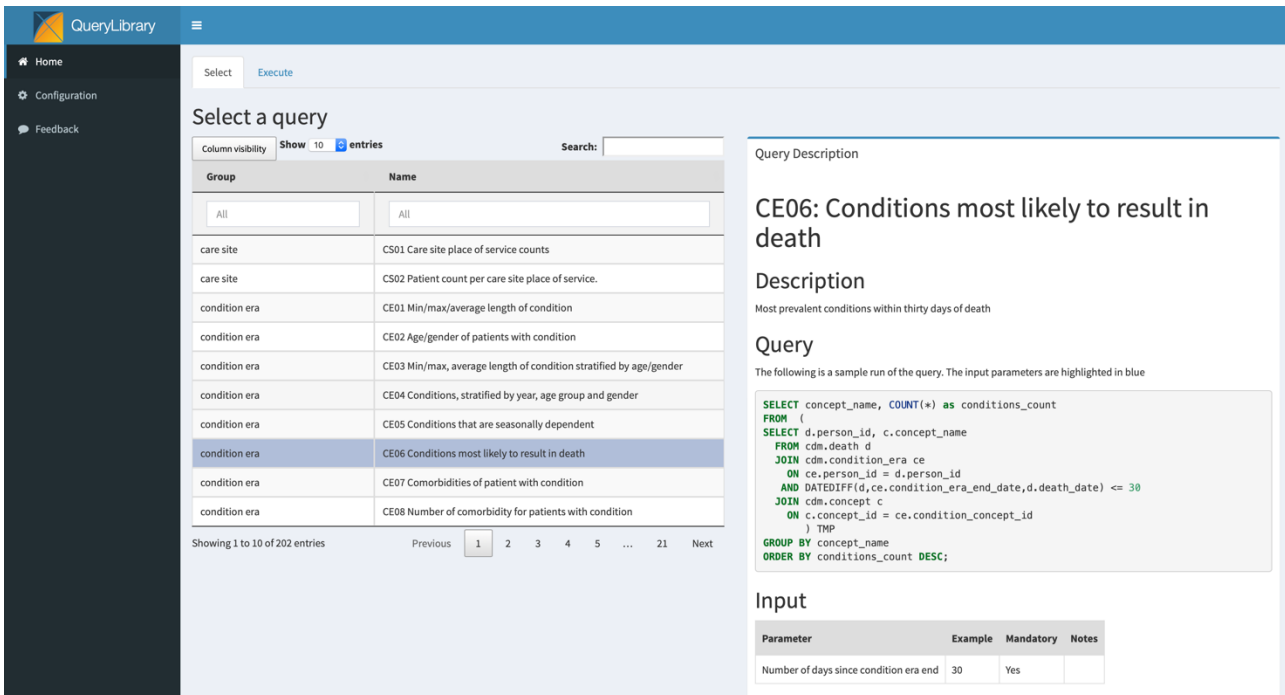
6.2 Virtual training environment

OHDSI and EHDEN are actively collaborating with Amazon to further develop the OHDSI-IN-A-BOX solution. This virtual machine can be easily instantiated by trainees and contains many OHDSI tool, a CDM database with simulated data and a source database that can be used in the EHDEN Academy to train on mapping. This VM should contain all the tools that a Data Partner would install locally, including Arachne Node, Atlas, All the ETL tools etc. For more information about this tool we refer to www.github.com/OHDSI/OHDSI-IN-A-BOX. Besides providing a powerful training environment, this tool will also be useful for EHDEN partners that do not have a CDM database and want to be more involved. A course is developed in the EHDEN Academy to train users on installing the OHDSI-IN-A-BOX environment.

6.3 Query Library

For training purposes, we are developing a library of frequently-used queries against the OMOP-CDM (<http://www.github.com/EHDEN/QueryLibrary>). The approved queries will improve quality of study execution and lower the barrier for new OMOP-CDM users.

	D4.1 - Technical framework design and architecture		
	WP4 - Technical implementations		Version: v1.3 – Final
	Author(s): Antje Hottgenroth, Gregory Klebanov, Kees Van Bochoven, Nigel Hughes, Maxim Moinat, Michel Van Speybroeck, Sebastiaan van Sandijk, Peter Rijnbeek		Security: PU



The screenshot shows the QueryLibrary web interface. On the left is a navigation menu with 'Home', 'Configuration', and 'Feedback'. The main area is titled 'Select a query' and contains a search bar and a table of queries. The table has columns for 'Group' and 'Name'. The query 'CE06 Conditions most likely to result in death' is selected and highlighted. Below the table, it shows 'Showing 1 to 10 of 202 entries' and pagination controls. On the right, the 'Query Description' panel for CE06 is visible, showing the query title, a description, the SQL query code, and an 'Input' table with parameters.

Parameter	Example	Mandatory	Notes
Number of days since condition era end	30	Yes	

Figure 4: Example screenshot of the QueryLibrary tool

The tool automatically renders the queries to the SQL dialect, CDM and vocabulary schemas specified by the user. It can execute the query against a user's database and show the results of the query. EHDEN has shared the QueryLibrary tool with OHDSI and will maintain this tool for the community. This includes the addition of queries and quality control.

6.4 EHDEN Environments


The EHDEN initiative will stand up multiple isolated physical tiers to be used during the implementation lifecycle to support various development, QA/QC, dry-run and production activities:

1. Development
2. Test / QC
3. Production
4. Sandbox

Each tier will only allow system to system integration and information flow with other systems located in the same tier but never across tiers. The infrastructure capacity and sample data in each tier will reflect the activities required to be performed.

The core hosting infrastructure is hosted on the AWS platform and co-located in the same region and data centre, with an exception of the ELIXR A&A system. The following are the key AWS components used by EHDEN:

- EC2 - ARACHNE, ATLAS and R-Server hosting
- RDS (Aurora / PostgreS) - SynPUF data hosting
- Athena – SQL Workbench for accessing and executing SQL queries against sample SynPUF
- S3 – raw SynPUF sample data storage

	D4.1 - Technical framework design and architecture		
	WP4 - Technical implementations		Version: v1.3 – Final
	Author(s): Antje Hottgenroth, Gregory Klebanov, Kees Van Bochoven, Nigel Hughes, Maxim Moinat, Michel Van Speybroeck, Sebastiaan van Sandijk, Peter Rijnbeek		Security: PU

1. (Build) Integration

The Development tier will support day-to-day code development and build activities:

- Latest development build (main code branch) – Software, Data Model and Database
- Automated Builds and Continuous Integration (CI)
- Unit Testing
- Regression Testing
- Integration Testing
- Backup and archiving

2. Test/QC

The Test/QC tier host of the Newest Release build (main code branch), including both Software, Data Model and Database. It will be built with Quality Control process requirements in mind, including:

- Non-functional Testing
 - a. Performance Testing
 - b. Load and Stress Testing
 - c. System Integration Testing
- Functional Testing
- User Acceptance Testing
- Backup and archiving


3. Production

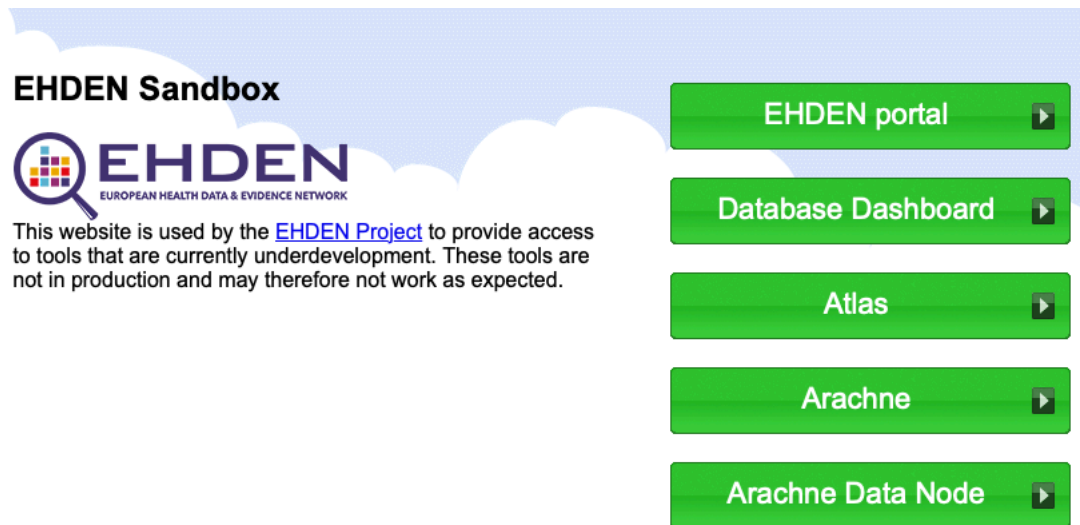
The production tier host of the latest tested production release build (production support branch), including both Software, Data Model and Database. It will be built with High Availability and Scalability requirements in mind, including:

- Redundancy and failover
- DR strategy and process
- Backup and archiving
- Availability Monitoring
- Smoke Testing

4. Sandbox

The Sandbox (i.e. a testing environment) tier will be used for performing dry-run test, demo or training studies on the latest production release without any risk of affecting data in a real production environment.

	D4.1 - Technical framework design and architecture		
	WP4 - Technical implementations	Version: v1.3 – Final	
	Author(s): Antje Hottgenroth, Gregory Klebanov, Kees Van Bochoven, Nigel Hughes, Maxim Moinat, Michel Van Speybroeck, Sebastiaan van Sandijk, Peter Rijnbeek	Security: PU	27/33



Questions about these applications from EHDEN consortium members can be posted on the [EHDEN Forum](#)

Figure 5. The EHDEN Sandbox (<https://test.ehden.eu>)

EHDEN will also host a publicly accessible OMOP CDM instance containing simulated data that can be used by EHDEN members for training and testing. This data set can contain the SynPUF and the Synthea data.

7. INFORMATION ARCHITECTURE

7.1 Information Models

The information model can serve multiple needs, including:


- 1) A definition of the master system and a golden truth of record for data elements.
- 2) A standardized approach to storing data (OMOP CDM), including detailed definitions of key attributes and constraints

The core, well defined information model used by EHDEN platform is the OMOP CDM.

The critical information elements used during the business process are the Study Information and business process as defined by ARACHNE and ATLAS, as well as Data Source as defined by EMIF Database Catalogue.

7.1.1 OMOP CDM

The latest version of OMOP CDM is version 6. However, in this initial implementation EHDEN will target OMOP CDM 5.3.1 since none of the current OHDSI tools provide support for version 6.0 yet.

	D4.1 - Technical framework design and architecture	
	WP4 - Technical implementations	Version: v1.3 – Final
	Author(s): Antje Hottgenroth, Gregory Klebanov, Kees Van Bochoven, Nigel Hughes, Maxim Moinat, Michel Van Speybroeck, Sebastiaan van Sandijk, Peter Rijnbeek	Security: PU

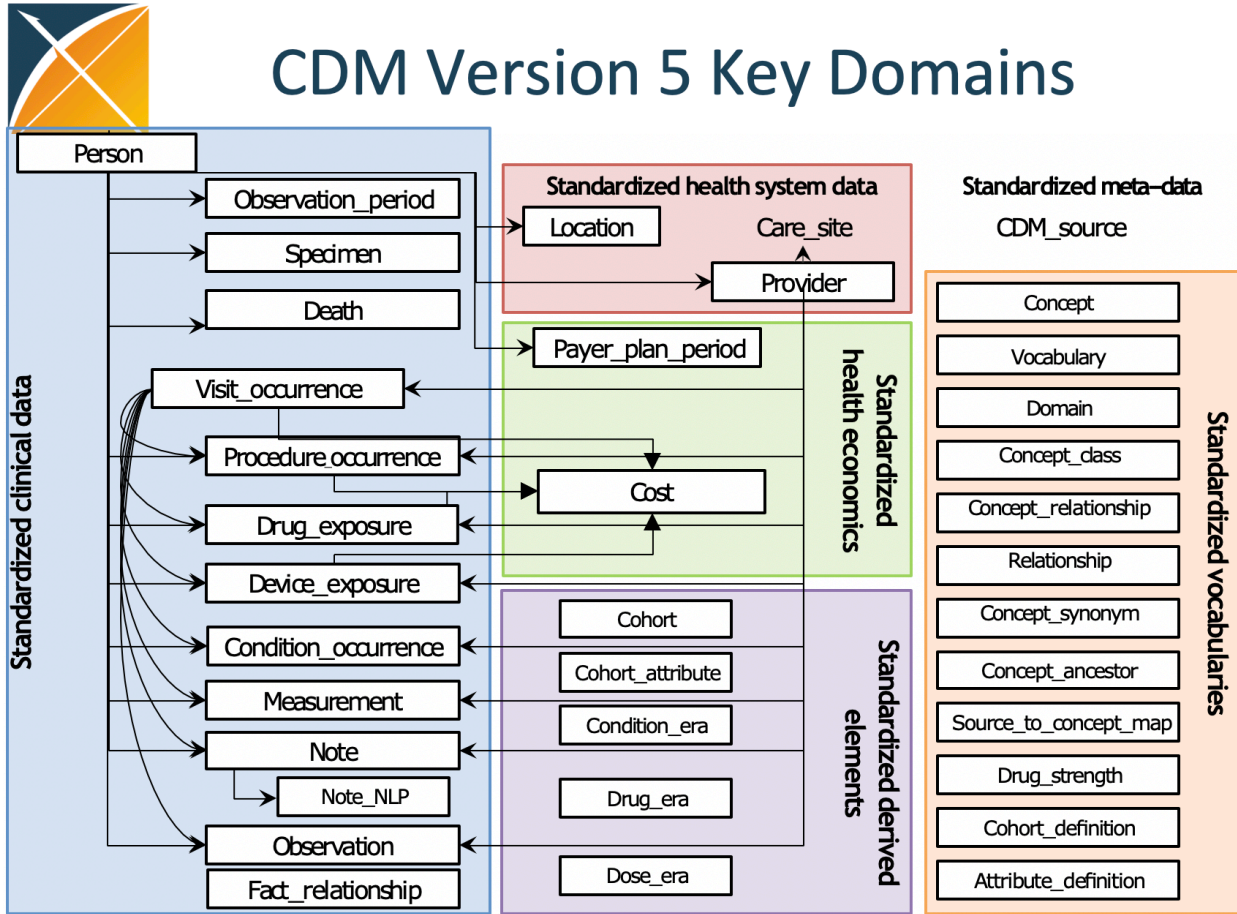



Figure 6: OMOP CDM model

The OMOP CDM model is used to store patient medical data in one standardized way across all participating data provider organizations. This will ensure that the business question defined in one organization can be executed across the whole network.

7.1.2 Study Information and Execution (ARACHNE)

The critical data elements as defined in ARACHNE are Study, Data Source, Study Participant, and Analysis including aggregated elements such as Analysis Design, Code and Results.

	D4.1 - Technical framework design and architecture		
	WP4 - Technical implementations		Version: v1.3 – Final
	Author(s): Antje Hottgenroth, Gregory Klebanov, Kees Van Bochoven, Nigel Hughes, Maxim Moinat, Michel Van Speybroeck, Sebastiaan van Sandijk, Peter Rijnbeek		Security: PU

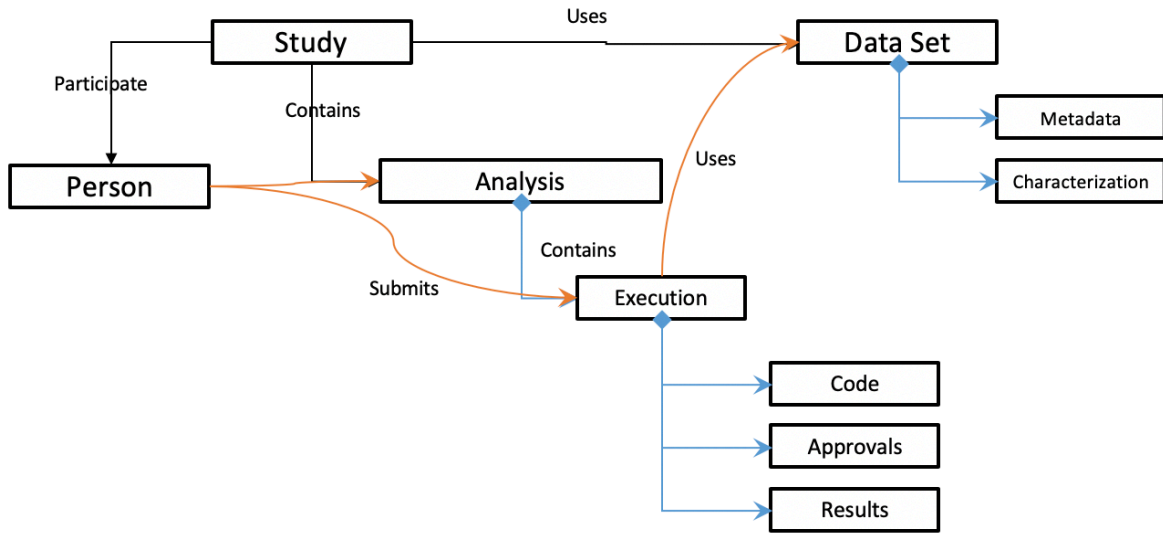



Figure 7: ARACHNE data elements (model)

The study analysis design and code will be defined and owned by ATLAS, while the core Data Set information will be defined, owned and provided by the Database Catalogue.

7.1.3 Analysis Design (ATLAS)

The core elements owned and defined by ATLAS are the Concept Set, Cohort, Analysis as well as specific descriptive, predictive and estimation analysis designs such as Characterization, Incidence Rates, TxPathways, Patient Level Prediction and Population Level Effect Estimation.

	D4.1 - Technical framework design and architecture		
	WP4 - Technical implementations		Version: v1.3 – Final
	Author(s): Antje Hottgenroth, Gregory Klebanov, Kees Van Bochoven, Nigel Hughes, Maxim Moinat, Michel Van Speybroeck, Sebastiaan van Sandijk, Peter Rijnbeek		Security: PU

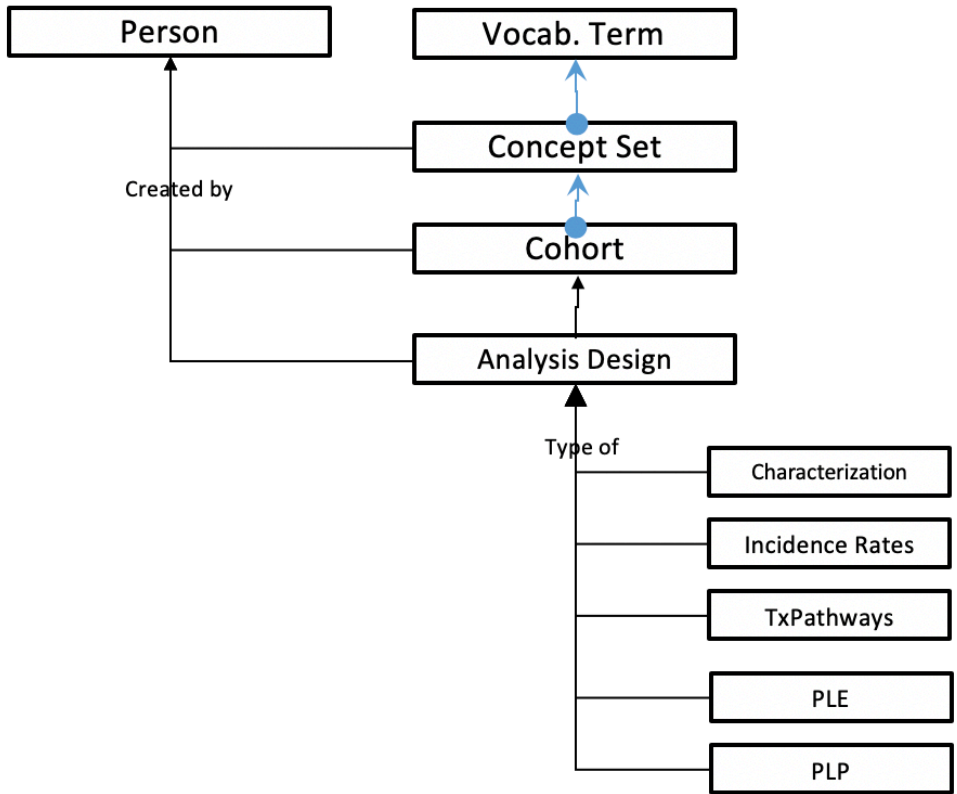



Figure 8: ATLAS/WebAPI data elements (model)

7.2 Information Flow

Multiple systems will be integrated to exchange relevant information, including ARACHNE, ATLAS and EHDEN Database Catalogue. The ARACHNE platform is already integrated with ATLAS, with integration of ARACHNE and the EHDEN Database Catalogue still to be implemented.

	D4.1 - Technical framework design and architecture		
	WP4 - Technical implementations		Version: v1.3 – Final
	Author(s): Antje Hottgenroth, Gregory Klebanov, Kees Van Bochoven, Nigel Hughes, Maxim Moinat, Michel Van Speybroeck, Sebastiaan van Sandijk, Peter Rijnbeek		Security: PU

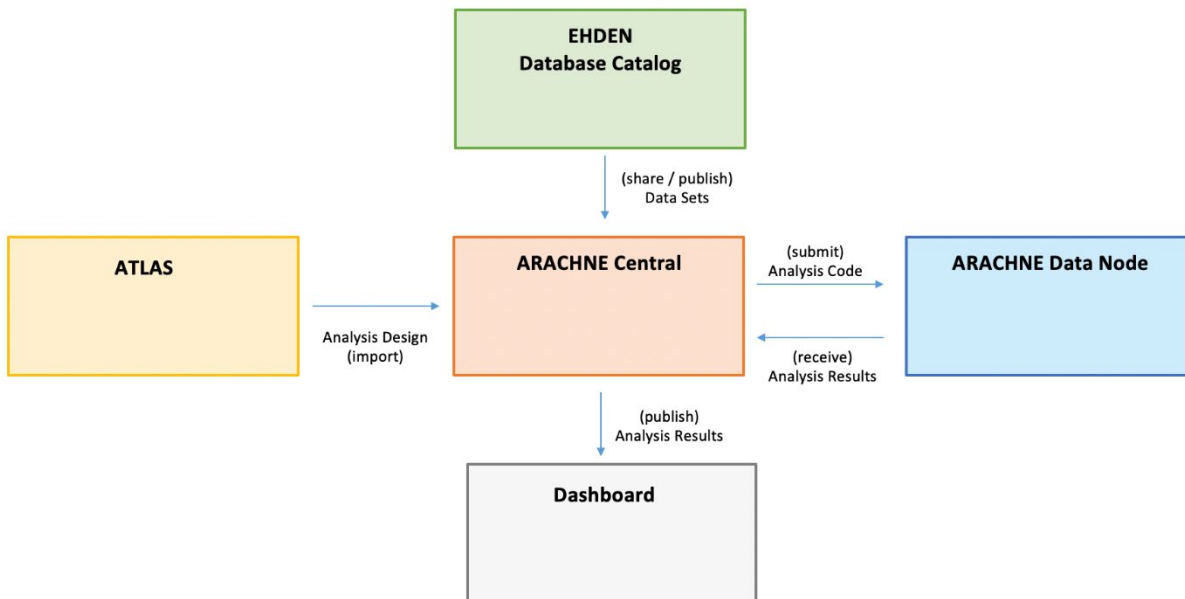


Figure 9: EHDEN platform high level data flow

There will be separate design documents outlining detailed integration technical architecture design for each of the integration points highlighted above:

- Database Catalogue – ARACHNE (new implementation)
- ARACHNE Central – Dashboard (new implementation)
- ARACHNE - ATLAS (implementation exist)
- ARACHNE Central – Data Node (implementation exists)

8. INTEGRATION ARCHITECTURE


The existing Application APIs (e.g. designed to enable a specific user interface) are not well suited for integration needs and thus should not be exposed to enable integration with clients outside of the specific application. Instead, each of the EHDEN core components participating in data and transaction exchanges will define, implement and expose a well-defined Integration API (REST based) to be used to implement the information flow as outlined above. Each of the end points will be secured.

Since all planned EHDEN integrations are point—to—point (e.g. one client to one consumer vs. multiple consumers or multiple clients), EHDEN will not deploy any specialized middleware - such as Enterprise Service Bus (ESB) - to enable communication between various components. Instead, each application component will host the corresponding API end point. Each API function implementation will implement and guarantee ACID transactions and guarantee message delivery.

Each integration point will be documented in a separate technical implementation document. All EHDEN Integration API functions will follow a consistent documented implementation best practices e.g. signature, naming conventions, security etc.

The standard message exchange will be based on JSON.

In case in future project phases it will be decided to deploy various re-usable API-based services, it is recommended to consider deploying a middleware that will provide further service de-coupling and routing, such as the one from WSO (<https://wso2.com/products/enterprise-service-bus/>).

	D4.1 - Technical framework design and architecture		
	WP4 - Technical implementations		Version: v1.3 – Final
	Author(s): Antje Hottgenroth, Gregory Klebanov, Kees Van Bochoven, Nigel Hughes, Maxim Moinat, Michel Van Speybroeck, Sebastiaan van Sandijk, Peter Rijnbeek		Security: PU

9. SECURITY

In EHDEN project we are currently working on an information security policy to state requirements for different security aspects of the whole architecture. The first deliverable is expected by the end of 2019. Therefore, here we describe only principles for authentication and authorization (A&A) of users in EHDEN ecosystem.

Authentication is the process of verifying who you are by asking a trusted resource. Authorization is the process of verifying that you have access to something.

Identity Propagation (IP) is a process of sharing the verified identity of a user with all components involved in a user transaction, e.g. web page(s), API(s), and database(s), where each layer will re-validate the identity and authorize to access resources at that level. Identity Propagation – while the ideal state – is often difficult to achieve in federated environments where each tier and component must support the same IP mechanism.

Considering the vast number of tools and technologies involved, EHDEN ecosystem will not be able to achieve a full identity propagation state. Instead, each tool/component will implement a consistent A&A framework to ensure that the user is at minimum authenticated and authorized to access the component and its features.

The Identity Management Provider (IMP) should support creation of group and assigning group memberships to identities. Then when authenticated, these groups are passed to the Service Provider (SP) (app, web service, etc.) and they can be mapped to internal app roles (as a part of RBAC) to allow access to various app resources.

The federated identity management system is linking and using identities of a user across several identity management systems and is being used to authenticate and authorize those users across participating systems.


EHDEN will use the ELIXIR Authentication and Authorisation Infrastructure (ELIXIR AAI) that will integrate with external Service Providers and will be used to manage EHDEN users and their assignments to high level roles.

ELIXIR is an intergovernmental organisation that brings together life science resources from across Europe – databases, software tools, cloud storage, etc. The goal of ELIXIR is to coordinate these resources so that they form a single infrastructure. Elixir AAI is one of its core components, allowing secure authentication and authorization layer for the whole ELIXIR infrastructure. Elixir AAI is a proxy service, which means that it relies on user authentication by external authentication providers (universities, Google, ORCID). Therefore, users do not need to create another user account and they can use their local institutional login credentials. Users are assigned to user groups of different services via open source Perun system either by SP administrators or automatically. The assignments are pushed to LDAP (OpenLDAP implementation), which is then used by the authentication system to get authorization information for the users. Elixir AAI is distributed geographically, it is monitored 24/7 and usually ready to solve the issues immediately. Elixir AAI is GDPR compliant and there are public documents how to make SPs capable of using Elixir AAI for A&A.

Currently, many EHDEN components – including ARACHNE, ATLAS (partial, with Open ID implemented) and Database Catalogue – do not provide complete support for the underlying protocols to be used by EHDEN (OpenID/SAML) and the implementation for those protocols must be completed.

NEXT STEPS

This blueprint document gives a high-level overview of the EHDEN platform and provides for each required functionality an overview of the available tools from other projects, especially OHDSI. The next step is to

	D4.1 - Technical framework design and architecture		
	WP4 - Technical implementations		Version: v1.3 – Final
	Author(s): Antje Hottgenroth, Gregory Klebanov, Kees Van Bochoven, Nigel Hughes, Maxim Moinat, Michel Van Speybroeck, Sebastiaan van Sandijk, Peter Rijnbeek		Security: PU

further develop the roadmaps for each of the individual tool. This will need the involvement of multiple stakeholders but especially the users of the tools. We need to establish and implementation timelines. The following are the critical artefacts:

- A prioritised list of requirements (backlog)
- Project schedule (Implementation timelines and critical milestones), including the roadmap for each technical component
- A&A solution technical implementation
- EHDEN Hosting blueprint
- Integration Technical Design (for each integration)

It will be crucial to perform these steps in close collaboration with the OHDSI developers since we need to make sure we get global uptake of our efforts and we need to guarantee interoperability. Fortunately, we are in a very good positions to make this a reality because multiple developers in EHDEN have leading roles in the OHDSI developers' teams.

Furthermore, considerable effort will be put on the integration of the tools into one common architecture (EHDEN Portal) and common security framework. This work will require the communication framework as described above that defines how information is shared between components, e.g. how the information from the data base catalogue is used in the Arachne tool.

Finally, we will focus on the results dissemination track in which a framework will be developed that contains dashboards that show results from multi-database studies in the data network.

CONCLUSION

This first deliverable of WP4 shows that the EHDEN technical framework development can strongly benefit from the available tools in OHDSI and those developed in other projects such as EMIF. EHDEN has a big opportunity to bring all these valuable tools together and further improve them based on user feedback and the available expertise in the consortium. We look forward to this exciting journey.