DataTools4Heart

A European Health Data Toolbox for Enhancing Cardiology Data Interoperability, Reusability and Privacy

Milestone MS6

1st prototype of DT4H platform and user interface

Reference	MS6_ DataTools4Heart_UB_30092024
Lead Beneficiary	UB
Author(s)	Josep Lluís Gelpí, Laia Codó
Dissemination level	Public
Type	Demonstration (available at software repositories)
Official Delivery Date	30/09/2024
Date of validation of the WP leader	30/09/2024
Date of validation by the Project Coordinator	30/09/2024
Project Coordinator Signature	

DataTools4Heart is funded by the European Union's Horizon Europe Framework Under Grant Agreement No. 101057849.

Version Log

Issue Date	Version	Involved	Comments
23/09/2024	0.1	Josep Lluís Gelpí, Laia Codó	First draft
26/09/2024	0.2	Josep Lluís Gelpí, Laia Codó	Second draft with feedback incorporation
30/09/2024	Final	Josep Lluís Gelpí, Laia Codó, Xènia Puig	Revised and corrected final version

Executive Summary

This document serves as the justification for achieving Milestone 6, which marks the release of the first integrated prototype of the platform's technical infrastructure. The primary objective is to deliver an operational Minimum Viable Product (MVP) of the DataTools4Heart platform. The prototype is designed to support the core research cycle of researchers, enabling them to design, manage and analyse the results of distributed and federated learning (FL) experiments across the DataTools4Heart federation. The current implementation is publicly available via the project's central code repository at https://github.com/dataTools4Heart. Ongoing enhancements and refinements will continue across individual platform modules, alongside integration efforts, culminating in the final technical release of the platform by Month 36 (M36).

Table of Contents

Acronyms

Federated Execution Manager (FEM)

Secure Multi-Party Computation (SMPC)

CVD: Cardiovascular disease

CM: Consortium Meeting

MVP: Minimum Viable Product

FL: federated learning

FDN: Federated Data Node

RN: Reference Node

VA: Virtual Assistant

LLM: Large Language Models

NER: Named Entity Recognition

NEL: Named Entity Linking

FHIR: Fast Healthcare Interoperability Resources

ETL: Extract, Transform and Load

API: Application Programmatic Interface

MIP: Medical Informatics Platform

HBP: Human Brain Project

ML: machine learning

List of figures

1 Introduction

One of the project's objectives is offering an integrative computational infrastructure enabling the deployment of the DataTools4Heart data toolbox across the DT4H data federation. The toolbox components integrate tools for data standardisation, machine translation, federated learning analytics and data synthesis for cardiovascular diseases (CVDs) data-driven research. The platform is built on top of them to provide the necessary compute services, interfaces and interoperability layers that enable the entire data lifecycle under a private-by-design federated design.

1.1 Platform capabilities

The DT4H platform will be composed of a series of interoperable software modules, packed in software containers, including:

- Central orchestration service enabling secure and efficient communication channels among the distributed infrastructures, *i.e.,* the federated nodes.
- Management interface for controlling the distributed analysis processes across the federated nodes
- Data management interface providing data control and transference capabilities
- Virtual shared storage interface to generate a virtual storage platform shared among platform nodes
- Graphical user interfaces (GUIs) for data to provide a GUI for accessing data and tools in a virtual workspace
- Blockchain modules for building/managing a permissioned blockchain network to handle
- authentication/authorization, and process auditing
- Metadata catalogue for DT4H data sets to make these data sets 'Findable' in FAIR terms

2 Methodology

The development of the platform components follows an iterative and agile methodology, centred around a Minimum Viable Product (MVP) approach. This process is organised into focused, incremental development cycles, each resulting in MVP versions of the platform components. These MVP releases enable early validation, testing, and adoption of prototypes within the consortium. The progress and functionality of each prototype is showcased through live demonstrations at the Consortium Meetings (CMs) held every six months, fostering continuous feedback and collaboration across the consortium.

The initial development phase (M6) was devoted to compiling the list of functional and technical requirements (T2.1) helping to describe the essential functionalities of the platform.

Along the following months (M6-M18), iterative sprints lead to a prioritised implementation of the core features of the individual platform modules guided for the basic interoperability capabilities defined in the first draft of the overall integrated architecture (see below section Architecture).

This methodology ensures that platform components are incrementally improved, maintaining flexibility and adaptability to stakeholder needs, while staying on track for final delivery by M36.

3 Architecture

The DataTools4Heart platform is designed as a modular and interoperable system, consisting of independent yet complementary software components that work together to enable secure computing and data management across the federation. The platform operates within a hybrid hub-and-spoke topology, where a node indistinctly assumes a role based on the services it hosts:

Reference Nodes (RNs): These nodes host the transversal core services basic for the well-functioning of the overall platform, such as those responsible for the federation's coordination, the data discoverability services, etc.

Federated Data Nodes (FDNs): These nodes provide computation capabilities and usually store sensitive, private data, which remains accessible only *on-premises*. FDNs run local data processing and management services, like CogStack or onFHIR stack. But they also interact with RNs to perform federated tasks like federated learning (FL) experiments.

The diagram below summarises the interactions between the platform components. Broadly, components that directly interface with AI researchers and clinicians offer discovery and federated analysis services. These include the portal, the data catalogue, the Virtual Assistant, and the AI dashboard, all of them deployed in Reference Nodes. Middleware components, such as the Federated Execution Manager (FEM), the Permissioned Blockchain Lawyer, and the Secure Multi-Party Computation (SMPC) network, act as intermediaries, ensuring secure access to and processing of protected datasets within the Federated Data Nodes (FDNs).

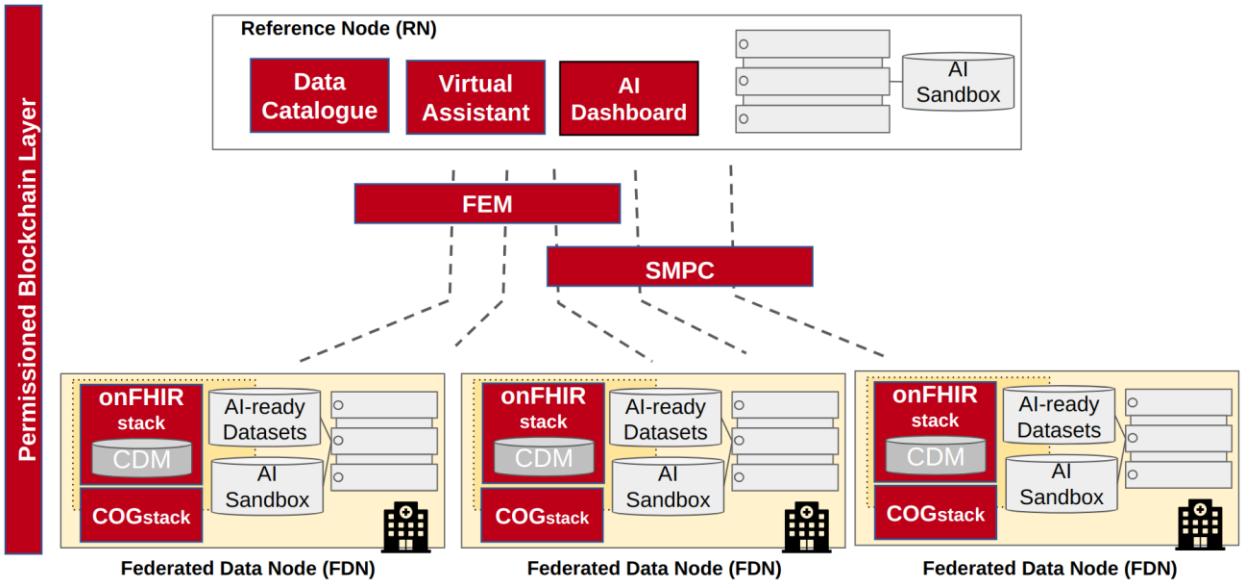


Figure 1. General diagram of platform component's interaction.

Single-node components deployed at the FDNs, such as the onFHIR stack and CogStack, are in charge of populating local patient repositories according to the DataTools4Heart Common Data Model (CDM), as well as the corresponding "AI-ready datasets". This volume serves as the primary data source for federated tasks being processed at the FDNs, while the "AI sandbox" volume provides a persistent workspace for them.

Together, these components ensure the secure and efficient deployment of the DataTools4Heart toolbox (seeMS7_ATH_30Sep24.docx) across the participant FDNs.

4 Components

This section offers a concise overview of the components that make up the first prototype of the DT4H platform. A summary of these components is provided in the table below:

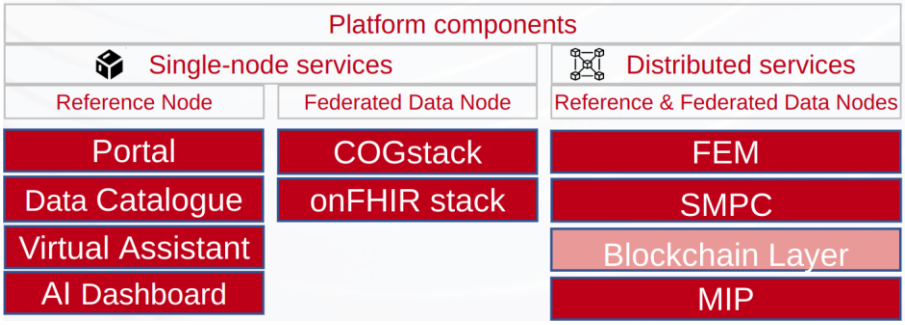
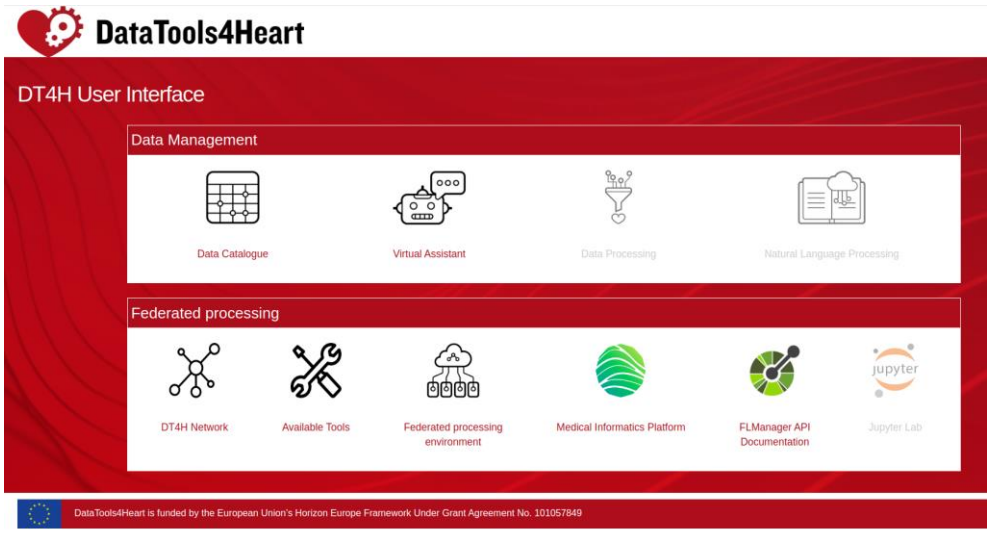


Figure 2. Platform's components summary. Note: The Permissioned Blockchain Layer is not part of the MPV. Currently no common authentication and authorization infrastructure is set.

The platform components can be categorised based on their deployment model. They fall into two main categories: **single-node services**, which operate independently on a single node (either a Reference Node or a Federated Data Node), and **distributed services**, which consist of multiple interdependent servers deployed across both Reference Nodes and Federated Data Nodes.

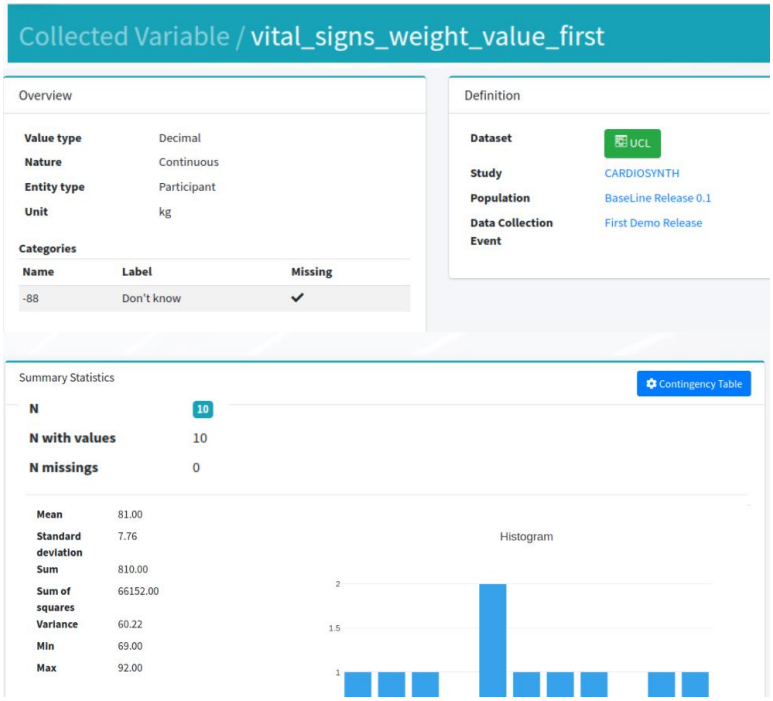
Below is a summary table for each platform component, including links to their corresponding source code repositories. With a few exceptions, most repositories are publicly accessible within the DataTools4Heart GitHub community. Users can easily filter these modules using the specific topic tag ("dt4h-platform-module") that annotates the relevant repositories across the GitHub domain. Repositories that are not yet open access will be made available at the same URL provided here as they reach a higher maturity level.

4.1 Portal	
Description	An open web-based interface providing users with integrated access to the full set of services and resources offered by the DT4H consortium. Typically installed centrally on a Reference Node, it serves as the primary access point for AI researchers and clinicians interacting with the platform. However, being designed as a fully portable system, it can also be deployed on a Federated Data Node (FDN), enabling local access to all internal DT4H services.

	
Task/WP	T6.1
Source code	Repository https://github.com/DataTools4Heart/portal Access public
Access	Test desplyment (BSC node): https://datatools4heart.bsc.es/
Documentation	

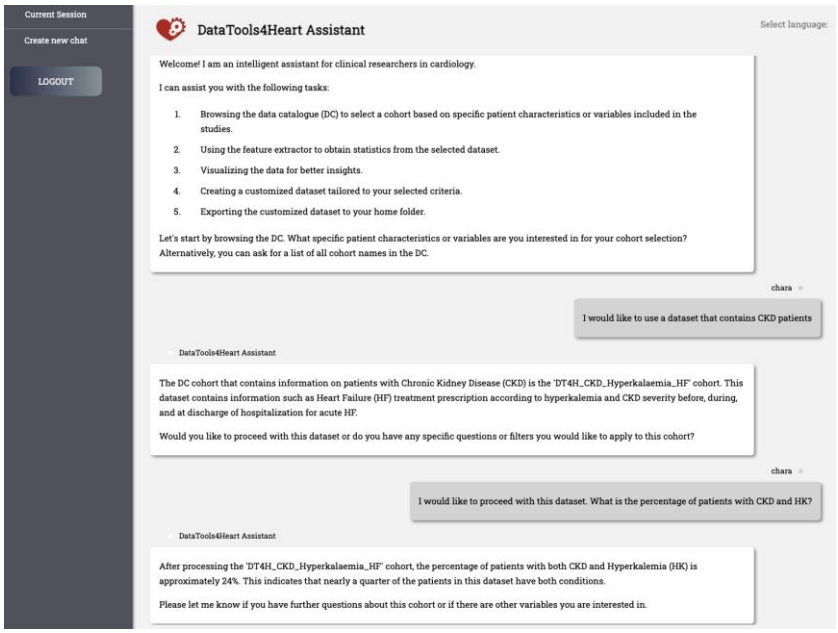
4.2 Cardiology Health Data Catalogue

Description	Centralised catalogue providing public metadata on the AI-ready cardiology datasets generated across the DT4H federation. It enables the discovery of DT4H use-case derived cohorts, allowing users to browse and filter by dataset features. While the catalogue does not reveal actual data values, it shares the structure of the variables and their availability across Federated Data Nodes (FDNs). Note: the current deployment is integrating only some exemplary datasets.

	
Task/WP	T2.5
Source code	Repository https://github.com/DataTools4Heart/obiba-data-catalogue Access public
Access	Test deployment (BSC node): https://catalogue.datatools4heart.bsc.es/
Documentation	Obiba official documentation: https://www.obiba.org/pages/infra/

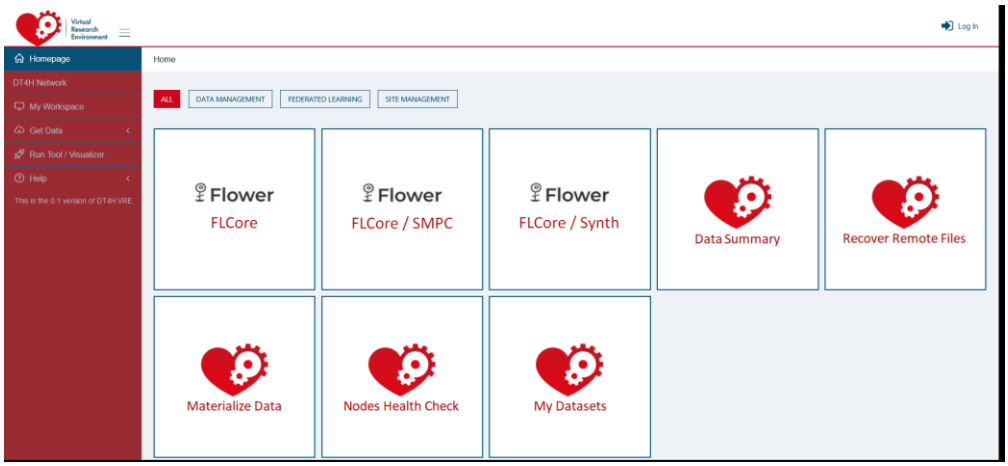
4.3 AI-powered Virtual Assistant

Description	The Virtual Assistant (VA) is a AI-powered chatbot to explore large, multi-source cardiology datasets through the onFHIR feature extraction module, which retrieves aggregated statistics of the features from FDNs. Utilising Large Language Models (LLMs) like GPT, the VA allows users to explore cohorts using natural language and evaluate their suitability for FL experiments.

	
Task/WP	T6.2
Source code	Repository Access https://github.com/DataTools4Heart/virtual_assistant private
Access	Test deployment (ATH node): https://chomsky.ilsp.gr:8643
Documentation	

4.4 AI Dashboard

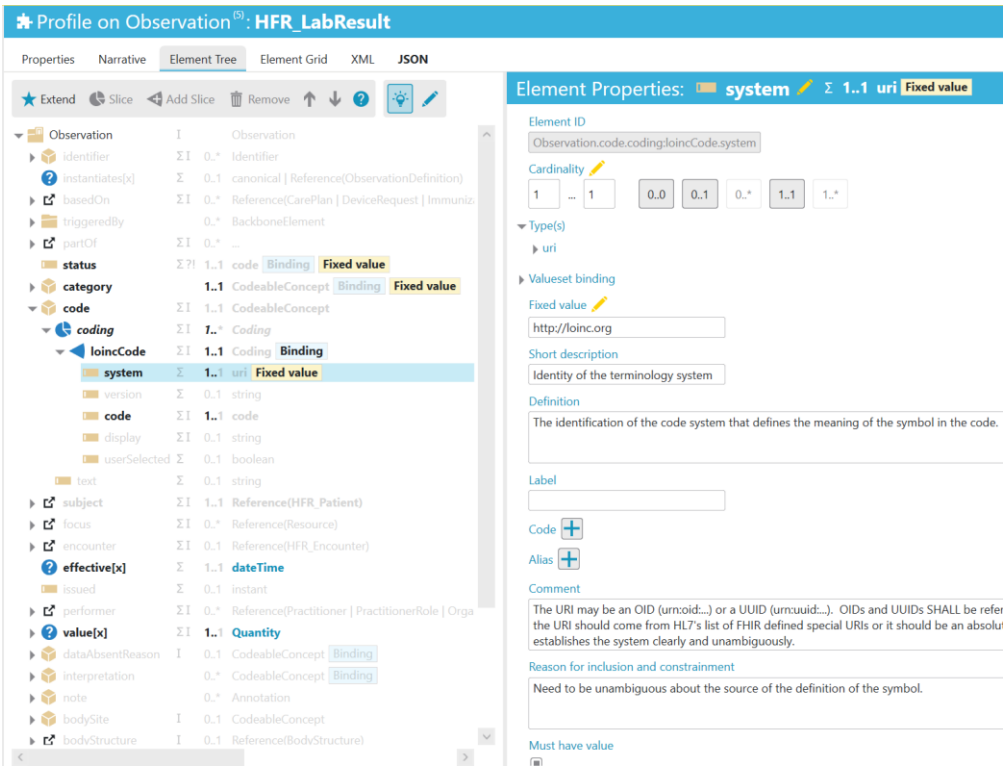
Description	The AI Dashboard is a web-based interface designed to assist AI researchers and clinicians in the design, management, and execution of FL experiments across the DT4H federation. It provides a user-friendly environment to interact with various DT4H toolbox components, primarily those implementing AI workflows, such as training and inference of AI models, or those allowing the preparation of AI-ready datasets using privacy-preserving means. The AI Dashboard is a crucial component that bridges the gap between AI researchers and the underlying programmatic access of the federation, *i.e.*, the FEM-orchestrator API.

	
Task/WP	T6.1
Source code	AI Dashboard compute platform (openVRE based) Repository https://github.com/DataTools4Heart/AI-dashboard-platform-demo Access public AI Dashboard customised front-end (openVRE based) Repository https://github.com/DataTools4Heart/AI-dashboard-frontend-demo Access public AI Dashboard tool submitting federated tasks via the FEM-orchestrator Repository https://github.com/DataTools4Heart/AI-Dashboard-FEM-runner Access public
Access	Test deployment (BSC node): https://datatools4heart.bsc.es/vre/
Documentation	

CogStack text analytics platform	
Description	CogStack is an open-source framework designed to extract, integrate, and process unstructured clinical data using natural language processing (NLP). As a modular platform, it includes advanced tools like MedCAT for automatic named entity recognition and linking (NER/NEL) and NiFi for flexible data flow orchestration. CogStack serves as the backbone for the DT4H Toolbox components implementing NLP modules and data preprocessing pipelines, enabling uniform and efficient deployment across FDNs.
Task/WP	WP3
Source code	CogStack NiFi Repository https://github.com/CogStack/CogStack-NiFi.git

	Access public CogStack MedCAT library Repository https://github.com/CogStack/MedCAT Access public
Access	Restricted access Currently Installed and tested at: BSC, SIEM, UMCU, AMC, KCL
Documentation	DT4H compiled documentation: Cogstack documentation Official documentation: https://github.com/CogStack/CogStack-NiFi

onFHIR stack	
Description	The onFHIR stack is a comprehensive data management solution designed to facilitate the integration, transformation and exchange of healthcare data according to the FHIR (Fast Healthcare Interoperability Resources) standard. The stack ensures data harmonisation to the DT4H CDM by means of the “Data Ingestion Suite”, comprising the FHIR server and a ETL tool (toFHIR) implementing FHIR mappers. On the other hand, the “Feature Extraction Suite” extracts relevant features and statistics from the FHIR server for cohort discovery and generation of AI-ready datasets

	
Task/WP	WP2
Source code	Feature Extraction Suite Repository https://github.com/DataTools4Heart/feature-extraction-suite/ Access public Data Ingestion Suite Repository https://github.com/DataTools4Heart/data-ingestion-suite Access public Common Data Model Repository https://github.com/DataTools4Heart/common-data-model Access public
Access	Restricted access. Test deployment (SRDC): https://matrix.srdc.com.tr/dt4h/feast/api/Dataset
Documentation	onFeast guideline: https://github.com/DataTools4Heart/feature-extraction-suite/blob/main/docker/DT4H_Feature_Extraction_Guideline.docx

Federated Execution Manager (FEM)	
Description	(FEM is a software component designed to coordinate and manage the execution of distributed processes in a FL or decentralised compute environment where multiple devices or clients, i.e. the FDN, provide compute and data resources. FEM is responsible for controlling and orchestrating the lifecycle of the federated tasks. It consists of two parts, (a) an orchestrator module that ensures tasks reach the multiple FDNs in a secure and coordinated manner, and (b) a client module installed at the participating FDN that pulls such tasks and allocates them in the local infrastructure.
Task/WP	T5.1/T6.1
Source code	FEM orchestrator Repository https://github.com/DataTools4Heart/FEM-orchestrator Access public FEM orchestrator setup for DT4H Repository https://github.com/DataTools4Heart/dt4h-FEM-orchestrator-config Access private FEM client Repository https://github.com/DataTools4Heart/FEM-client Access public FEM client setup for DT4H Repository https://github.com/DataTools4Heart/dt4h-FEM-client-config Access private
Access	Orchestrator (BSC): https://fl.bsc.es/flmanager/API/v1/ Tested clients: BSC, UB, GEM, KUH, UCL, UMCU
Documentation	openAPI documentation: https://fl.bsc.es/flmanager/API/v1/docs

Secure Multi-Party Computation (SMPC) cluster	
Description	The Secure Multi-Party Computation (SMPC) module enables secure collaborative computations across the federation employing cryptographic techniques. Built on the SCALE-MAMBA framework, the SMPC network consists of three key components: the coordinator, which manages API requests; the SMPC nodes, which execute the computation protocols; and the client nodes,

	which handle secure data importation. The platform integrates an SMPC cluster for providing an alternative federated learning aggregation strategy with strong guarantees for input privacy. It is employed by FLCore, one of the Toolbox components.
Task/WP	T4.5
Source code	SMPC cluster deployment Repository https://github.com/DataTools4Heart/smpc-for-mip/ Access public
Access	
Documentation	https://github.com/GPikra/smpc-for-mip/blob/main/Documentation/Documentation.pdf

Medical Informatics Platform (MIP)	
Description	The Medical Informatics Platform (MIP) is a decentralised e-infrastructure developed under the Human Brain Project (HBP) for clinical data federation and performing large-scale analysis. It offers advanced statistical tools and machine learning (ML) algorithms for feature extraction and predictive modelling. Within the DT4H platform, MIP serves as an alternative to the FEM, implementing its own federated learning network and tools. 
Task/WP	T6.1
Source code	Exareme2 custom engine for MIP platform

	Repository https://github.com/madgik/exareme2 Access public MIP portal backend Repository https://github.com/HBPMedical/portal-backend/ Access public MIP front-end Repository https://github.com/HBPMedical/portal-frontend/ Access public MIP frontend-backend middleware Repository https://github.com/HBPMedical/gateway Access public MIP deployment tools Repository https://github.com/HBPMedical/mip-deployment Access public
Access	http://88.197.53.15/access
Documentation	https://gitlab.ebrains.eu/hbp-mip/mip-docs