Call: HORIZON-HLTH-2021-TOOL-06
Topic: HORIZON-HLTH-2021-TOOL-06-03
Funding Scheme: HORIZON Research and Innovation Actions (RIA)

Grant Agreement no: 101057062

AI powered Data Curation & Publishing Virtual Assistant

Deliverable No. 2.3

Solution Design Document for G1

Approval by the European Commission Pending

Contractual Submission Date: 30/06/2023

Actual Submission Date: 30/06/2023

Responsible partner: P1 - University of Maastricht (UM)

Funded by
the European Union

Grant agreement no.	101057062
Project full title	AIDAVA - AI powered Data Curation & Publishing Virtual Assistant

Deliverable number	**D2.3**
Deliverable title	**Solution Design**
Type[1]	DEM
Dissemination level[2]	PU
Work package number	2
Work package leader	P1-UM
Author(s)	Isabelle de Zegher, P2-b!lo Louis Powell, Remzi Celebi, P1-UM Bela Bihari, Daniel Dallos, P9-GND
Keywords	'virtual assistant', 'data curation', 'solution architecture' 'knowledge graphs'

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HaDEA).

Neither the European Union nor the granting authority can be held responsible for them.

Document History

Version	Date	Description
V1	30 June 2023	First submission of the D2.3

List of Definitions

The definitions used in the deliverable are based on the AIDAVA Glossary [ref].

[1] **Type**: Use one of the following codes (in consistence with the Description of the Action):
R: Document, report (excluding the periodic and final reports)
DEM: Demonstrator, pilot, prototype, plan designs
DEC: Websites, patents filing, press & media actions, videos, etc.

[2] **Dissemination level**: Use one of the following codes (in consistence with the Description of the Action)
PU: Public, fully open, e.g. web
SEN: Sensitive, limited under conditions of the Grant Agreement

Table of Contents

Executive Summary 4
1. Introduction 5
2. Description of Activities 6
3. Problem to be solved and data in scope 8
    3.1 Overview 8
    3.2 Actors (people and systems) 9
    3.3 Type of data sources to be managed 10
4. AIDAVA Solution Design 11
    4.1 High level architecture diagram 11
    4.2 Backend: foundational repositories & related Epics 12
    4.3 Backend: Orchestration module – including Ingestion and Curation 20
    4.4 Non functional Epics 24
    4.5 Initiatives and Epics 26
5. Human in the loop 30
    5.1 Principles of user interaction 30
    5.2 User profile supporting user interaction during curation 31
6. Formal representation of Solution Design: C4 model 33
    6.1 Context level 34
    6.2 Container level 35
7. Deployment, Maintenance and Support 37
    7.1 Deployment & configuration of the system 37
    7.2 Customer support 38
8. Market considerations for the AIDAVA "product" 40
    8.1 Approach to market: mass market or target market? 40
    8.2 Potential customers 40
9. Conclusion and next steps 42
Annexes 44
    A.1 Wireframe extracts 44

Executive Summary

Availability of integrated, high-quality personal health data (PHD) remains limited, with impact on quality & cost of care and limiting possibilities for research and analytics. Indeed, PHD is currently distributed, heterogeneous, captured through different modalities, with variable quality. Findability and Accessibility of this data – following the FAIR principles – is addressed in numerous projects; Interoperability and Reuse remains a challenge due to several factors that are addressed by the intelligent virtual assistant being prototyped in the AIDAVA project. Concretely, the objective of the project is to maximise automation in data curation & publishing[3] of heterogeneous PHDs while empowering individuals – patients or their deputies and data curators – when automation is not possible due to lack of contextual information. Through the data curation workflows (*Deliverable 2.2. Data curation and publishing process*), the AIDAVA virtual assistant prototype is expected to transform each patient's PHD into a Personal Health Knowledge Graph (PHKG). All PHKGs will be generated in compliance with the AIDAVA Reference Ontology (*Deliverable 2.1. AIDAVA Reference Ontology as a Global Data Sharing Standard*). From the PHKGs and the mapping information contained in the Reference Ontology, the publishing module will generate different target outputs as required by different use cases (see *Deliverable 1.1. Description of Use Cases*).

This deliverable focuses on the solution design of the AIDAVA prototype virtual assistant. The solution includes a **backend** and a **frontend**. The **backend** includes foundational components such as the user directory, a master data reference repository, the catalogue of data sources with metadata supporting automation, a library of curation tools used in the automation workflows, the reference ontology which is the standard of reference for each PHKG, the repository of patient data – from raw format to PHKG and published format in the target standard – and the overarching orchestration module that supports automation and interaction with the end users. The **frontend** is the module that interacts with the end users. In Generation 1 of the AIDAVA prototype, user interaction will be minimum; in Generation 2, the user interface will build on advanced technologies from human-computer interaction, with explainability to facilitate understanding of the questions raised by the virtual assistant during the curation process. Explanations will be tailored to users categorised through different user profiles as identified in *Deliverable D1.2. Report from user survey with personas canvas.*

These different components are described in the solution architecture, with the related Epics, in turn consolidated in Initiatives for the development team. In addition, the first 2 levels of a formal description of the system - based on the C4 model - is provided; an in-depth description of lower levels of the C4 models will be developed in *Deliverable D3.1. VA Architecture*.

This deliverable also describes the proposed support model to be implemented when evaluating the prototype across the different sites.

Finally, this deliverable introduces potential target customers. A full market analysis, with in-depth analysis of customers, market size and market potential for a product that could be developed on the result of the AIDAVA project will be provided in *Deliverable 2.4,* as an updated version of this deliverable, after evaluation of the first generation of the prototype.

[3] By "***data curation & publishing***" we mean the integration, harmonisation and quality enhancement (curation) and the transformation into a target format (publishing) of multimodal data, collected from various sources, to make it more usable by humans and machines.

1. Introduction

This document is the result of the work performed within Task 2.3.

- Section 2 highlights the different activities that took place during the first part of this task.
- Section 3 builds on the use cases (Deliverable D1.1) to confirm the key actors of the systems and the different types of data we can expected to have as input
- Section 4 builds on the business requirements (Task 1.3/Deliverable D1.3), the automation requirements (Task 2.1/Deliverable D2.2) and background information from P2-b!lo to define the key components of the backend and the frontend of the AIDAVA prototype including :
    - User Directory with metadata on users supporting tailored user interactions
    - Master Data repository
    - Catalogue of data sources with metadata on data sources, supporting automation
    - Library of data curation tools, including conditions of use
    - Patient Data store with different levels, from raw data, to PHKG, to published data
    - AIDAVA Reference Ontology serving as a standard of reference to ensure interoperability across PHKGs
    - Orchestration of the whole process including ingestion, automated curation and publishing

  It then identifies the different initiatives and related Epics to be used by the development team for G1 and G2.
- Section 5 builds on the personas developed in Task 1.2/ Deliverable D1.2 and on wireframes developed previously by P2-b!lo and expands on the high-level requirements for the user interface to be mainly developed as part of Task 5.3.
- Section 6 provides a formal description of the AIDAVA prototype based on the 2 highest levels of the C4 model, to be expanded in the technical architecture (D3.1).
- Section 7 describes the deployment approach as well as the proposed customer support model during evaluation.
- Section 8 introduces the market analysis, by identifying different target customers. This analysis will be further developed with the AIDAVA Sustainability Advisory Board – including senior executives across regulators, pharma, authorities, hospitals and vendors as well as senior representatives from the partners – and will be described in the update of this deliverable (D2.4) due in December 2024.

2. Description of Activities

Development of the solution design took place through a set of parallel activities, building on background information from different partners that was subsequently compiled into a common strategy.

- **Consolidation of the problem to be solved and scope of the solution.** The first step was to consolidate the problem to solve from the use cases (D1.1), and more specifically to confirm the type of data sources and the different types of actors. This is described in Section 3.

- **Validation of conceptual solution design.** Building on the conceptual architecture provided in the proposal, and background information from P2-b!lo on the architecture and the data flows, we consolidated the proposed components of the solution and data processing workflows. This was initiated at the Tallinn business workshop (December 2022) where the data workflows were discussed with a team of clinicians, business analysts, data stewards and data privacy experts, and further clarified through regular meetings. This helped to refine the foundational components of the solution architecture of AIDAVA (see Section 4.2), as well as to clarify the data privacy requirements which were taken into consideration in the Data Management Plan (D4.2) and the Generic Data Protection Information Assessment (DPIA) checks included in D4.4.

- **Validation of automation potential.** A critical part of the AIDAVA project is the potential for automation based on the orchestration of different workflows with supporting tools, developed in Task 2.1/Deliverable D2.2. Integration of the workflows and orchestration layer was discussed through bi-weekly meetings across WP2 partners and consolidated at the Sofia technical workshop (March 2023). This is further described in Section 4.3.

- **Human-In-The-Loop and User Interface.** In AIDAVA, we aim to minimise user intervention (to simplify the whole curation process for patients) and to maximise clarity when the system requires an answer.
  To gather input from end users and support formalisation of the user interface, we used wireframes developed by P2-b!lo.
  To ensure alignment between the automation process and human-interaction, we ensured that each workflow defined in Task 2.1/D2.2 includes automated and manual steps where the manual step includes a request for content or for decision from the human. In addition - in collaboration with Task 1.2 - we extracted the characteristics of the user profile needed to ensure targeted human interaction. The integration of these different aspects (user profile, automation, tailored human interaction) was discussed during the Vienna workshop in June 2023, on WP2 and WP5 alignment, as WP5 will be responsible for implementing the automation workflows (Task 5.2) and the human interaction during curation (Task 5.3)

- **Formalisation of the solution and transfer of knowledge to the development team.** A solution design remains conceptual, prone to interpretation by the software development team. To avoid any misinterpretation, we decided to adopt the C4 model. In this document, we describe the 2 first levels of the C4 model (See Section 6), which were developed through regular meetings between P1-UM and P2-b!lo and validated by P9-GND; this is laying the ground for a smooth transition to the two more detailed levels to be further expanded in WP3.
  In addition, the solution architecture and the Epics were organised in Agile Initiatives (see Section

4.5) that were discussed with the development team during the Brasow Workshop in June 2023; this allows to clarify multiple questions from the developers and ensure they understand the product intent.

- A first draft of the proposed model for **Deployment, Maintenance and Support** of the prototype is provided in Section 7. It was validated through regular meetings with the "site administrator" of 5 partners who will be participating in the evaluation of the prototype (P1-UM University of Maastricht, P7-MUG University of Graz, P8-NEMC Northern Estonian Medical Center, P13- MID MIDATA and P14-DME DIGI.me).

- In terms of potential customers and potential market, a first analysis was provided by P2-b!lo based on background work. There were initial discussions with the two partners (P13-MID and P14-DME) who act as **Health Data Intermediaries (HDI)** and are key actors to support the patients in managing and controlling their data and are an inherent part of the AIDAVA project and the evaluation of the prototype. Discussion on how to integrate the HDI started during the Tallinn workshop in December 2022, expanded during the discussion related to the evaluation of the prototype (Task 1.4) and were followed by meetings with the HDI partners.

3. Problem to be solved and data in scope

3.1 Overview

The goal of the AIDAVA project is to increase Interoperability and Reusability – following the FAIR principles – of Personal Health Data (PHD) through integration and semantic enrichment within multimodal personal health knowledge graphs (PHKG) based on the AIDAVA Reference Ontology. This will be achieved by optimising the data curation & publishing process and by orchestrating complementary AI technologies into a virtual assistant that maximises automation and helps users – clinical staff and potentially patients – when automation is not possible, while adapting to users' preferences and skill levels.

Existing data curation tools support transformation, integration and quality enhancement but are siloed, i.e. they solve a specific curation step (e.g., metadata management by Data Catalogue, transformation by ETL, concept extraction by Text Mining, entity reconciliation by Master Data Management, image feature extraction through Deep Learning, imputation of missing data through statistical methods), when in fact all these steps might be necessary. In some cases, these tools are insufficient (e.g., no semantic enrichment by linking different data sources, limited amount of concepts extracted with Text Mining). As a result, data curation & publishing remains a costly, time-consuming process managed by expert data stewards; as a direct consequence, much available PHD is not curated and not reusable – or portable – in clinical care and clinical research.

AIDAVA proposes to solve the problem by orchestrating multiple tools through a structured workflow (see Deliverable D2.2) to maximise automation in the transformation of the heterogeneous PHD into a multimodal Personal Health Knowledge Graph (PHKG) as displayed in Figure 1, enriched and curated gradually as new data sources are added and compliant with the AIDAVA Reference Ontology. A PHKG can in turn be transformed into multiple different target formats – following different needs – based on mappings predefined within the AIDAVA Reference Ontology.

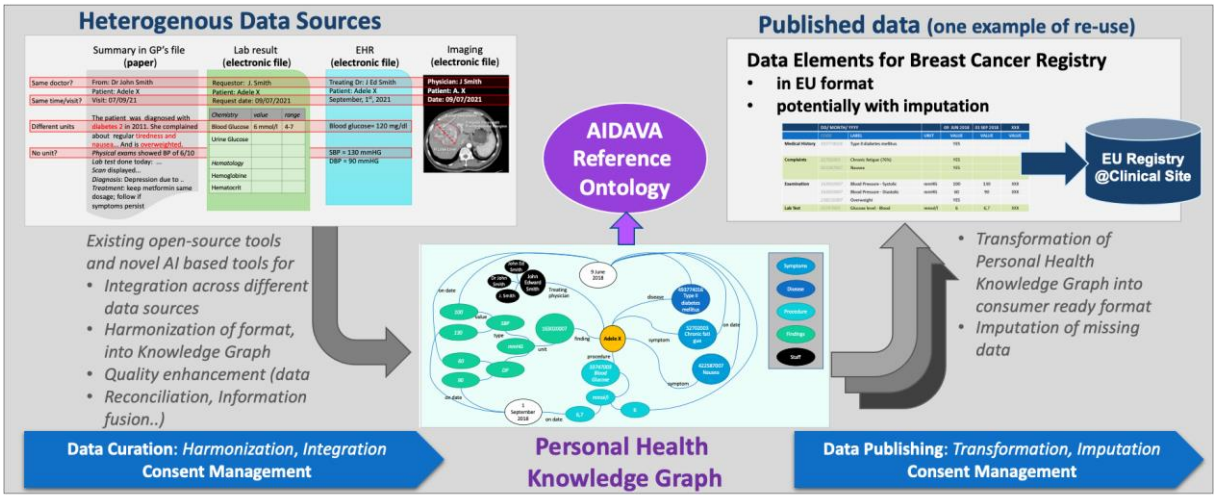


Figure 1. Each PHKG is constrained by the AIDAVA Reference Ontology

3.2 Actors (people and systems)

Based on the detailed description of the 2 use cases provided in Deliverable D1.1 and additional analysis during regular WP1 meetings, we derived a set of actors, including software systems (in light blue).

User	Roles
AIDAVA Administrator	Responsible for configuring – and testing – the solution with all components reusable across sites, before actually deploying the system in local sites. ● Perform onboarding of data curation tools for each release of the prototype and test their execution with a set of agreed test data ● Define and set-up default template with user profile ● Define output of publishing supporting the use cases (extract for Breast cancer registry, extract to compute CVD score, HL7 IP extract to transfer to HDI)
SITE Administrator	Responsible for configuring the solution at local level ● Execute onboarding of local data source in data catalogue ● Register and manage local users (based on template user profile) ● First level support
SITE Source System	System – Hospital Information System – in production in a local site that includes Data Sources to be curated in AIDAVA. There must be a Data Sharing Agreement (DSA), including technical Data Transfer Specification (DTS), for each Source system to support transfer to AIDAVA.
HDI – as Source System	HDI is linked to a local site that includes Data Sources to be curated in AIDAVA. There must be a DSA/DTS for each Source System to support transfer to AIDAVA.
Patient	As data curators, patients ingest their personal data from the different systems, initiate the curation process and answer questions from the system whenever needed. As data consumers, the patients request to forward their personal data to the local HDI for further visualisation. *Note: Before using the AIDAVA prototype, the patients need first to sign an Informed Consent Form (ICF) by which they agree that the system will access and process their personal data, and that curators will also have access to their personal health data. This ICF also allows the patient to stop using the system at any point in time. In this case, their personal data included in the system will be either transferred to them or deleted.*
Curator	The expert curator (also called data steward in some site) supports patients in the curation of their data, with the consent/agreement of the patient
Data User	Data Stewards, health care provider, scientific staff having access to extract published from AIDAVA (with patient consent) ● Breast Cancer specialist who can perform analytics on a "Breast Cancer" registry spread across the 3 sites ● Cardiovascular specialist who has access to an automatically computed risk score (instead of having to compute it manually) and can more effectively monitor CVD risks.
HDI – as Data	HDI system linked to a local site that consumes data extracted from AIDAVA.

User	Roles
User System	There must be a formal agreement between AIDAVA and the HDI to send them data; the same DSA/DTS used for ingestion into AIDAVA can be used for this. For the prototype the data consumed by the HDI will be in HL7 IPS format.
AIDAVA VA	Virtual assistant supporting the user with a maximum of automation and requesting information only when needed.

3.3 Type of data sources to be managed

Following the use cases – and the list of potential data sources identified for these use cases – we worked with the different partner hospitals to extract de-identified examples of the data sources with personal health data information.
These different data sources include different interoperability issues further described in D2.1.

	Example	Description (and where it is coming from)
1. Unstructured	Discharge summary GP notes Clinical notes	Reports with clinical narratives; either typed directly by a clinical staff or – most often – auto generated from different parts of the dossier
2. Semi-structured	Medical record	Data resulting from capture by a human – contains a combination of free text, numeric and coded values
3. Structured (small)	Lab results Medical device (limited data points)	Data resulting from procedures executed by a machine and providing data in specific format
4. Structured (Image – DICOM)	*Rx* *MRI*	*Data resulting from reports on imaging*
5. Structured (large – TBytes)	Medical device (movelets)	Out ot scope for AIDAVA
6. Structured (OMICS)	Genomics Metabolomics	Out ot scope for AIDAVA

4. AIDAVA Solution Design

4.1 High level architecture diagram

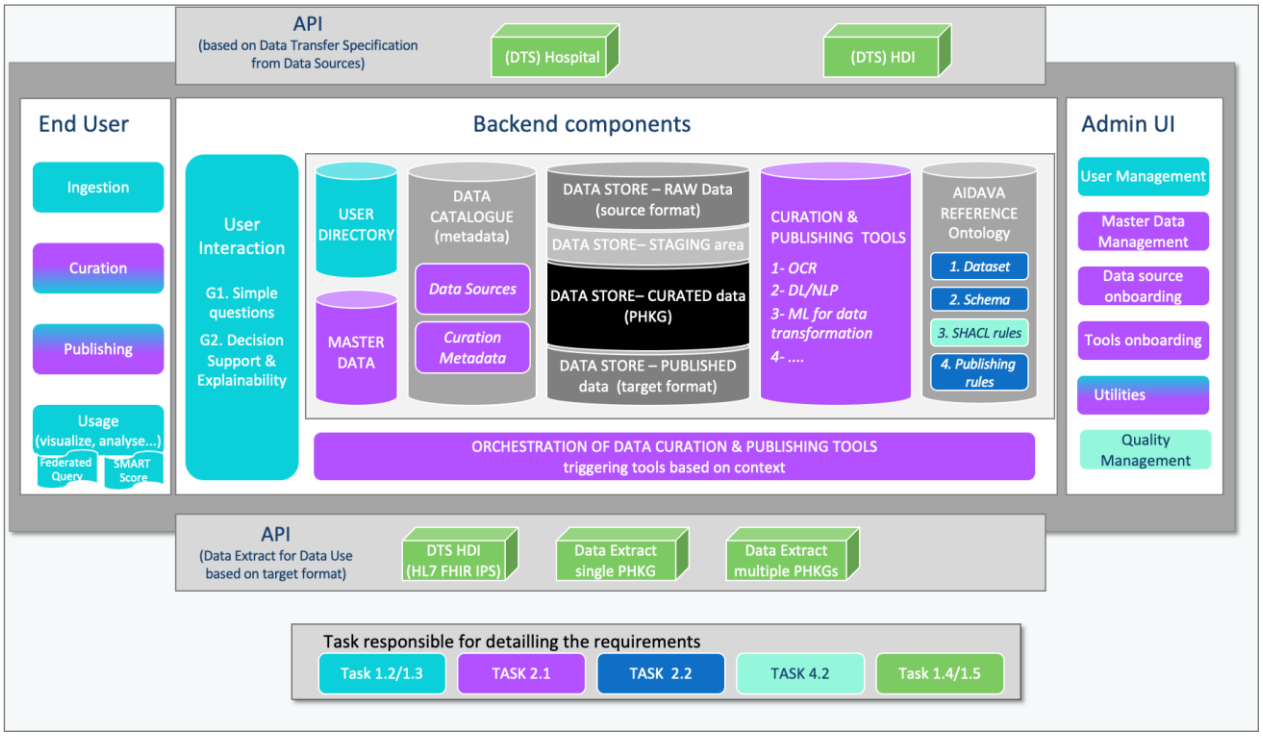


Figure 2. Architectural overview of the AIDAVA prototype with the relationship to the specific tasks defining the requirements for the respective component

The AIDAVA prototype architecture includes the following components.

1. The backend includes
    a. foundational data- and metadata repositories of the system (Section 4.2)
    b. the orchestration module that is responsible for automation (Section 4.3)
    c. the human-in-the-loop decision support and explainability module enabling adaptive interaction – with explanations – with the user (Section 5)
2. The Admin functionalities support the system administrator to configure the system to maximise automation and user adaptability (included in Section 4.2)
3. The End User functionalities support user interaction around ingestion, curation and publishing (included in Section 4.3) as well as the one related to visualisation (Section 5)
4. API that enables ingestion from connected data sources (Section 4.4)
5. API and additional functionalities that support the use of published data (Section 4.4) .

These components are described in the sections below, please note the legend for the data flows.

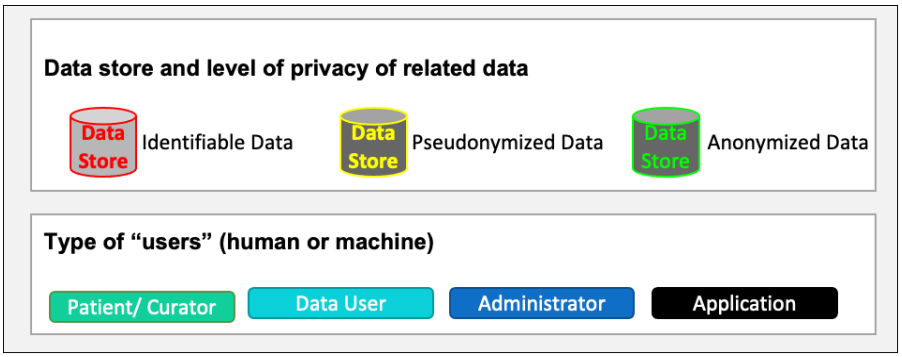


Figure 3. Legend used in the data flows

4.2 Backend: foundational repositories & related Epics

4.2.1 User Directory and User Management

COMPONENT	**USER DIRECTORY** Repository which contains data related to all users of the systems (including administrators) and including the following information. ● Default profile & roles (provided at AIDAVA level across sites): ○ Specification of the role with related access rights (i.e access to different functional components) ○ Template profile of typical users of the system – based on Personas identified in Task 1.2 – with their roles and related access rights ● User profile of each registered user (provided at Site level) with ○ Identification information that allows to uniquely identify the user – including *a pseudonym to be used when processing individuals data and supporting linkage across data sources.* ○ Information about the user relevant in the context of data curation (e.g. skill set including health literacy and digital literacy) ○ Specification of default profile of the user as well as their preferences of enabling adaptability in the interaction; the user profile should be updated as the system – through the explainability module – learns from the interaction with the user
EPIC	**USER MANAGEMENT** ● *AIDAVA Administrator:* Define/manage templates with different user profiles and related access rights, based on roles (AIDAVA admin level). Typical user profiles are described in the AIDAVA personas (D1.2) ● *Site Administrator:* Manage registration and update of local users by linking the user with a default profile and confirming access right based on their role ● *AIDAVA VA:* Creates the root PHKG with identification information about the patient and core identifiable information such as gender, ethnicity… ● *End user (patient, expert curator, data user)* is able to manage /update their preferences

	USER INTERACTION ● Login into the system ● Support dialogue with the user to capture information needed to advance the workflow (G1 = simple dialogue, G2 = dialogue with explanations) ● The number of questions raised to the patient/data curators should be limited (e.g. for NLP output need to specify an upper limit of number of questions per data source). Whenever user input is required, the AIDAVA prototype will ○ decide if the question should be raised to the patient or to an experienced data curator based on the type of question and skill set of the user ○ provide explanations to the patient and user – at their level of understanding according to their profiles ○ *for G1, description of the content of questions (T2.1/5.3) and to whom those should be addressed, based on user profile (T5.3)* ○ *for G2, extension to the content of the questions with explanations (T5.3)*
Assumptions for the prototype	● The prototype will come with predefined user profiles defined at AIDAVA level across sites ● User accounts – for business users – will be created when the system will be deployed in each site, following an information session, specifically for the patient and the curator, during which he/she will be asked to answer a questionnaire to assess their skills (e.g. medical and computer literacy) that will drive user interaction. Additional aspects such as preferences of interaction, languagewill also be added when creating the user account.
Data Flows (see figures below)	The data flows below describe the registration of local users by the Site Administrator. There are two main types of users ● The administrators, the patient and the local curators, identified as persons with their personal credentials. (Note: local data curators will automatically be assigned to the Curator role for the patients in the context of the prototype and following signature of the Informed Consent Form by the patient). ● The data users: they can be an organisation or an individual acting in a specific role but not as a private individual. They should provide credentials as data users, not as private individuals.

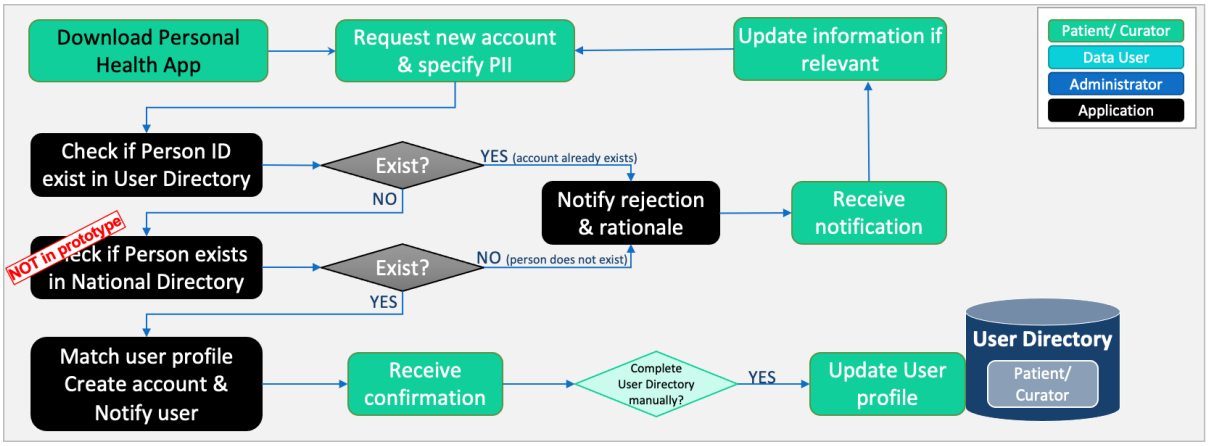


Figure 4. Onboarding Users (Patients, Curators and Administrators)

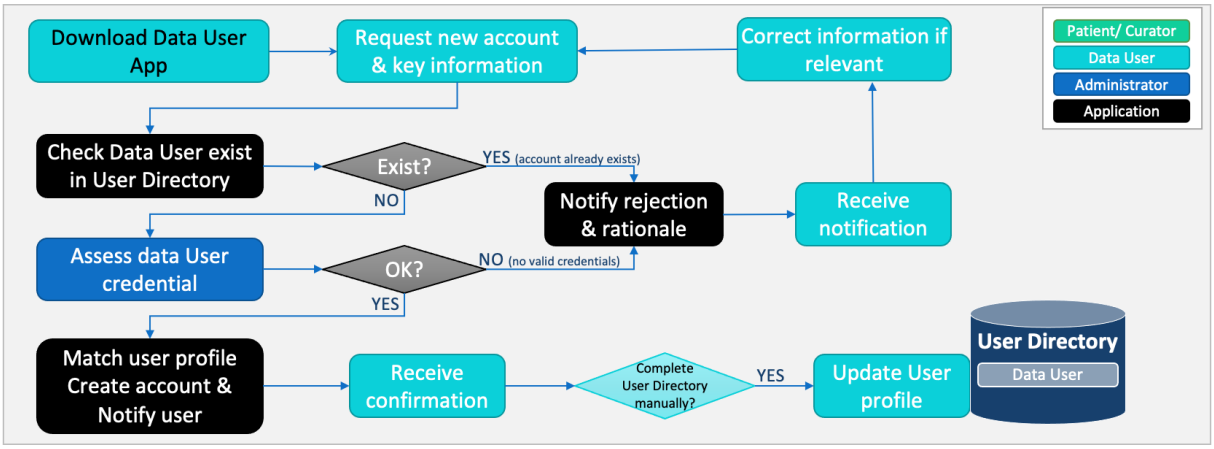


Figure 5. Onboarding Data Users

4.2.2 Master Data repository and Master Data Management

COMPONENT	**MASTER DATA REPOSITORY** Represents data about the business entities that provide context for business transactions across multiple systems & data sources They are typically used as a "look up" table and are dynamic, i.e. business entities can be added on an ongoing basis. This information is important to support entity deduplication of master data (e.g., name of a Healthcare Provider). In AIDAVA, there are different types of "master" business entities ● Organisations which can be a hospital, a department within a hospital, a health data intermediary, a data holder (app provider), and potential national authorities. ● Clinical Staff that intervene at different points in time in the life of a patient within the organisation. ● Lab test description – including unit- and reference ranges, which can be specific per organisation but also per type of instrument.

	Note: In the long term, this information should be available at the national level; this is an area that will be developed as part of the deployment of the European Health Data Space (EHDS)
EPIC	**MASTER DATA MANAGEMENT** ● The Site Administrator updates the repository with relevant information for the site before starting the evaluation, as part of the system configuration. ● In addition, the Site Administrator should be able to update the repository during the evaluation/running of the application, if/when a new occurrence of an entity is detected.
Assumptions for the prototype	● Information about the local hospital and the related HDI will be entered manually when deploying the system on the local site ● Information about the health care professionals active in the TA within the hospital will be entered as well when deploying the system on the local site. However, we cannot expect that all HCP identified in the patient medical record will be included; the system will allow the data curator to add new HCPs if/when a new occurrence is detected.
Data Flows	Not applicable: Master Data will be managed directly by the Site Administrator

4.2.3 Catalogue of Data Sources and Data Source Onboarding

COMPONENT	**CATALOGUE OF DATA SOURCES** This catalogue includes metadata on the data sources – including data potentially across multiple patients – that needs to be curated. The required metadata has been identified in Task 2.1 / *D2.2 Details on curation and publishing process,* and will be implemented as an extension of the metadata described for the FAIR Data Points; it is key to supporting the automation of the curation process. Indeed, this metadata allows to identify the interoperability issue of the source component to be curated and to select the appropriate curation workflow (and related tools).
EPICS	**DATA SOURCE ONBOARDING** *Site Administrator:* onboard data sources following workflow described in Deliverable D2.2 and update the catalogue of data sources with all descriptors (i.e. metadata) of the data sources to maximise the potential for automation.
Assumptions for the prototype	Before deploying AIDAVA, a data sharing agreement, one with Data Transfer Technical Specification (DTS), needs to be signed among the different parties (hospital product system to hospital-based AIDAVA prototype; hospital-based AIDAVA prototype and Health Data Intermediary / HDI). The DTS serves as the basis for the onboarding process of the data sources when deploying the system locally. The data sources in scope have been identified in Deliverable *D1.1. Use case description.* Data Sources of a single patient are coming from the Hospital as well as from the HDI through asynchronous file transfers, compliant with the schema provided in the Data Transfer Specification (DTS) mentioned above.

	The Data Source Onboarding will follow the workflow described in Deliverable *D2.2 Details on curation and publishing process*. The expectation for the prototype is that as soon as a patient signs the consent to participate, their data will be transferred into the prototype
Data Flows (see figures below)	The onboarding process is supported by the AIDAVA Virtual assistant – but requires input from the Site Administrator.

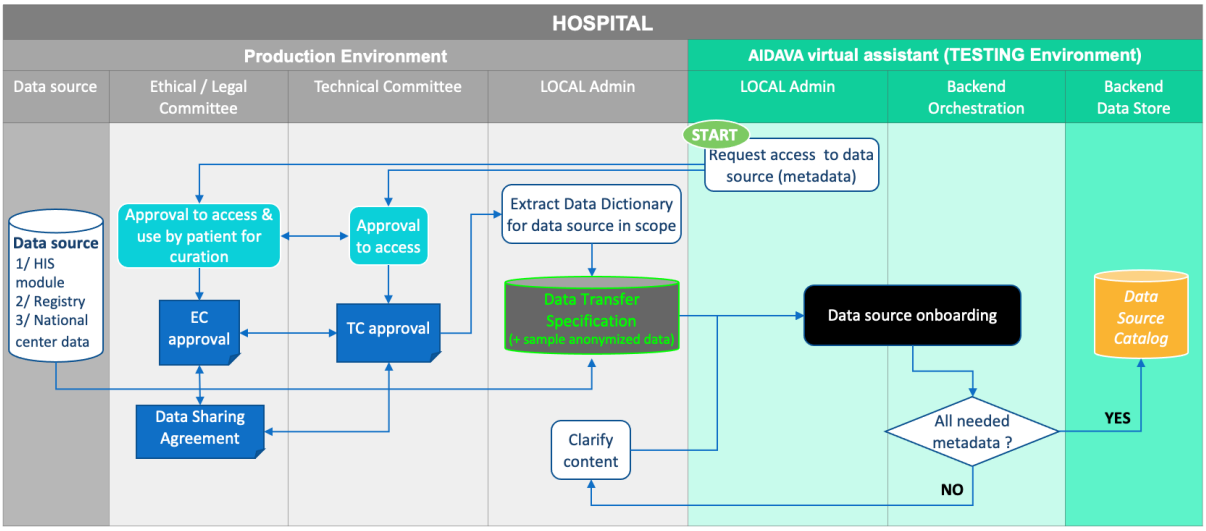


Figure 6. Onboarding Data source from Hospital

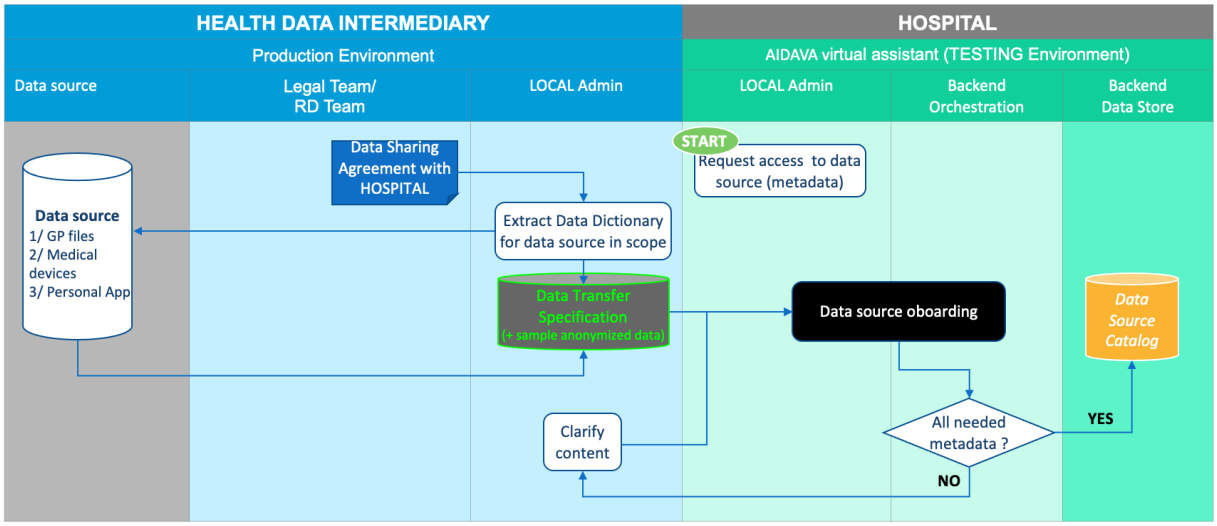


Figure 7. Onboarding Data Source from HDI

4.2.4 Patient Data store

COMPONENT	There are 4 levels of data stores, with different versions of the same data. ● **RAW DATA.** The RAW Data stores source data as received, in any electronic format (could be SQL dump or any type), following the DTS mentioned above. ● **STAGING area**. The Staging Area stores data sources being processed up to the curated format in knowledge graph / RDF format. Each source related KG is then integrated within the individual's Personal Health Graph (PHKG) and transferred to the Curated Data Store when all issues related to integration across data sources have been solved. ● **CURATED area.** Final PHKG. The Curated Data is stored in the integrated PHKG of each individual patient. ● **PUBLISHED area**. Stores any type of data (personal identifiable data as well as pseudonymized/anonymized data in structured format) derived from the patients' PHKG in response to a specific data request. Note: In the prototype we will keep the published data to ensure auditability. In the long term, we could keep just the query to recreate the extract
EPIC	This component is used in the INGESTION and CURATION Epics In the context of the Data Curation, functionalities can be limited to data storage and -retrieval functions. (Automatic) archival should be included in a later stage.
Assumptions for the prototype	Empty at deployment time
Data Flows	Not applicable (this component is updated as part of the curation process)

4.2.4 Library of curation- and publishing tools, and onboarding of tools

COMPONENT	**LIBRARY OF CURATION & PUBLISHING TOOLS** This is a library of tools to support the data curation & transformation process Several tools related to matching & ranking and ETL are long-standing existing products on the market; AI-based tools are emerging. This list should be constantly updated: the wider and more powerful the tools, the more efficient end users will be in curating data.
EPICS	**CURATION TOOL ONBOARDING** *AIDAVA Administrator* onboards data curation & publishing tools that will be triggered by the system to perform specific curation steps to solve data interoperability issues (see *D2.2. Data curation and publishing process)*. This includes the following ● Update the library of curation tools for any tool for which a need has been identified in the curation and publishing process ● Integrate the tool so that it can be executed from the curation and publishing process with the appropriate input and return the appropriate output In case of error, the system should return the name of the tools with the provided input data and any additional information available from this curation tool ● Test the tool and its integration
Assumptions for the prototype	In AIDAVA, this library will contain ● for G1: existing products or open-source software identified in Task 2.1 ● for G2: same as G1 where some tools are replaced with novel tools developed in the project
Data Flows	Not applicable – done by **AIDAVA Administrator** (within the Development team)

4.2.6 Reference Ontology and Reference Ontology Management

COMPONENT	**Reference Ontology** Components – supported/ implemented within AIDAVA as described in *Deliverable 2.1. AIDAVA Reference Ontology as a Global Data Sharing Standard*) ● The **Dataset** with a collection of discrete items (concepts) in the domain in scope that describe the semantic (meaning) of these concepts. The meaning of a concept is typically expressed by value sets, terminologies and ontologies (e.g. SNOMED CT, LOINC). ● The **Ontology Schema** is the machine-usable format – in Resource Description Framework (RDF) and maintained in GraphDB from Partner Ontotext – representing the semantics of the concepts defined in the dataset; it is maintained by the ontology experts, based on the concepts identified in the dataset. It includes the CORE ontology and any extension

	needed for target format mapping. It is used to ensure interoperability across different data sources and to facilitate data exchange. ● The **SHACL rules –** also managed within GraphDB – include additional constraints on the concepts – typically scientific/medical constraints (e.g. a biological woman cannot be diagnosed with prostate cancer) related to the domain in scope, resulting from the consensus of scientific societies, and formalised within the Ontology to ensure consistency across data. Note: implementation of SHACL rules is further described in *Deliverable 4.6. Definition of Data Quality metrics & checking rules.*
EPICS	**Reference Ontology Management** The AIDAVA Administrator must be able to upload the reference ontology (including potential updates with version control) and ensure integration with the rest of the system to support data sources onboarding, ingestion, curation and publishing.
Assumptions for prototype	The content of the AIDAVA Reference Ontology is expected to evolve throughout the project, following the governance mechanism described in Deliverable D2.1, and through G1 and G2 of the AIDAVA prototype. The 2 generations of the prototype will therefore be deployed with the last version of the ontology with the components needed for the supported use cases
Data Flows	The governance process to update the AIDAVA Reference Ontology is defined in *D2.1 AIDAVA Reference Ontology as the basis for a Global Data Sharing Standard*

4.2.7 Utilities Management

COMPONENT	NA
EPICS	**UTILITIES MANAGEMENT** This includes computation of the quality score on the PHKG. The systems should allow for metrics and benchmarks on usage on the following components ● Data Catalogue / Data Source ● Data Catalogue / Curation Metadata ● Library of Data Curation tools ● Results of Quality Assessment
Assumptions for the prototype	● Support the users of the systems to visualise the quality of the PHKG based on quality metrics ● Support the federated queries across BC data stores ("BC registry") across the 3 hospitals
Data Flows	NA

4.3 Backend: Orchestration module – including Ingestion and Curation

4.3.1 Ingestion of personal data

COMPONENT	Ingestion is using the User Directory, the Catalogue of Data Sources and the Raw Data Store Note: For paper input from the patient, a "curation workflow" has been defined, starting from the scanning or picture of the paper document, followed by OCR step and then processing of the content)
EPICS	**INGESTION** Support the patient – or their designated local data curator – to ingest their own personal data from different data sources – in the available format. This includes ● Ingestion of paper-based data sources – by using OCR tools ● Ingestion of electronic data sources available across different data holders such as clinical care providers (hospitals, GP, nursing homes etc.), device- and app providers or Health Data Intermediaries (HDI) mediating data for a patient. Ingestion of electronic data sources requires a Data Transfer Agreement between the holder of the data sources and the AIDAVA prototype, as well as onboarding of the data sources (see Section 4.2.3).
Assumptions for the prototype	For the prototype we will assume 3 types of ingestion (see data flows below) ● Paper documents through OCR ● Data extracted from the Hospital (NEMC, UM/MUMC, MUG) based on the data sources identified in the use cases (see Deliverable D1.1) and agreed DTS; formats can be .csv, .txt,.json, .xml, .pdf, .jpg, .png. No direct integration with a Hospital Information System: ○ Data source onboarding to be performed on the Data Transfer Specification (DTS), with results stored in the Catalogue of Data Sources ○ Data is to be transferred from a hospital into the hospital's AIDAVA secure data store in accordance with the sharing agreement and the DTS. ○ Traffic only possible on agreed channels ● Data extracted from the HDI (MIDATA for MUG and NEMC, DIGI.me for UM/MUMC) – based on the data sources identified in the supported use cases, same approach as for hospitals (with DTS and data source onboarding)
Data Flows	See Data Curation

4.3.2 Curation of personal data

COMPONENT	The objective of AIDAVA is to maximise automation in the curation process and to minimise the need for user intervention. Curation is based on workflows, using the catalogue of data sources, the library of curation tools and the patient data stores. During the curation process, the system captures a set of metadata with context information supporting further steps in the automation process
EPICS	**CURATION** Support the patient or their designated local data curator – to curate their personal data and build the patient's Personal Health Knowledge Graph (PHKG) integrating all the data ingested into one common semantic model. *Deliverable 2.2. Data curation and publishing process* identified a set of data interoperability issues. Each issue is linked with a workflow that includes one or more tools to be triggered to solve the interoperability issue automatically, if possible. Management of the workflows for the different data sources is done through an orchestration workflow. Automation of curation requires that ● Data sources have been onboarded (see Data Source Onboarding Epic in Section 4.2.3)) ● Curation tools have been onboarded (see Curation Tools Onboarding Epic in Section 4.2.4) In case the workflow requires additional information, the frontend is activated (see User Interaction Epic in Section 4.2.1) to capture the needed information.
Assumptions for the prototype	The system must be able to orchestrate execution of the different workflows for the onboarded data sources and to call on the curation tools included in the library of curation (& publishing) tools.
Data Flows (see figures below)	Three different data flows have been identified 1. Paper ingestion and curation – starting from a hard-copy paper document provided by the patient 2. Electronic file (except for imaging) coming from the hospital 3. Electronic file from the HDI **Note 1.** At the time of writing this document, processing of imaging was not finalised; it will be added in the updated version of this deliverable.

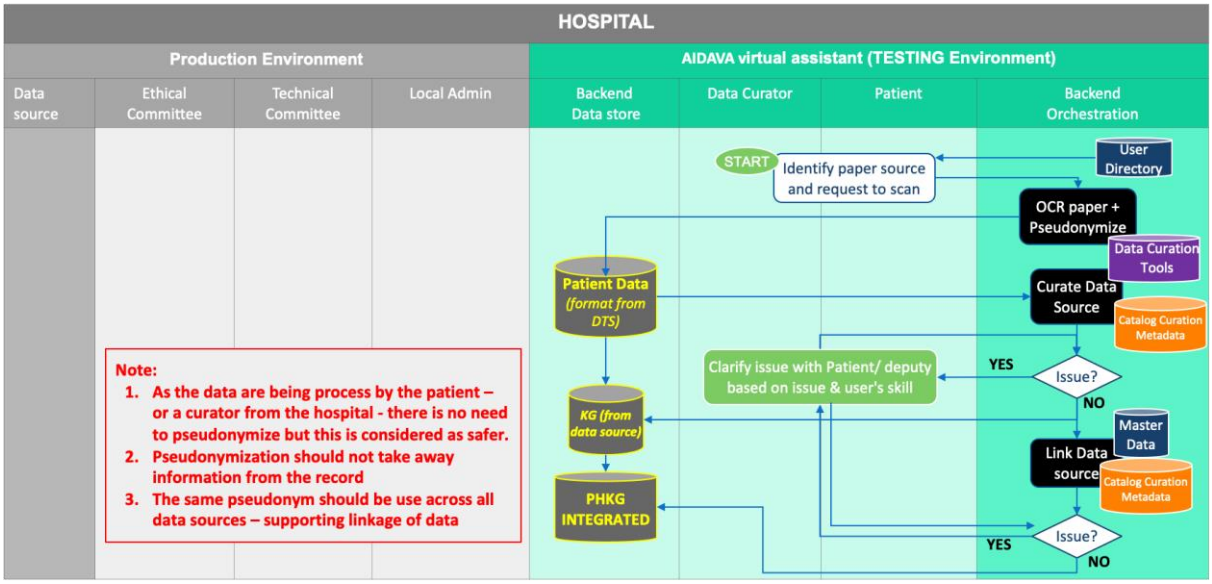


Figure 8. Ingestion Flow of Paper Data from the Patient into the PHKG

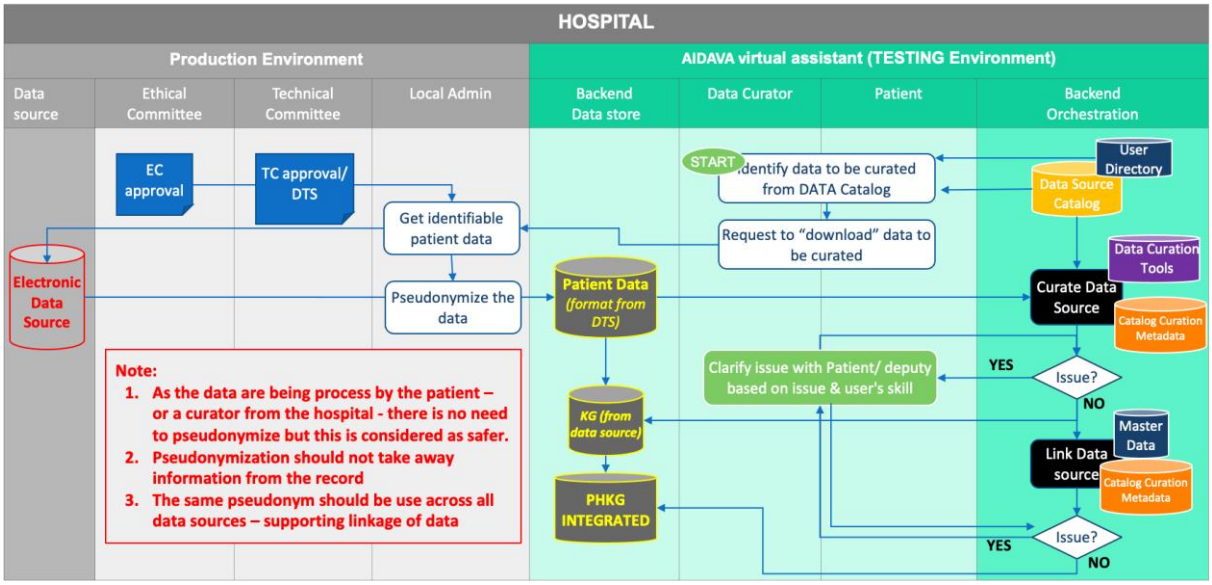


Figure 9. Curation Flow: HIS/Hospital electronic file to PHKG

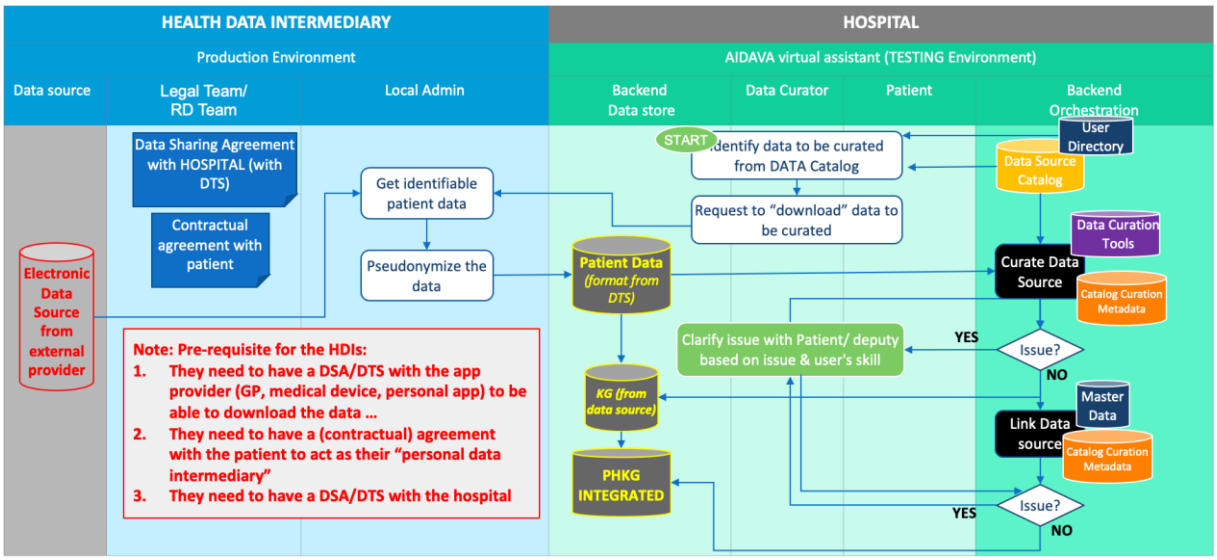


Figure 10. Curation Flow: HDI to PHKG

4.3.2 Publishing (identifiable and non-identifiable data)

COMPONENT	Publishing is based on the CURATED data store and the PUBLISHED data store, as well as the library of curation tools
EPICS	**PUBLISHING** Support the user – patient or any other consented data user – to extract from a patient-specific PHKG, or from a set of PHKGs, a data set that is further transformed into the target format required for the 3rd party application that will use the data set. The extracts generated are kept within the AIDAVA prototype, ready to be shared.
Assumptions for the prototype	In the context of the AIDAVA prototype, the extract will be predefined to support the 3 uses cases ● **Extract 1**. Patient IPS (HL7 FHIR format) extracted from the patient PHKG to be transferred to the HDI – with patient consent, ● **Extract 2** (breast cancer use case). List of data elements identified as part of the "EU breast cancer registry", extracted from the PHKG of the breast cancer patients, in the specified format. ● **Extract 3** (Cardiovascular use case). Data elements needed to compute the SMART score extracted from a patient's PHKG.
Data Flows (see figures below)	There are 2 main workflows ● One for identifiable data for a single patient/ PHKG: (1) extract to be transferred to the HDI and (2) extract to compute the CVD SMART score ● One for non-identifiable data (pseudonymized or anonymized) across multiple patients/ PHKGs to deliver the Breast Cancer Registry at each site

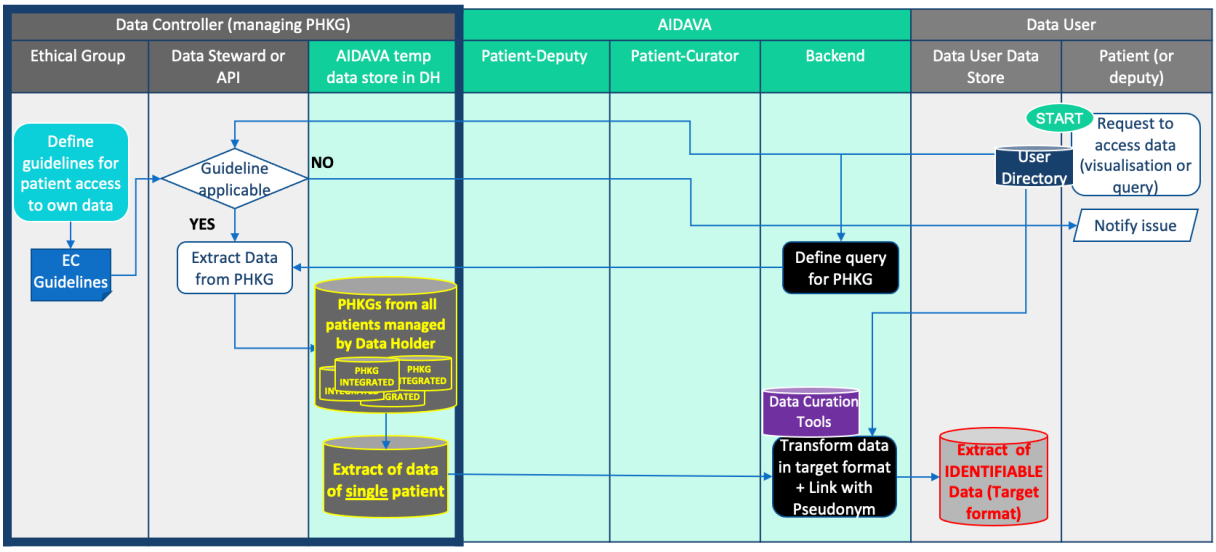


Figure 11. Publishing. Single PHKG to identifiable data set for 2 data uses

1. *HDI: with HL7 IPS to be transferred*
2. *Hospital HCP: CVD extract to compute the CVD score*

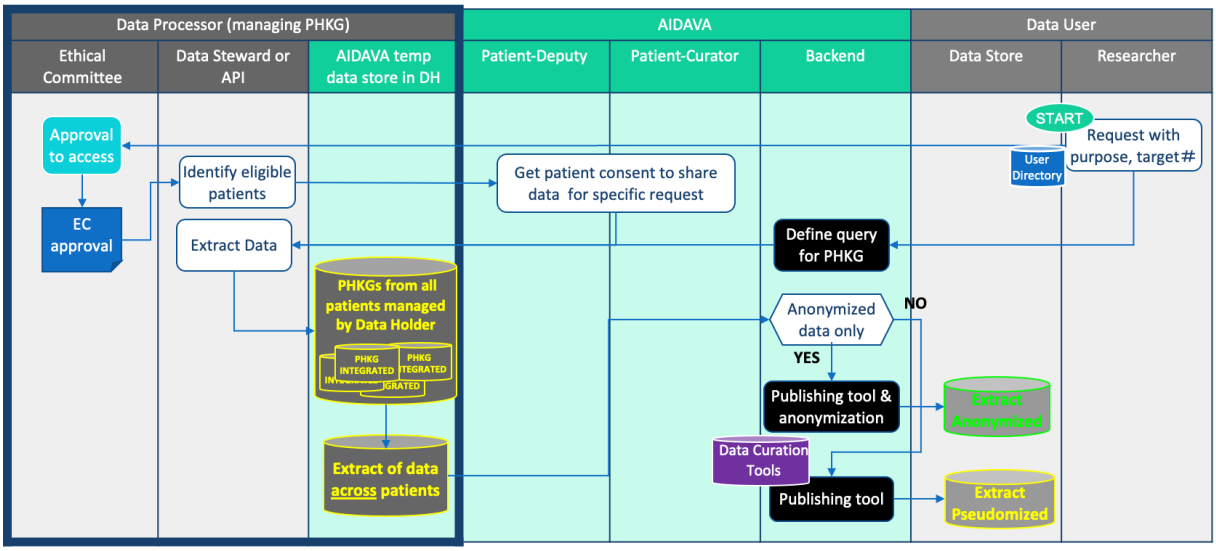


Figure 12. Publishing. PHKG to pseudonymized & aggregates – to be used for the Breast Cancer extract

4.4 Non functional Epics

D1.3 Business requirements is providing a detailed list of functional and non-functional requirements. The table below contains the most critical ones

EPICS	Objectives
Integration	The AIDAVA will function standalone. It will however require data transfer based on a formal Data Sharing Agreement (DSA) with Data Transfer Specification (DTS) ● DSA/DTS from the hospital to AIDAVA to support ingestion of patient data ● DSA/DTS from the HDI to the AIDAVA prototype to support (1) ingestion

EPICS	Objectives
	of patient data into AIDAVA and (2) transfer of personal data – in the form of HL7 IPS – back to the HDI for visualisation and use by the patient In addition, AIDAVA requires the integration of 2 software components ● A federated query (see Section 3.2.1 in Deliverable D1.4 – Annex with Study Protocol) – to compute parameters across sites – that is to be executed in all 3 sites and to return the consolidated results in each of the 3 sites (Peer to Peer request) ● The SMART score algorithm (see Section 3.2.2 in Deliverable D1.4 – Annex with Study Protocol) that is executed locally and produces a score for the physician (internal in each hospital site)
Audit trail	As the AIDAVA prototype is managing personal data, and is intended to become a fully fledged medical device, any transaction should be linked to an audit trail (who did what when).
Security	As the AIDAVA prototype is managing personal data, the system should support the following requirements ● Data stored within the hospital environment and transferred following pre-agreed data sharing agreement/ data transfer specification ● Data stored in a secure environment, with no data loss ● Role-based security access: access must be authorised based on access rights (e.g. an administrator should not have access to the personal data of a patient) ● Disaster recovery for the prototype – not to be used for clinical care and decision making – this is not critical
Parameters	● **Performance**. Response time: less than 3 seconds ● **Availability** = min 8 hours per day/ 5 days per week (working hours) with 90% to be operational during this period ● **Reliability**/ Mean time between failures (MTBF) = 90% with a Mean Time To Repair (MTTR) of 48 hours
Quality Management	Support the users of the systems to visualise the quality of the PHKG based on quality metrics as introduced in the Study Protocol (Annex 1 of Deliverable 1.4) and to be further detailed in *Deliverable 4.6 .Definition of Data Quality metrics & checking rules*

4.5 Initiatives and Epics

Figure 13 below groups the different components of the solution architecture, identified in Figure 2, in initiatives on which the development team can further concentrate

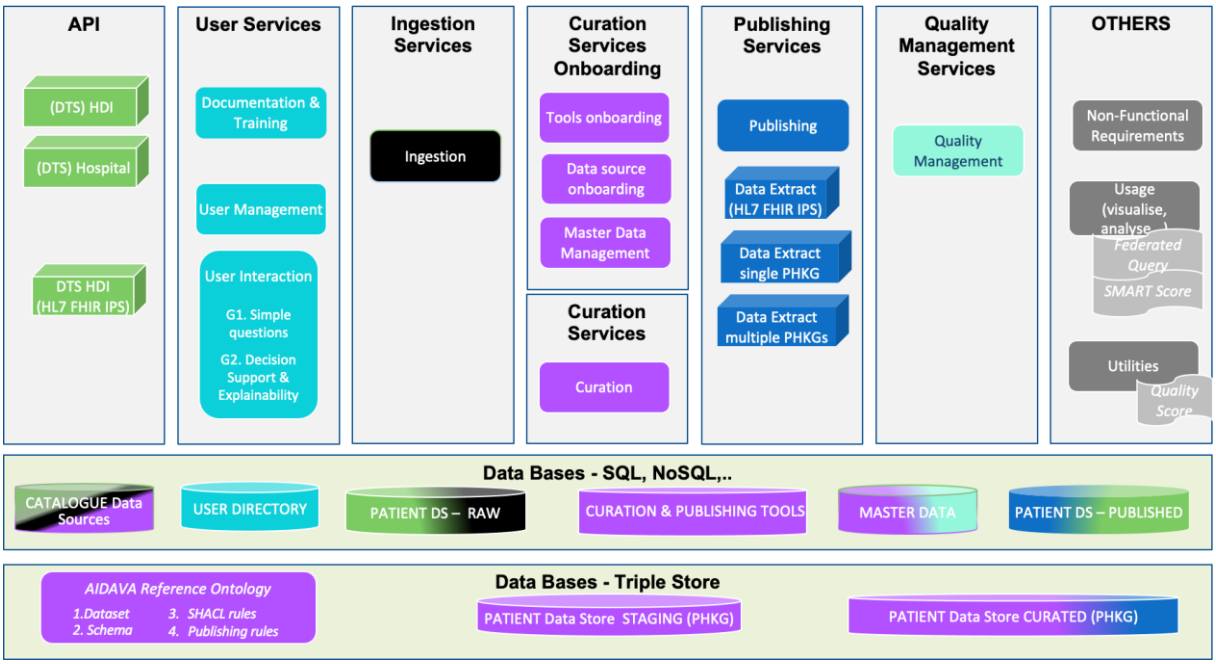
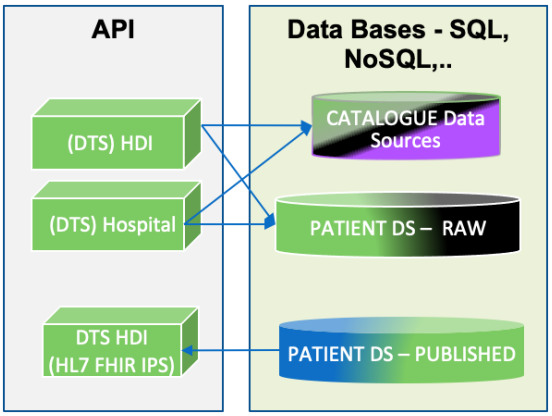


Figure 13. Grouping of Epics through initiatives

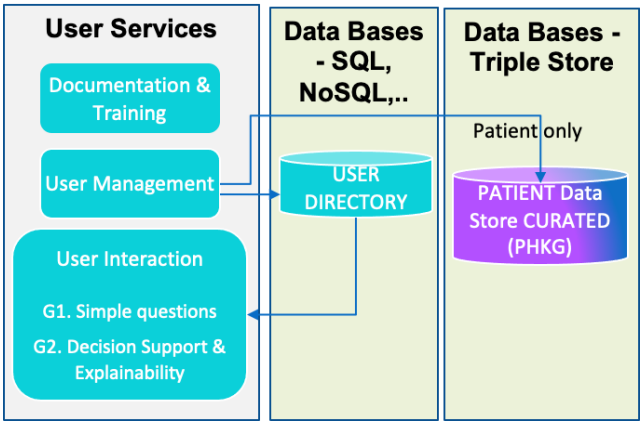
The colour code indicates how the different databases relate to the different initiatives

The following initiatives have been identified

- **"API"** relates to interfaces to access patients' data from the hospital and HDI, through asynchronous data transfer based on formal specifications. The data transfer specifications of each data source are included in the "Catalogue of Data Sources"; the patient data transferred from the different data sources is stored into the "Patient DS – Raw". API also relates to interfaces transferring published data – and stored in "Patient DS – Published" – to the HDI of the patient.

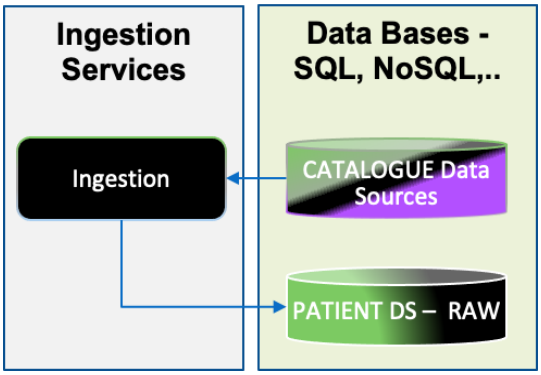


- **User Services** includes documentation and training, as well as User Management (including account management, secure login, user preferences). As part of account management, when registering a patient the "root" part of their PHKG will be created with basic identifiable information such as birthdata, gender, ethnicity. User Services also includes

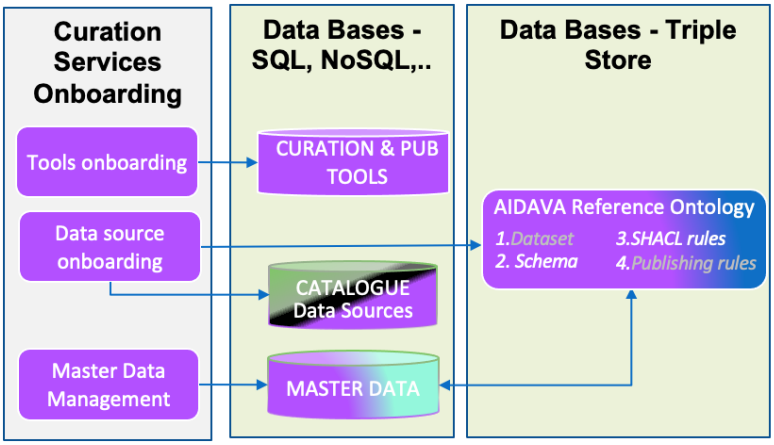


management of the user interaction specifically during the curation process with simple questions for the Generation 1 of the prototype and additional decision support during Generation 2.
This initiative relies on data stored in the User Directory with user profiles and preferences,a s well as the Patient DS - Curated PHKG.

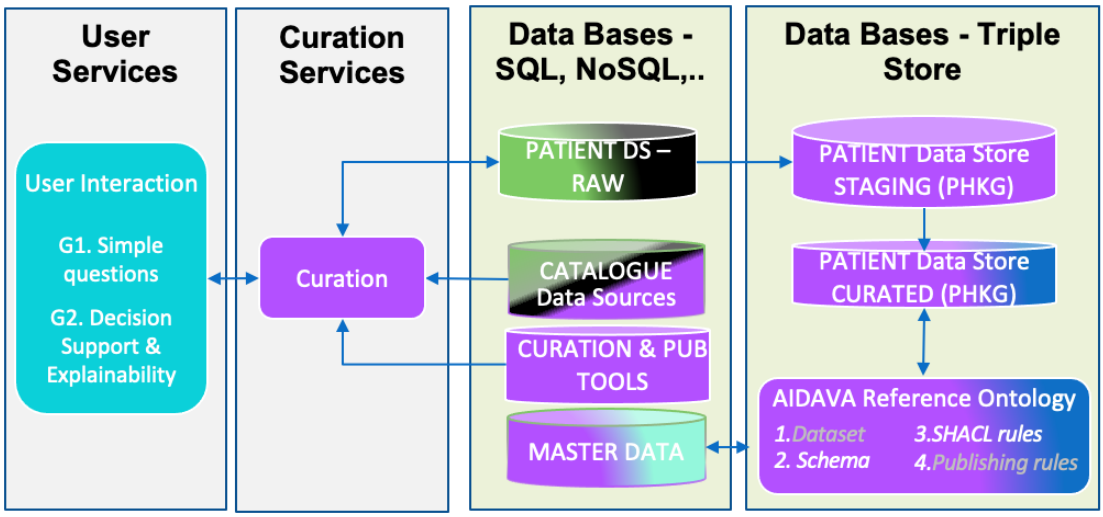
- **Ingestion Services** relates to the ingestion of the different data sources of the patient – documented in the Catalogue of Data Sources – by the patient themself, or the curator, from the hospital and/or the HDI. In real life, this would activate the data transfer from the data sources (hospital or HDI) within AIDAVA; in the context of the prototype, this will mark the data available in the "Patient DS – Raw" SQL DB as ready to be curated.



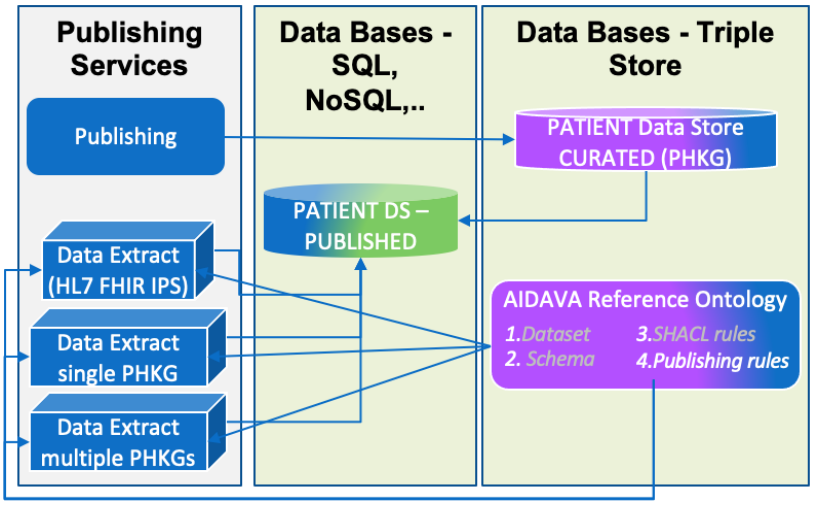
- **Curation Services Onboarding** is the first part of the curation process: it focuses on entering information needed to support the process with maximum automation.
  - The curation tools onboarding Epic is responsible for updating the "Curation & Publishing Tools" library with information on the tools identified in the curation workflows (Task 2.1 and WP5).
  - The data source onboarding Epic (by the Site Administrator) ensures updates to the "Catalogue of Data Sources" in each site, with mapping information between the data source and the AIDAVA Reference Ontology (Schema).
  - The master data management Epic ensures that master data is properly updated (by the Site Administrator) when deploying the system locally. The Master Data includes parameters (e.g. reference range) that support the execution of SHACL rules.



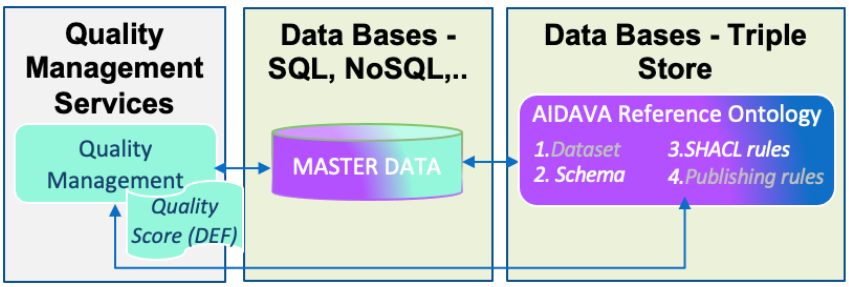
- **Curation Services** relate to the automation of the curation of the patient data stored in the "Patient DS – Raw". Each data source is curated – through different workflows, triggered by the metadata describing this data source in the Catalogue, and executing the relevant curation tools onboarded previously. Whenever user input is required, the curation workflow manages the user interaction through the User Interaction module (simple in G1, advanced in G2). The curated data source is stored in the "Patient DS – Staging" Triple Store; it is then integrated with the patient's PHKG, checked for completeness and consistency through the related workflow and then stored in the "Patient DS – Curated" Triple Store. While checking for completeness and consistency, SHACL rules may be invoked.



- **Publishing Services** are responsible to transform the data included in the "Patient DS – Curated" triple store (PHKG) into the required target format based on publishing rules contained in the Reference Ontology. The resulting data extracts are stored in the "Patient DS – Published" data store (from where they can be further processed).

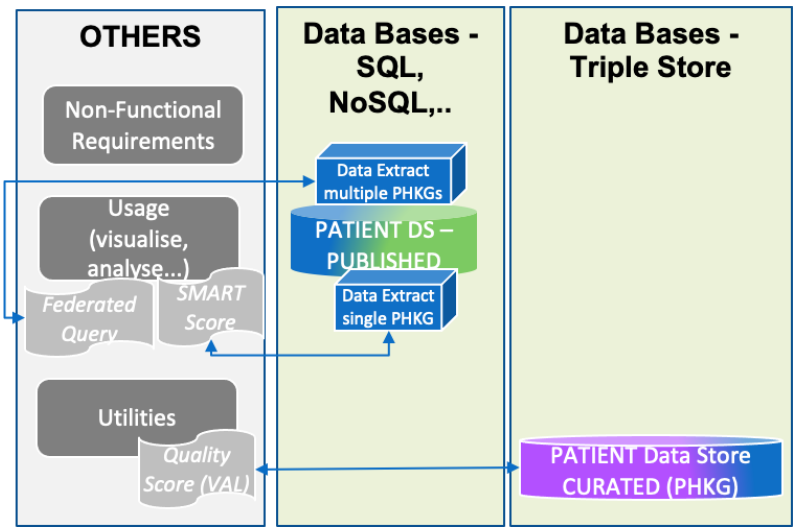


- **Quality Management Services** relates to the definition and use of quality rules (in the form of SHACL rules) to check and assess quality. Task 4.2/ Deliverable 4.6 is concerned with the definition of these rules, including scientific rules extracted from community-wide consensus scientific guidelines, correctness rules linking measurement values with ranges specified in the Master Data. Quality Management also relates to the definition of an overall quality score of the curated data.



- **Others** includes the non-functional requirements (security, performance, traceability, provenance,...), the data usage functions related to the use cases (execution of a federated

query across extracts from all Breast Cancer patients in the site, computation of the SMART risk score from the patient's PHKG extract). This initiative also includes the computation of the quality score for data reuse of the patient's PHKG.



5. Human in the loop

5.1 Principles of user interaction

We identified different users in Section 3.2; user interaction with the prototype will be different for each of them.

User	Expected user interaction
AIDAVA Administrator	Update information on curation tools (Curation Tools Onboarding) with simple tools (e.g a spreadsheet to be uploaded or a simple user interface) requiring advanced computer literacy and/or programming skills
SITE Administrator	Update information on data sources (Data Source Onboarding) with simple tools/ user interface requiring advanced computer literacy and/or programming skills
Patient	End users for ingestion and curation ● For data ingestion, the UI should be as simple and as intuitive as possible in compliance with user stories described in *Deliverable D1.3. Business requirements* ● For data curation, the objective of AIDAVA is to maximise automation and to minimise user interaction. Automation is achieved through a set of related workflows defined in Deliverable 2.2. Each workflow includes different steps that can be a. a call to execute a curation tool b. a decision taken by the system based on context metadata c. a request to an end user, requiring information to further process/curate a data source d. a human decision The last two steps will activate the "Human In The Loop" module that interacts with the end user taking into account the context and the profile of the user.
Curator	
Data User	End users (patient, curator or any other agreed actor) for publishing – based on the same principles as data ingestion

To ensure acceptance and usability, the AIDAVA user interface for the patient, curator and data user must have the following features

- Multi-lingual, supporting at least Estonian, Dutch and German. The Administrator user interface can be in English
- Desktop or mobile/touch based, where interaction through mobile with patients will be prioritised as much as possible. Data Curator and Administrator interfaces can be desktop based.
- Adaptive to different user profiles and skill sets: the explanations provided to the curators (patient and expert curator) whenever automation is not possible and input from the curator is needed, must be adapted to the user profile. Indeed, different patients will have different levels of data, health and digital skillset.

P2-b!lo designed a set of wireframes before the project (see Annexes). These wireframes were presented first to the patient representatives who confirmed their preference to have a mobile-based

application (smartphone or potentially tablet computer) as they may not always have access to a desktop; they also confirmed that it was more important for them to have something "simple, intuitive and useful" than something "beautiful".
The wireframes were also presented and discussed with the software development team to explain the principles of the system.
To manage user interaction, AIDAVA will take into account the user preferences and the user profile stored in the User Directory.

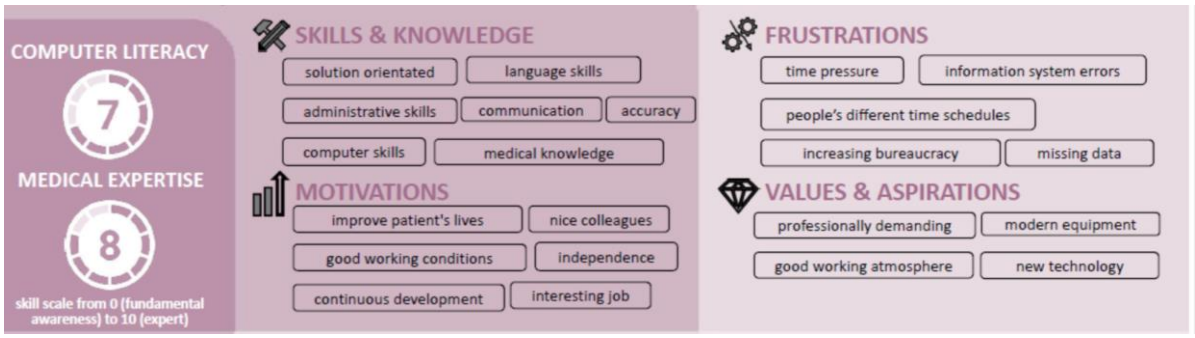
- The user preferences include information on the user such as preferred type of interaction (smartphone, tablet or desktop), preferred language, request to delegate certain actions (like ingestion or curation) etc.
- The user profile includes information on the role, access rights and skills of the user. The latter is particularly important to direct curation requests to the appropriate user and to tailor explanations during the curation process as described below.

5.2 User profile supporting user interaction during curation

A patient should not be considered as "one size fits all". Patients are first and foremost citizens with different skill sets that will impact their capacity (and interest) to manage and curate their data. One patient can be a medical data scientist, with a high level of health- and computer literacy; another patient can as well be in a profession that does not require much proficiency in health and computers. They are both patients. As a consequence, we decided to link the questions to be raised not to the type of users but rather to the level of skills described in the user profile.

We identified 2 major skill sets (see Deliverable 1.2.)

- Medical expertise (scale 1 to 10). This attribute will be used in the AIDAVA prototype to decide if a question that occurs in the curation workflow can be sent to the patient or should be sent to the expert curator
- Computer literacy (scale 1 to 10). This attribute can be used in the AIDAVA prototype to provide guidance and information to the user in a customised level of technical complexity



As part of more extensive work to be done in Task 5.3, we will consider using other attributes such as motivation, frustration and aspirations (all based on a set of predefined values).

As mentioned above, the curation process is based on a set of workflows (see Deliverable 2.2) that

- attempt to curate the data automatically by calling and executing the appropriate curation tools

- request input from the user, whenever the system requires additional semantic information to solve a data curation issue. It can be a missing date for a specific event that is required (as most of health data is time-stamped e.g. date on which a diagnostic was made) or a missing unit of a measurement procedure (e.g. missing unit for blood pressure measurement in the narrative of the physician). This information can only be provided by a human who understands the context of the patient and, for the second example, has some medical knowledge.

AIDAVA is providing two generations of the prototype.

- Generation 1 (G1) is based on existing curation tools. From a user interaction perspective, it is expected to be a clear – though simple – question.
    - The content of questions is described in Deliverable 2.2
    - The way to formulate the question in a simple and intuitive way, as well as the level of medical expertise and computer literacy required to answer the question, will be developed in Task 5.3.
      AIDAVA will then require input from the patient by default, unless they don't have the health- or computer literacy level required
- Generation 2 (G2) will include novel tools developed in the project. User interaction will be more extensive with true interaction with the user – with explanations – through a NLP interface. Speech Recognition will be considered as well.

6. Formal representation of Solution Design: C4 model

The C4 model is a hierarchical approach to software architecture modelling that provides a set of diagrams for visualising the components, containers, and context of a system. It is designed to be simple, lightweight, and scalable, making it easy to communicate and understand the system's architecture across different stakeholders. It is often used in agile and DevOps environments, where quick and efficient communication and collaboration among developers, architects, and stakeholders are essential.

It consists of four levels, where each level represents a different level of abstraction:

1. **Context**: This level provides an overview of the entire system and its interactions with external systems, showing the system's boundaries, users, and external dependencies.
2. **Containers**: This level describes the high-level components of the system and their interactions, showing how the system is decomposed into smaller, more manageable parts.
3. **Components**: This level describes the internal components of each container and how they interact with each other, showing how the components are organised and connected within a container.
4. **Code**: This level represents the lowest level of abstraction, showing how the components are implemented in code.

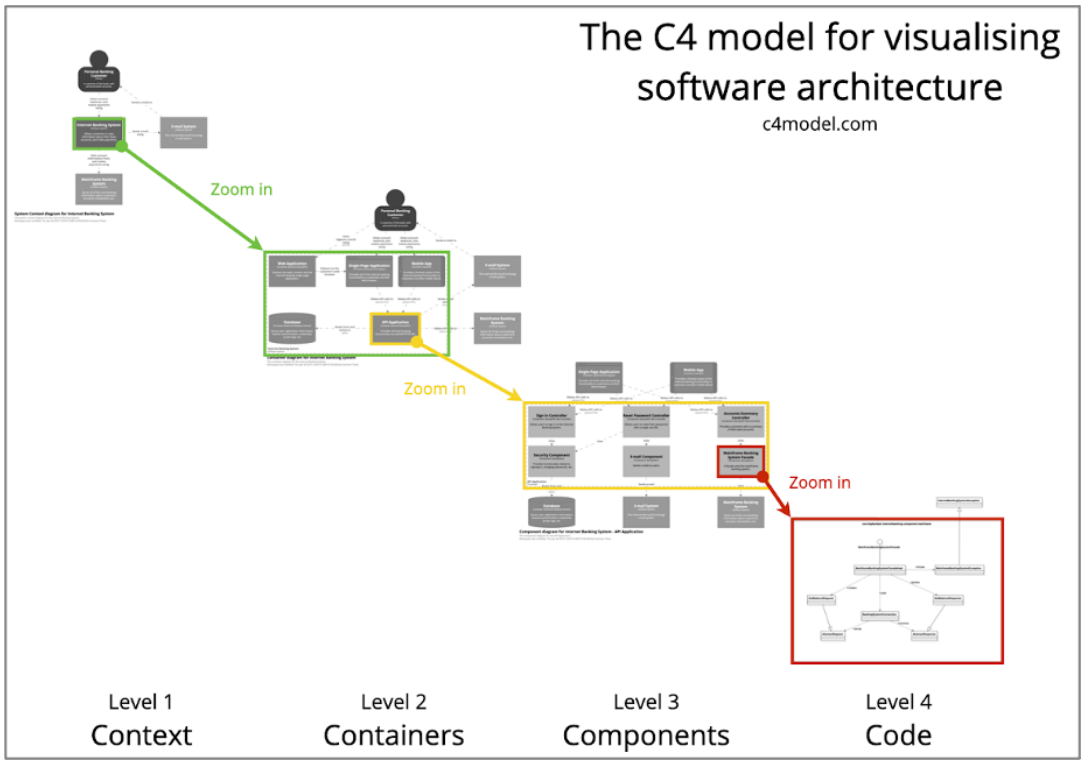


Figure 14. Levels of C4 description

This deliverable is focusing on level 1 (Context) and level 2 (Containers), which have been mapped to the solution design described in Section 4.1
Level 3 (Components) is part of the technical architecture in Task 3.1 and level 4 (Code) is developed in the detailed requirements – Task 3.2, and development – Task 3.3.

6.1 Context level

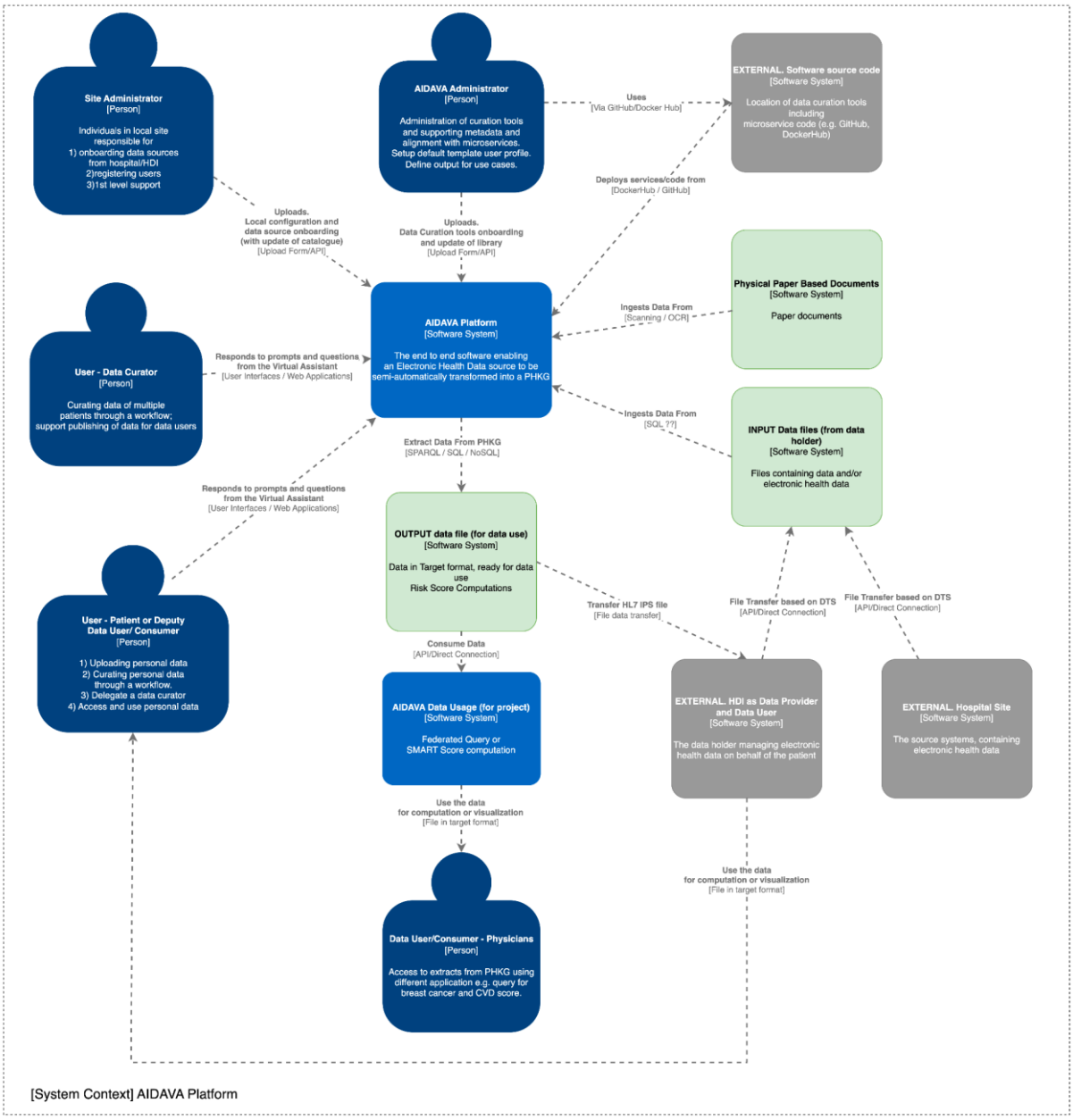


Figure 15. AIDAVA – C4 Context level

The AIDAVA platform will have different types of users as described before (Administrators and End Users). It will take as input either

- a physical paper copy[4] of the patient's health related documents, that will be scanned by the patient themself or their relative.
- electronic documents coming from the hospitals or from Health Data Intermediaries; the documents in scope of the prototype are described in Deliverable D1.1.

To support data curation, AIDAVA will use a set of curation tools – managed within the Library of Curation Tools – and integrated into the system as microservices by the AIDAVA Administrator.

[4] Paper documents are still important in certain countries, certainly for legacy documents.

AIDAVA will produce different outputs in different formats. For the prototype, this includes

- HL7 IPS of the patient will be transferred to the relevant Health Data Intermediary
- Extract for Breast Cancer registry and SMART CVD score which will remain within AIDAVA for further processing.

The data usage function will include among others

- the algorithm support computation of the SMART risk score for individual patients,
- A federated data query that can be executed across the 3 hospitals participating in the evaluation.

6.2 Container level

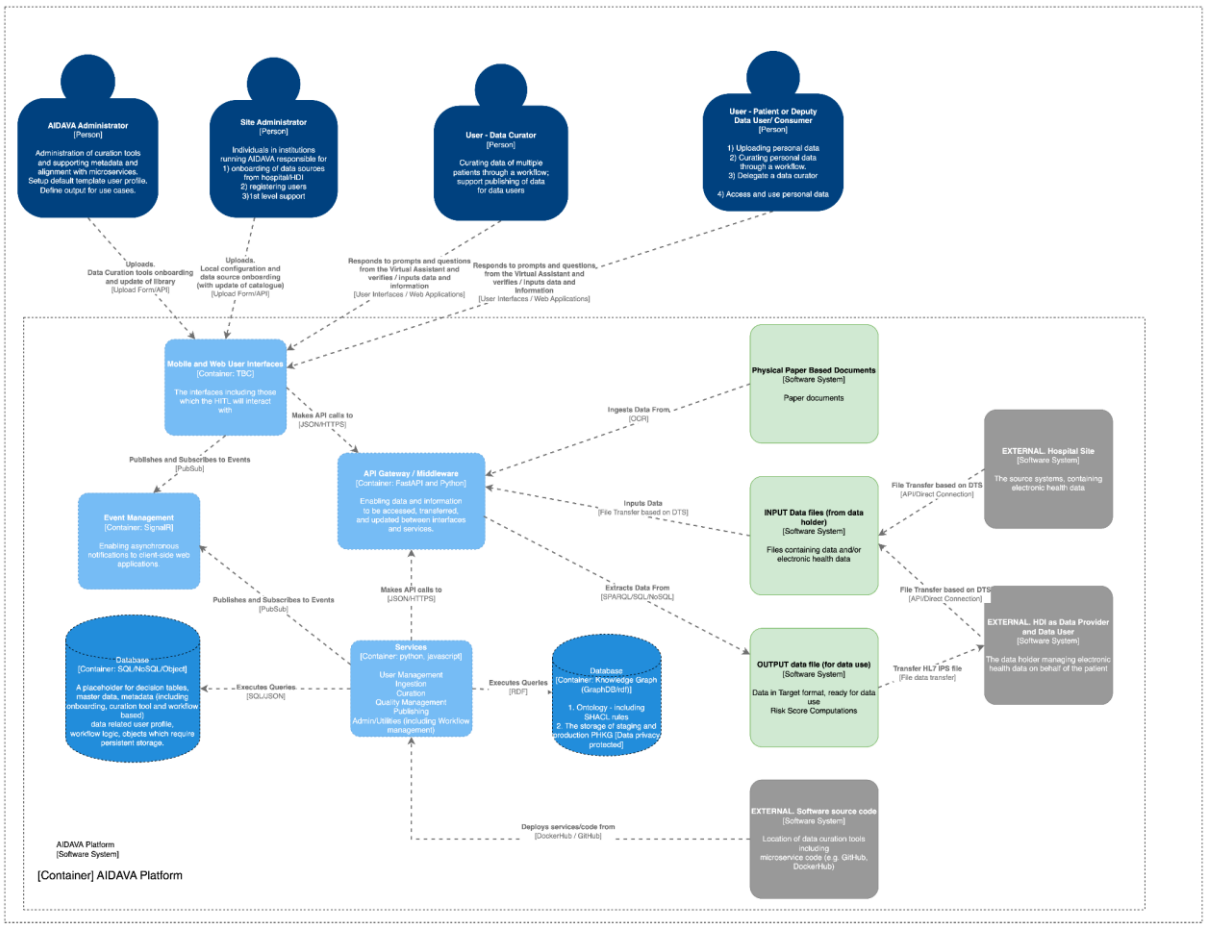


Figure 16. AIDAVA – C4 Container level

AIDAVA will follow a microservice architecture integrated on the publish/subscribe principle.

- API Gateway: The API gateway acts as the entry point for all external requests into the microservice architecture. It provides a single point of entry for all external clients, and routes requests to the appropriate microservice(s). The API Gateway can also perform authentication, rate limiting, and other security-related tasks.
- Event Manager: The event manager is responsible for managing the communication between microservices. It receives events from publishers, routes them to the appropriate subscribers, and ensures that events are delivered to all interested parties.
- Services: Services are the individual components of the microservice architecture. Each service performs a specific business function and can communicate with other services via the event

manager. In a pub/sub architecture, services can act as publishers, subscribers, or both. When a service publishes an event, other services that have subscribed to that event will receive it and can perform any necessary actions.
The services are responsible for their own data storage, retrieval, and management. Each service encapsulates its own data and provides an API for other services to access it. This means that each service can use the database technology that is best suited to its specific needs, rather than being constrained by a single technology that is used by all services.

- External systems will be simulated for the prototype that will access them through an API, with data transfer specification. In the prototype, the API will access files provided asynchronously; for the product, these APIs will need to be changed to allow direct access to the data source

7. Deployment, Maintenance and Support

This section describes the deployment of the prototype, which is strictly constrained by a study research protocol. A full product, running in real life, will have to be more flexible and should be more tightly integrated with data holders and data users and offer a wider support(24/7 instead of 8/5). **Potential deployment models for the product will be discussed in Deliverable 2.4 after evaluation G1.**

7.1 Deployment & configuration of the system

Deployment of a prototype solution like AIDAVA – managing sensitive personal data – must be done in a strictly controlled way, to ensure data protection compliance, including the following steps.

7.1.1. Preparation – before deploying the prototype

This phase is being managed within Task 1.4. It is provided here for information.

Step	Description	Coordination with local sites
Research Protocol	Definition of the **Study Assessment Protocol** (D1.4) with the clinical teams, to be used as the basis of the submission to each local Ethical Committee	NEMC
Data Privacy checks	● For the patient, and included as annex to the Study Assessment protocol: **Informed Consent Form** and S**tudy Information Package** validated by the AIDAVA Project Data Protection Officer (DPO – Partner IHD) and translated in the local language for each site ● **Data Privacy Impact Assessment (DPIA) –** to be performed in each site – on the basis of an information governance checklist provided by the AIDAVA DPO (Deliverable 4.4)	NEMC
Data Sharing Agreement (DSA)	Formal contractual agreement between the different parties to share data; a template contract is proposed by the AIDAVA DPO as a basis to be used in each site. The DSA includes a Data Transfer Specification (DTS) which captures detailed information on data sources to be curated; this information is used as the basis of the data source onboarding process. The DSA with the DTS needs to be provided in each site (hospital and HDI) – see integration aspects in Section 4.4.	b!lo
Ethical Committee Approval	Each site will have to submit the Study Assessment Protocol and the DPIA to the local Ethical Committee for approval before starting the actual evaluation. It is expected that this takes between 4 to 6 months and should therefore be done well in advance (i.e. started in Feb 2024 for an evaluation foreseen in Aug 2024)	NEMC
Training	Organisation of training sessions for the different types of users (before and during evaluation)	NEMC

7.1.1. Deployment of the AIDAVA prototype

Step	Description	Lead
Testing and Validation (GND)	Software must be tested and validated within GND and must pass the acceptance threshold	GND
Preparation of Documentation	Details on ● Infrastructure description ● Guidelines for installation and configuration – and local testing	GND
Build and Release	Software is built and packaged into a release package, which includes all the necessary files and artefacts needed for installation. The release package may be stored in a repository, such as a version control system or artefact repository.	GND
Environment Setup	Target environment must be set up with the necessary infrastructure and resources to support the software. This to include setting up the hardware, network connectivity, and software dependencies This includes clarification of the local servers for ● Raw data transferred from hospitals and HDI ● Curated PHKG (Staging and Curated areas) ● Published output for consumption for BC query, CVD score ● Published output for transfer to HDI	Local Admin
Local Configuration	● Onboarding of different components: ○ Data Source Catalogue ○ Master Data Repository ○ User Directory with users identified for the evaluation (patients should be added as they are recruited) ● Setup of data transfer of patient (HL7 IPS) data back to the HDI for further use and visualisation	Local Admin
Testing and Validation	The software must be tested and validated in pre-production environments locally to ensure that it works as expected and does not cause any issues or conflicts with existing systems.	Local Admin

7.2 Customer support

Level	WHO	WHAT	HOW
L1	Site Admin	Initial level of support provided to users locally – taking into account local constraints and data privacy. It typically involves basic troubleshooting and issue resolution. The goal is to resolve customer issues as quickly and efficiently as possible ● Answer common questions, provide instructions, or direct customers to self-help resources. ● **User management:** add, update or delete user profiles ○ Register new user – including personal address – with role and respective access rights; allocate users to roles ○ Take care of users who forgot their password ● Log request in Issue Log when not possible to solve the	● Personal email, chat, or phone ● Local language ● 8 to 5 local time, working days Input in Issue Log, required when transferring to L2

Level	WHO	WHAT	HOW
		problem directly, and follow up on resolution Note: to send an email from the local site, the AIDAVA local system needs to have access to the local email server (or to have a local account from the hospital email system) ● Transfer issue to level 2 when not possible to solve locally	
L2	GND	Involves more complex issue resolution and may require specialised technical knowledge or expertise. Level 2 support is usually provided by a specialised team, such as a technical support team or an engineering team, and may involve more in-depth troubleshooting or debugging Note: need access to the local version: console level access / IP Secure access with admin rights for maintenance purposes => GND will be an agreed stakeholder (as defined in the Study Protocol)	● Email and GND phone (remote support with local access) ● English ● Working days – 8 to 5 local times (CET and CET + 1) = 8 to 6 (CET+1) for GND (NOT 24/7)
L3	GND	Involves the most complex and critical issue resolution, such as bug fixes or product enhancements. Level 3 support is usually provided by the product development team or senior technical experts and may involve code changes or product updates.	
L2	HDI	For issues related to access to personal records (IPS) sent to the HDI	● Email ● English ● Working days – 8 to 5 local times (CET and CET + 1)

Function	Description	Value (min-max)
Issue Log	Centralised Service management system to be provided by GND to register and track all the issues	GND will provide a central issue tracking system to collect and supervise the issues reported during the evaluation and feedback phase. Also we will provide an issue collector page into the system, which will let the users create issue tickets directly into the tracking system.
Time to resolution	Time it takes between the first entry of an issue in the log file and the resolution	4 hours in average (for the tools developed by GND, no engagement for tools developed by other parties such as curation tools)

8. Market considerations for the AIDAVA "product"

The section introduces market considerations for the AIDAVA virtual assistant product, to be certified under the Medical Device Regulation. A more in-depth analysis will be developed based on the results of the G1 evaluation and the input from the Sustainability Advisory Group - to be initiated mid 2023 – and documented in *Deliverable D2.4 Update to Solution Design Document for G2* at the end of the project.

8.1 Approach to market: mass market or target market?

The first question to ask when developing a product is to identify if this should focus on a mass market (as wide an audience as possible) or a target market with specific needs. While centred around the patient/citizens, AIDAVA is a product aiming to solve data interoperability and data quality issues i.e a very specialised topic which points to a targeted approach, for the following reasons.

- The audience for this type of software is likely to be relatively small and specialised, i.e. organisations managing data (on behalf of, with and for the patient like Health Data Intermediaries (HDI), though not necessarily limited to this). Attempting to market the software to a mass market would result in a lot of wasted effort and resources, as the majority of the general population would not understand the value of the software, without additional training and services.
- A targeted approach allows for more precise targeting of marketing channels that are most likely to be effective for the specific audience. For example, if the audience is primarily made up of professionals in a particular public health domain or in the hospital sector, targeting industry-specific conferences or trade publications would likely be more effective than generic advertising channels. In addition, if successful, the approach developed in AIDAVA could be applied – against target customers – in other sectors, not directly related to healthcare, and we could more easily reach these other sectors through a targeted approach.
- Focusing on a targeted approach can also help to establish the software as a specialised and highly valuable solution within the industry. This can help to differentiate the software from other generic solutions on data quality enhancement and position it as the go-to solution for organisations looking to enhance their data quality and interoperability.

8.2 Potential customers

This section provides an initial analysis of potential customers within healthcare. In Deliverable D2.4 we will further analyse these different customers, including an assessment of market readiness and size, and expand the analysis for potential outside of healthcare.

The AIDAVA virtual assistant as developed in the project should be of interest to any healthcare organisation that needs to integrate and curate heterogenous personal health data (PHD) and wants to reuse these data for different purposes in clinical care as well as in clinical research.

1. **Hospitals** running a Hospital Information System composed of multiple different, often non-integrated systems, across departments. They would benefit from deploying and running a system like AIDAVA to integrate and curate data coming from these different systems at different levels.

- First and foremost by providing to their own physician an integrated dossier of the patient. Currently physicians have to look across systems to build a view of a patient. This can be time consuming and increases the burden of already overloaded clinical staff.
- By supporting local clinical research by automatic extraction and transformation of relevant data for specific clinical registries, maintained by scientific staff for their own research. Different departments maintain different registries for different therapeutic areas; they may curate - most often manually - the same data of the same patient multiple times (and inconsistently). With AIDAVA, data will be curated once and reusable for all registries.
- By transforming data into different formats required to support different external users.
    i. Transfer of patient data across healthcare organisations (primary, secondary and tertiary care) in a specific format to support continuity of care (e.g. HL7 V2.5 or HL7 CDA).
    ii. Transformation and transfer in HL7 FHIR (e.g. IPS) to meet regulatory requirements (in alignment with EHDS) and ensure continuity of care across borders.
    iii. Transfer to regional/national healthcare authorities such as National Contact Points for Digital Health (NCPDH) as described below; this may become more stringent with the emerging EHDS regulation
    iv. Transfer to research organisation/ pharmaceutical companies in CDISC format as part of clinical trial (electronic data capture) or in OMOP format to support development of so-called "real world data sets".

2. **Software vendors** that serve healthcare organisations, and mainly EHR vendors who could benefit from the functionalities of AIDAVA in their overall solution. They would first benefit from a most effective generation of HL7 FHIR content. They would also benefit by supporting more effectively the increasing demand of hospitals to generate data sets in different formats for clinical research as described above: (1) clinical registries for internal research in proprietary format; (2) patient data in CDISC CDASH format for agreed clinical trials, or (3) (anonymized) real-world datasets for commercial partners such as pharma companies, and software companies developing AI/Machine Learning models for the healthcare sector.

3. **"Patient Hub" or National Contact Points for Digital Health (NCPDH)** introduced in the EHDS – and building on infrastructure already in place in several EU countries (e.g. Health Data Hub in France, eHealth.gov in Belgium, HELGA in Austria, TEHIK in Estonia,..) – are centralising health data from individuals, across multiple data sources, including healthcare organisations such as hospitals, GPs, nursing homes, vaccination centres but also research infrastructure such as biobanks/genome DBs, and clinical trial data databases, and increasingly data collected through personal apps and medical devices.
While there is a major value of "pooling" all patient data in such a central hub, this data has limited value as it is in many different formats, not interoperable and therefore not reusable for further analytics without time-consuming curation. Within the EHDS regulation, it is proposed that other organisations called **Health Data Access Bodies (HDAB)** will process and curate data based on a request by a data user, for specific research or policy making purposes.

It can be expected that different data users will require different standards formats (OMOP, CDISC, HL&, ..) requiring separate cleaning and transformation from the data sources. It would be more effective to have a tool like AIDAVA that would curate the data into an interoperable and reusable format - the Personal Health Knowledge Graph - within the NCPDH and then transfer the needed information into the HDAB in the required format.

4. As identified during the use cases analysis, **Health Data Intermediaries (HDI)** are emerging actors supporting the patients in managing and controlling their data. In the context of the prototype, it was decided – for security, data privacy issues and ethical consideration – that the prototype managing sensitive personal data will function only within the Hospital environment to avoid that any data collected by Healthcare providers would leave the hospital. In the long run however, we envision that AIDAVA might run within HDI so that patients can truly manage and curate, if needed, all their personal health data from different sources - including hospital data - through a single actor. AIDAVA would benefit the HDI at different levels.
    - Provide better direct services to their patients, including tools supporting visualisation of the health record based on integrated longitudinal data.
    - Provide new services to their patients, allowing them to valorize their data for the common good, by sharing these data for clinical research of public interest, but also enabling the patient to share their data with commercial organisations - like pharma companies - and receive some benefit out of their data.
    - Building on the previous point, AIDAVA would allow HDIs to become highly effective private National Contact Points for Digital Health, under the condition that they are certified by the national governments.

5. **Pharmaceutical companies** (or the **Clinical Research Organisations** supporting them) could also benefit from the AIDAVA virtual assistant, for clinical trial management and for internal purposes (reporting, regulatory audit/query and/or for internal AI/analytics). We can envision the following uses.
    - Development of real world data sets across multiple organisations and countries, based on heterogeneous data. Today this is a time consuming, often manual process.
    - Answers query from regulators on the safety aspect of a drug. This may require analysing the data collected - with different proprietary standards - more than 10 years ago; it can take several months to transform these legacy data into an interoperable data set. The worst case known to one of the authors of this document was 6 MUSD, 6 months for one single query from the authorities.
    - Pharma companies realise that their (legacy) clinical trial data – and related information contained in an electronic Trial Master File (eTMF) – is a major asset for clinical research, not usable currently because of its disparity. Some pharma companies have started data “transformation” projects to solve this, curating all their clinical trials over the years. While large companies have the bandwidth to develop their own solution, smaller companies will not, and would heavily benefit from a solution like the AIDAVA Virtual assistant.

9. Conclusion and next steps

Together with the deliverables on personas (D1.2) and on business requirements (D1.3), this document will be used as the basis of the Technical Architecture (D3.1) and further support development in WP3.

After evaluation of G1 of the prototype, we intend to revise this deliverable – to produce D2.4 – on the following aspects.

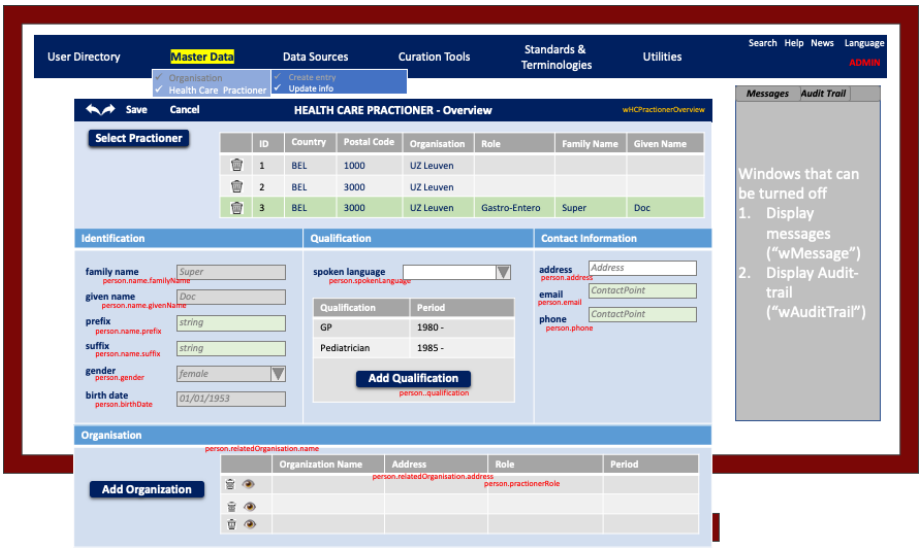
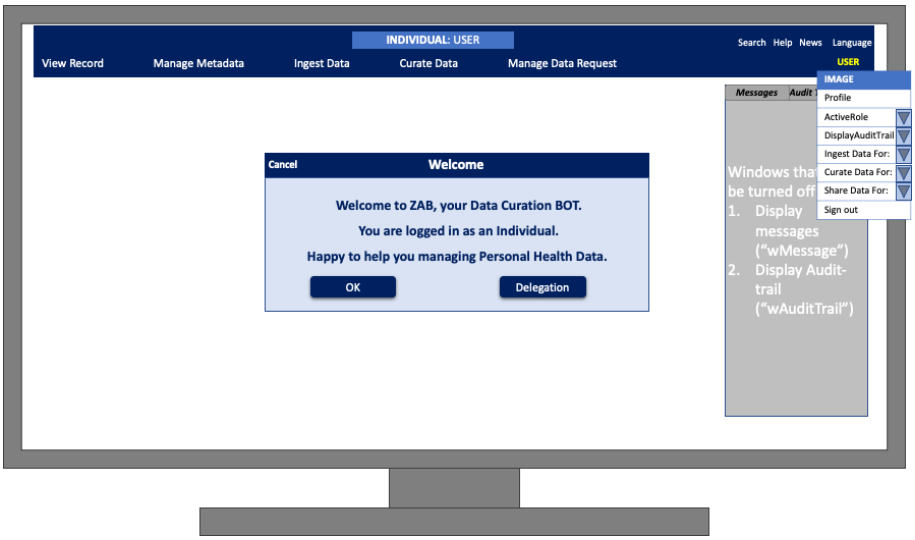
- Impact on the solution architecture of the Human-in-the-Loop (Section 5) solution being developed in Task 5.3.
- Further exploration of market approach, potential and readiness. This later topic will be the core focus of the revision, building on results from the evaluation of G1 and subject to discussion and validation with the Sustainability Advisory Board (SAB).

Annexes

A.1 Wireframe extracts

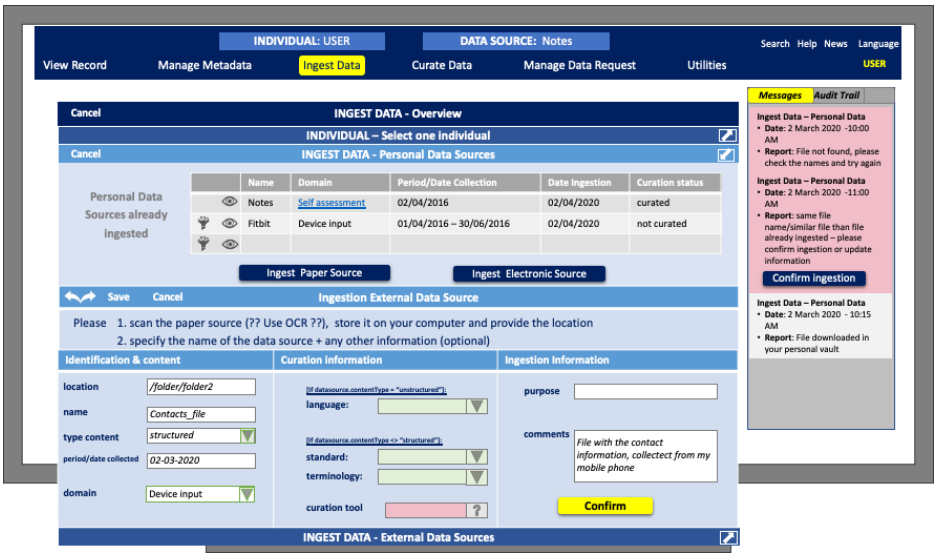
The wireframes below are derived from background information from Partner b!loba; the wireframes will be further used during the development to discuss human interaction aspects.

A.1.1 Web-related

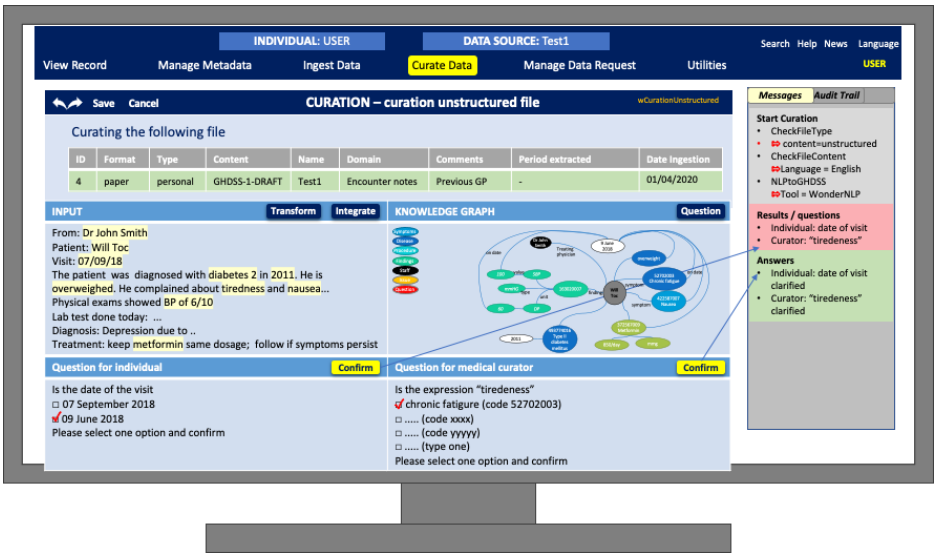
<table>
<tr>
<td> 
</td>
<td>Colour legend of attributes used across the wireframes</td>
</tr>
<tr>
<td> 
</td>
<td>Admin screen to enter Master Data (Healthcare Providers). The element in red indicates the name of the attribute in the Data Architecture</td>
</tr>
<tr>
<td> 
</td>
<td>Welcome screen for individuals. Please note the different roles the user can assume for a patient. (A data curator would not have access to "Ingest data").

The window on the right is also used to display information during the execution of different tasks.</td>
</tr>
</table>

Screen for users to display the content of the data source (raw format and PHKG). At the source level, a KG is still limited. It is expected that the full PHKG will not be visualised as such (too large)

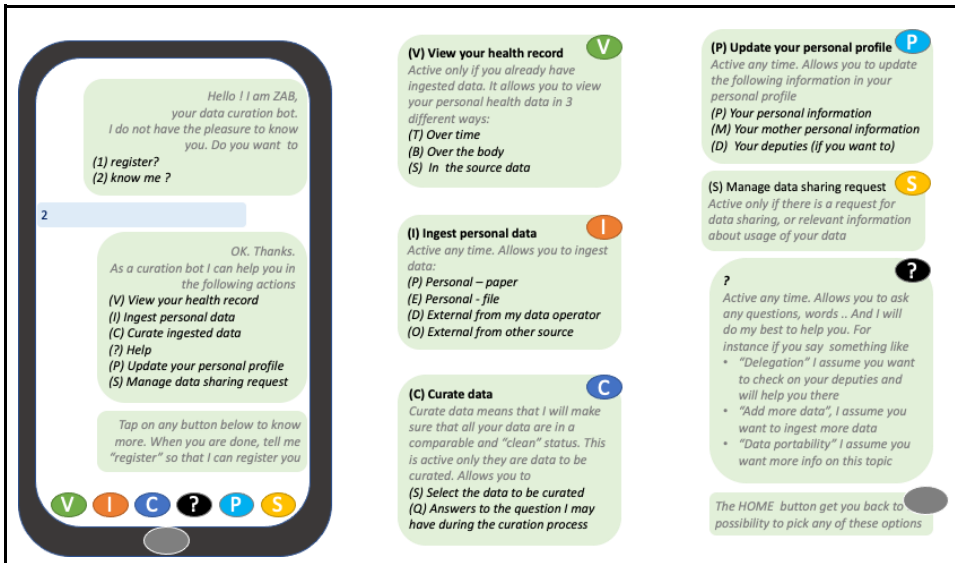


Screen for patient to ingest their data (either paper data with OCR, or data available in the data source catalogue)
Please note messages on the right-hand side.

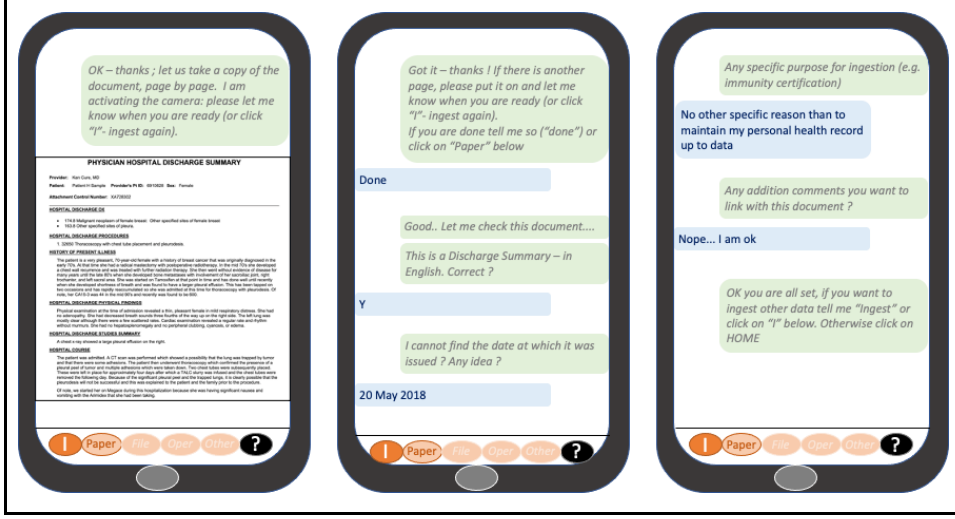


Screen for the patient to curate their data – selecting one data source that is then transformed. A question is asked to the user to clarify semantics.
Please note messages on the right-hand side.

A.1.2 Smartphone-related



Explanation of the different actions that can be provided through the smartphone app



Example of a dialogue to insert a paper document , with addition of information requested by the system