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

AIDAVA

AI powered Data Curation & Publishing Virtual Assistant

Deliverable No. 4.2

Data Management Plan (DMP)

Contractual Submission Date: 28/02/2023

Actual Submission Date: 28/02/2023

Responsible partner: P6: IHD

Funded by
the European Union

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

Deliverable number	**D4.2**
Deliverable title	**Data Management Plan**
Type[1]	OTHER
Dissemination level[2]	PU
Work package number	WP4
Work package leader	P6-IHD
Author(s)	Nathan Lea (P6-IHD) Isabelle de Zegher (P2-b!lo)
Keywords	Data Management Plan, FAIR, Open Data, Security, Confidentiality, Governance, Compliance

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
V1.0	28/02/2023	First submission

[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

List of Abbreviations and definitions

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

Summary

This Deliverable provides the Data Management Plan (DMP) for AIDAVA. It is based on the European Commission Template for Horizon 2020 projects available at https://ec.europa.eu/research/participants/data/ref/h2020/gm/reporting/h2020-tpl-oa-data-mgt-plan_en.docx. AIDAVA has populated this Data Management Plan in line with recommended EC guidelines. It will be updated as the project proceeds.

A DMP is an important component of any data intensive programme because it imposes a need for balance between protection of data, success of the programme and the potential for reuse of data. AIDAVA is unique as a project because the primary data handling is focused on data ingestion and curation as a tool to assist citizens in managing their own health data.

The approach to developing the data management plan has included workshop discussions with partners at the October 2022 Kick Off Meeting in Maastricht and a dedicated data flow workshop held in Tallinn in December 2023.

The details gathered were compared with the proposal and obligations on the partners as described in the consortium agreement. They were also compared with the developing Research Protocols for both the Breast Cancer and Cardiovascular Disease (CVD) use cases developed in Task 1.4

The results of the details gathered are presented as the Data Management Plan in Section 3 of this Deliverable. It concludes with the next steps and specification of updates in time for M40's second version of the Data Management Plan.

1 Introduction

This Deliverable provides the Data Management Plan (DMP) for AIDAVA. It is based on the European Commission Template for Horizon 2020 projects available at https://ec.europa.eu/research/participants/data/ref/h2020/gm/reporting/h2020-tpl-oa-data-mgt-plan_en.docx. AIDAVA has populated this Data Management Plan in line with recommended EC guidelines. It will be updated as the project proceeds.

The purpose of the Data Management Plan is to provide a basis upon which data will be handled and used to achieve the purposes of AIDAVA as well as ensure that maximum benefit can be derived from the project, including the adherence to the Findable, Accessible, Interoperable and Reusable (FAIR) Principles, Open Science agenda and informational self-determination for the individual citizen.

This must be balanced with the need to ensure that data relating to citizens, especially their health records, is protected in line with regulatory requirements, professional duties of confidence and the reasonable expectations of the citizens to have their personal data protected whilst also receiving the highest possible standards of care.

The DMP therefore provides a basis of how AIDAVA plans to achieve this balance and will contain details around AIDAVA's purposes, the data that will be required, how it will be used, how it will be protected and how it might be shared more widely where permissible. The DMP forms part of the overall governance and risk assessment framework aligned with approaches defined in the General Data Protection Regulation (GDPR) and wider research governance processes, including development of research protocols and receiving appropriate Research Ethics Committee (REC) approvals.

This Deliverable forms the basis of the initial DMP and will be updated later in the project as a separate Deliverable.

2 Description of Activities

2.1 Reasons for a DMP

A DMP is an important component of any data intensive programme because it imposes a need for balance between protection of data, success of the programme and the potential for reuse of data. AIDAVA is unique as a project because the primary data handling is focused on data ingestion and curation as a tool to assist citizens in managing their own health data. Their data are being transformed into an interoperable and reusable Personal Health Knowledge Graphs (PHKG). The data handling is for the benefit of the patient and treating physicians in clinical care, as well as for researchers, innovators and decision-makers in secondary data use. While the end goal is reusability, the scope of AIDAVA is first and foremost on data quality (there is limited value in sharing low quality data) to support the "curate once, use many times principle". To demonstrate this principle the project is evaluating 3 cases of reuse: 1) delivery to a clinical registry in support of clinical research in BC, 2) computing a CVD risk score for treating physicians in clinical care, 3) display the IPS to the patient.

The DMP further addresses adherence to data protection and security regulations as well as addressing concerns around security. These in turn relate to the need to ensure that data reuse is part of the respect for individual autonomy and privacy for participants. Whilst the DMP alone cannot provide assurance that these are addressed, it does provide a basis to inform the activities that do, including the Data Protection by Design and Default and Data Management Plan approaches. These will be documented under other deliverables.

2.2 Approach

The approach to developing the data management plan has included workshop discussions with partners at the October 2022 Kick Off Meeting in Maastricht and a dedicated data flow workshop held in Tallinn in December 2023. These workshops allowed AIDAVA to gather plans and intentions around data handling and start to explore some of the particulars around how this would be achieved.

The details gathered were compared with the proposal and obligations on the partners as described in the consortium agreement. They were also compared with the developing Research Protocols for both the Breast Cancer and Cardiovascular Disease (CVD) use cases developed in Task 1.4

The findings and conclusions are presented in Section 3.

3 Results & Discussion

3.1 Data summary

Integrated, high-quality personal health data (PHD) represents a potential wealth of knowledge for healthcare systems, but there is no reliable conduit for this data to become interoperable, AI-ready and reuse-ready at scale across institutions, at national and EU level.

AIDAVA will fill this gap by prototyping and testing an AI-powered, virtual assistant maximising automation of data curation & publishing of computable knowledge derived from unstructured and structured, heterogeneous data. The assistant includes a backend with a library of AI-based data curation tools and a frontend based on human-AI interaction modules that will help users when automation is not possible, while adapting to users' preferences. Two versions of this virtual assistant will be tested with hospitals and emerging health data intermediaries (HDIs), around breast cancer patient registries and longitudinal health records for cardio-vascular patients, in three languages.

The data in scope are described in detail in Deliverable D1.1 [ref] under Section 4.1 and 4.2 (subsection data source) for each of the use cases. This includes

- Medical Records from GP and mainly from hospitals, including structured and unstructured / narratives and diagnostic images, but excluding omics data;
- Patient held data from personal apps and/or medical devices

The data will be collated and curated into a Personal Health Knowledge Graph (PHKG), containing an interoperable patient longitudinal health record. The data curation process will be automated as much as possible through the orchestration of multiple tools, AI-based tools and others; whenever automation is not possible - by lack of semantic - a "human in the loop" module will assess if additional information should be requested to the patient or to an expert curator.

The curated PHKG can be reused for many different purposes. In the project, we will focus on

- extracts contributing to an "EU'' federated breast cancer registry that supports analytics across different sites - without transferring the data (use case 1),
- extracts to compute a cardiovascular risk score supporting monitoring of CVD patient (use case 2)
- Extracts - composed by the international patient summary (IPS) with visualisation of the data (both use cases).

AIDAVA will develop 2 generations of virtual assistants as a prototype and will perform a formal assessment study as described in D1.4 to evaluate whether the prototype behaves as expected based on clear criteria and assess whether data are reusable. It will also provide a foundation for eventual Medical Device Regulation certification. During the assessment study, the recruitment is limited to a small cohort of participants who can be consented to enrol in the study if they wish. The assessment of the virtual assistant will occur in parallel with usual care and will not be used to impact care (in case participation reveals any findings of clinical significance, the treating physicians will be directly informed to take actions independently of the tool).

For these reasons, AIDAVA does not anticipate that the data gathered will be of utility outside of the assessment study beyond any utility that participants may find for data that will be held by the HDI's.

In this DMP we address these points.

Aspect	Response/explanation
Purpose of the data collection/ generation and its relation to the objectives of the project	Most of the data will be retrospective; only personal app and medical device app will be collected prospectively All data - retrospective and prospective - will be used to test the performance and acceptability of the prototype Virtual Assistant to transform dispersed and heterogeneous data source into an interoperable and reusable, Personal Health Knowledge Graph that can be reused for multiple purposes in clinical care (building a breast cancer registry) and clinical research (monitoring risk in CVD patients)
Types and formats of data generated or collected by the project	Prospective (to be confirmed): patient fitness and activity apps, and those supporting smart devices. These data will be managed by Health Data Intermediaries and ingested in the AIDAVA virtual assistant following a Data Transfer Agreement with the relevant HDI and a formal onboarding of the data sources, documented with metadata in a Data Source Catalogue.
Any re-use of existing data and how this will be done	Retrospective data: Medical records from primary and secondary care, all formats, including imaging and items excluding omics. These data will be managed by ingesting in the AIDAVA virtual assistant following a Data Transfer Agreement with the hospital and GP Practices, and a formal onboarding of the data sources, documented with metadata in a Data Source Catalogue.
The origin of such data	Care provision, registry participation and personally collected data by the participants
Expected size of the data	The "data" in scope of AIDAVA includes potentially the whole electronic medical record - without images themselves (just the DICOM metadata) and without omics information - for breast cancer and CVD patients; there will be an expected 90 patients (45 breast cancer, 45 CVD). The size of this record (including hospital stays, hospital visits, GP follow-up, device and app data.) will be dependent upon the age of the patient, the length of time of their medical conditions and the seriousness of their condition. At this point it is difficult to provide an appropriate estimate.
Likely users of the data	Users of the systems ● Patient themselves (undertaking and validating the curation), recruited for assessing the prototype; they will be trained ● Community Curators - i.e., Patient Deputies ● Data Curators within the hospital sites who are responsible today for - often - manual curation Users of the curated data (Data users in the sense of DGA) ● Breast cancer specialist

Aspect	Response/explanation
	● CVD treating physician ● Patient themselves (visualisation),

Table 1: Data Summary

3.2 FAIR data

The main purpose of AIDAVA is to build a medical device prototype that will take personal health data from heterogeneous data sources and make them Interoperable and Reusable.
When speaking about data in the context of AIDAVA, we speak about 2 types of data

1. The data sources (DS) - which include the raw data from the patient - coming from the hospital and - through the Health Data Intermediary - from GP, personal app and medical devices
2. The Personal Health Knowledge Graph (PHKG) which is generated from the data sources by the prototype medical device, into a curated, interoperable format. The PHKG can in turn be transformed into different target formats to support multiple different uses.

In the table below we make the difference between the 2 for each question

3.2.1 Making data findable, including provisions for metadata

Aspect	Response/explanation
Are the data produced and/or used in the project discoverable, identifiable and locatable by means of a standard identification mechanism	● DS: no, as the project is working with retrospective data ● PHKG: not in scope but should be considered as the project proceeds
What standard identification mechanism used (e.g., persistent and unique identifiers such as Digital Object Identifiers)	● DS: no, as the project is working with retrospective data ● PHKG: not in scope but should be considered as the project proceeds where https://w3id.org/ may be a candidate for use.
Is meta-data available through catalogue?	● DS: yes, the project will develop a data catalogue for the data sources that will be curated into the PHKG ● PHKG: not in scope but should be considered as the project proceeds
Can meta-data be used for search?	Yes

Aspect	Response/explanation
Naming conventions used	This will be agreed upon later in the project (where the internal Consortium processing may require a consistent naming convention as clinical variables are agreed upon)
Clear versioning supported?	● DS: not specifically in the project as they are retrospective data ● PHKG: not in scope but should be considered as the project proceeds
Additional keyword search supported?	It will be.
What metadata will be created using which standards?	● DS metadata: DCAT-AP is being considered as an initial standard for metadata description; it will need to be enriched ● PHKG: considering using a knowledge graph of metadata

Table 2: Making Data FAIR

3.2.2 Making data openly accessible

Aspect	Response/explanation
Will data be made openly available as the default?	● DS: No (out of scope) ● PHKG: The data will only be useful for AIDAVA purposes. Before considering using outside of AIDAVA for clinical care and/ or clinical research we will need to assess the final data quality and validate the device following MDR. Note that the entire premise of AIDAVA is to apply the FAIR principles so that individual citizens can ensure their data is Findable, Accessible, Interoperable and Reusable for them but can also join in making it FAIR for others, and always within their control.
Which datasets will NOT be openly available and why?	● DS: No (out of scope). ● PHKG: will not be made openly accessible as mentioned above
How will the data & meta-data be made accessible (e.g., by deposition in stated repository)?	● DS: The cataloguing will follow standards for internal use. However, use of these data will be strictly limited by the Data Sharing agreement ● PHKG: not considered (see above)
If known repository, what arrangements are explored?	Not applicable
If project-specific access, then:	In line with the particulars of the Consortium wide Data Protection Impact Assessment and where appropriate, its Data Sharing Transfer and Data Licensing Agreements.,

Aspect	Response/explanation
Data Access Committee	None planned, but an Ethical Advisory Board is being appointed and the project has Patient Association Organisations and Patient Consultants as equal partners.
Any conditions for access (i.e., a machine-readable license)	Likely N/A
What methods or software tools will be needed to access the data?	● DS: extract from data source and transfer to secure server (as specified in Data Sharing Agreement). ● PHKG: extract published in specific format/standard though the AIDAVA publishing module
Documentation for software	This will be developed in line with requirements for MDR certification.
Availability of software	Likely accessible by GitLab or GitHub internally.
Institution and researcher vetting process/procedures - describe	In line with the particulars of the Data Protection Impact Assessment where in the first instance data sharing will be assessed by the Data Sources in line with their existing requirements and thereafter once the data is assembled centrally a decision can be made regarding which group can vet.

Table 3: Making Data Openly Accessible

3.2.3 Making data interoperable

Aspect	Response/explanation
Are the data produced in the project interoperable	● DS: will be use as such (heterogenous, not standardised) ● PHKG: yes, all PHKGs will be an instance of a reference KG which will be based on HL7 FHIR resource profiles, SNOMED, LOINC, ATC…
If not, explain why not	N/A
Data and metadata vocabularies, standards or methodologies used	HL7 FHIR resource profile, SNOMED CT, LOINC, ATC, (others as needed)
Standard vocabularies used	See above
Mappings from uncommon or project-specific ontologies or vocabularies to more commonly used ontologies	We do not anticipate the need to adopt uncommon ontologies or vocabularies; however, if needed to complement SNOMED for instance we may consider to add some therapeutic areas specific terminologies

Table 4: Making Data Interoperable

3.2.4 Increase data reuse (through clarifying licensing)

Aspect	Response/explanation
Will data be available for onward data-sharing/re-use?	● DS: no ● PHKG: yes - the end purpose of AIDAVA is to ***curate data, reuse many times.*** In the context of the prototype there will be 3 type of extract ○ Mimicking an EU breast cancer registry ○ Input for assessment CVD risk score ○ Support the International Patient Summary (IPS) for patient visualisation
Approach to data licensing for onward use	N/A as the PHKG from the patient are resulting from an invalidated/ uncertified prototype. The Project will explore the possibility of keeping the data in the control of the participant at its completion as well.
Likely date for data availability for onward use	N/A as the PHKG from the patient are resulting from an invalidated/ uncertified prototype
Explain any restriction on date of availability	N/A
Possible restrictions on onward data-sharing	Data is unlikely to be shared onward, as the PHKG from the patient is resulting from an invalidated/ uncertified prototype. The end objective however of AIDAVA is to ensure high quality data so that the patient can share high quality data
Data retention policy (including availability for data-sharing)	Will follow data retention policies as determined at sites. Evaluation/open access data will be retained according to Consortium feasibility checks and wider EC guidance. **NB: as the PHKG data are not for further used after then end of the project; AIDAVA will need to agree whether they should be kept or deleted after the end of the project depending on regulatory requirements or any Medical Device Certification, or where the participants may be able to retain the data at the end of the study.**
Description of data quality assurance processes	*Task 4.2 will develop statistical assessment methodologies and tools to define metrics, to help flag up issues of bias in the data quality enhancement and FAIRificiation process. This will assure that the PHKG will be based on high-quality data.* For the PHKG: A dedicated task within T4.2 is assigned to assure data quality using specific methods developed by partner P6-IHD to assess and label the data from a quality perspective

Table 5: Increase Data Re-use (through clarifying licences)

3.2.5 Allocation of resources

Aspect	Response/explanation

Estimated project costs for making data FAIR	● DS: N/A ● PHKG: to be estimated in terms of workload as part of the assessment
Data management responsibility across the project	Each clinical site/ hospital will take lead responsibility for the data managed within their hospital - including the resulting PHKG managed within their site. They will liaise closely with participants and HDIs in discharging this responsibility, where the responsibilities of participants and HDIs will be established in the sharing agreements. The developer of the curation tool will support deployment and use of the prototype; they will be responsible for ensuring secure processing not for the content of the data.
Resources required for long-term preservation (costs and potential value, who decides and how what data will be kept and for how long)	To be determined later in the project, under the responsibility of the coordinator as the data cannot be reused, they will most probably be discarded.

Table 6: Allocation of Resources

3.2.6 Data security

Aspect	Response/explanation
Data security measures used (including data recovery as well as secure storage and transfer of sensitive data)	● DS: data transferred by DTA will be managed within a Docker instance deployed in each clinical site and under their local responsibility ● PHKG: as above. A number of information governance and data protection and information security instruments are being used in this and forthcoming project deliverables.
Where data will safely be stored (in certified repositories for long-term preservation and curation). Provide detail	● DS: data transferred by DTA will be managed within a Docker instance deployed in each clinical site ● PHKG: dito For the existing modelling and risk stratification work within the Consortium, the security and certification requirements are being assessed as part of the processes described in the DPIA

Table 7: Data Security

3.2.7 Ethical aspects

Aspect	Response/explanation
Any ethical or legal issues that can have an impact on data sharing	Independent ethical oversight will be achieved via Independent Review Boards for each of the Clinical Partner sites where participants will be recruited. An EAB will also be appointed.
References to ethics deliverables and ethics chapter in the Description of the Action (DoA) – if relevant	WP4
Questionnaires dealing with personal data	Participants will be given user satisfaction questionnaires as part of the Assessment study
How is informed consent for data sharing and long-term preservation sought in such questionnaires?	This will be handled pursuant to Research Ethics Approval and sought at clinical sites according to an ethically approved standard operating procedure to be specified in the Study Assessment Protocol, approved by each local ethical committee

Table 8: Ethical Aspects

3.3 Information Asset Register

AIDAVA will keep an Information Asset Register in the form of a metadata catalogue as part of the curation process. This will account for each of the data holder partners' contributions, data flows and sample transfers across the two use cases. The Register will be achieved through the Data Source Catalogue in Task 3.4 (Deliverable 3.3) and the requirements established in Task 1.3.

The importance of the Information Assets Register is to be able to establish the types of data being used for the project and what each of the parties within the project are doing with those data. This could include being a data source providing data, a recipient processing data or indeed both.

This register helps to link data items, activities and responsibilities to the agreements, approvals and oversight requirements (including entry into the DPIA). This way AIDAVA can account for not only the continued update of proposed data flows but also isolate where there may be delays or uncertainties around data transfers. This is also important for Machine Learning and VA validation, including the requisite Data Quality checks.

4 Conclusion

AIDAVA is a unique project whose primary goal is to empower citizens to be able to access and manage - and increase the quality of - their own health records and support their care teams in managing their conditions. It seeks to give citizens unprecedented control over their health information and use it as they see fit.

In evaluating the approach this DMP is a first attempt to balance the unique nature of the data processing with adherence to high standards of data protection whilst honouring the FAIR Principles, supporting Open Science where possible but also to enhance the data utility for citizens within the Health Data Intermediary model.

In order to achieve this, the approach that AIDAVA is taking to governance is reflecting the need to support the data use within the scope of the assessment study and the extent to which it empowers citizens. This DMP outlines how AIDAVA will:

- Ensure that the assessment of the prototypes is appropriately risk assessed and governed according to contract between the data handling parties and service providers;
- Enabling the participants in the assessment of the prototypes to control the curation of their health data and its reuse where this will be assured via review of the assessment protocol established in D1.4;
- Assessing how data generated by the project, including the Personal Health Knowledge Graphs, might be made available for further reuse at the behest of the participants and the Project with respect to the FAIR Principles and OpenData.

The DMP will be updated in M40, and its proposed approaches will be assessed in line with the protocols of the studies that are being defined under D1.2 and 1.4. It will also align with the requirements establishment in Task 1.3 and align Information Asset Register requirements with Task 3.3.

No issues are foreseen with the Data Management Plan and their execution, though the compliance work as part of the rest of Task 4.1 will help ensure this will be the case.

5 Next steps

This deliverable will be reviewed periodically, and an update as part of a forthcoming Deliverable is due in M40. The regulatory compliance measures as part of the rest of Task 4.1 will align with the design and development of the AIDAVA tooling throughout the other work packages. These include continuation of the DPIA and identification of the requisite data sharing agreements and any appropriate codes of conduct.

As the tooling is developed, it will be trialled in line with the protocols being developed as part of WP1 for each use case. The regulatory compliance requirements as part of Task 4.1 will continue to oversee the protocol development process for the trials where the particulars of the Data Management Plan will be made clear throughout the assessment study protocol, the study information package/patient information leaflet and associated consent forms for participants.

6 References (if applicable)

N/A

7 Annexes (if applicable)

N/A