



Safe and Explainable
Critical Embedded Systems based on AI

PhDMG0001 Data Management Guideline

Version 2.0

Documentation Information

Contract Number	101069595
Project Website	www.safexplain.eu
Contractual Deadline	31.03.2024
Dissemination Level	SEN
Nature	R
Author	Javier Fernández
Modified by	Lorea Belategi
Reviewed by	Irune Agirre
Approved by	Irune Agirre
Keywords	AI, Functional Safety, Data Management, Data Collection, Data Verification



This project has received funding from the European Union's Horizon Europe programme under grant agreement number 101069595.

Table of Contents

1	Review / Modification History	2
2	Objective	3
3	Scope.....	3
4	Introduction	3
5	Data Requirements Specifications	6
6	Data Collection.....	6
7	Data Preparation.....	7
8	Data Verification	7
9	Acronyms and Abbreviations	8

1 Review / Modification History

Version	Date	Description Change
V2.0	15/02/2024	Changes Applied as a result of TÜV Review 2024-01-19
V1.0	01/12/2023	First version after complete internal review
V0.4	20/11/2023	Modifications and improvements based on internal review
V0.3	01/09/2023	Modifications and improvements based on internal review
V0.2	30/06/2023	Modifications and improvements
V0.1	08/05/2023	First draft

Note: Since Artificial Intelligence - Functional Safety Management (AI-FSM) utilizes templates from both the traditional Functional Safety Management (FSM) and its own templates, this annex distinguishes the AI-FSM templates by color-coding them in orange and the traditional FSM templates in green. Additionally, the folders' names will be enclosed in quotation marks and the files' names created from the templates are written in italics and underlined. These files' names are preceded by "REF_", which should be changed to reflect the specific safety project reference.

2 Objective

The objective of this document is to guide the Data Management process required by Deep Learning (DL) constituents in the lifecycle of safety-related systems. It can be decomposed into 4 steps as can be seen in Figure 1. It is important to note that in the first iteration of the process, the data collection and data preparation steps do not need to be considered if previously generated and verified datasets are being employed for the specific application.

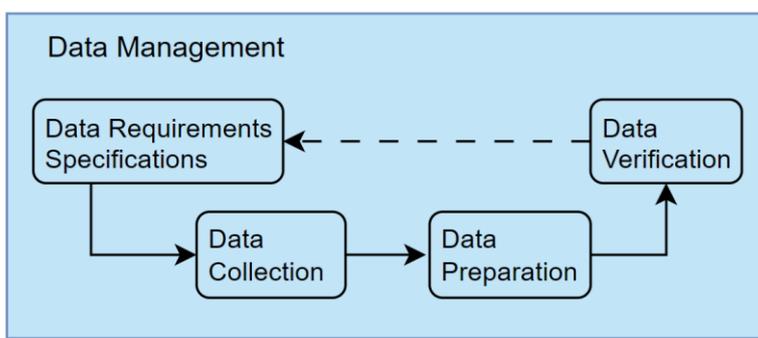


Figure 1. Data management steps

3 Scope

This guideline applies to all activities and documentation required to perform the Data Management phase.

4 Introduction

The final objective of this phase is the generation of the following datasets:

- Development dataset¹. This dataset is split into two additional ones:
 - Training datasets. Dataset employed to train the model.
 - Validation² datasets. Dataset used to evaluate if the model achieves a predefined performance and, in some cases, stops the training phase.
- Verification² dataset: This dataset expands upon the previous validation dataset to assess whether the model maintains its performance requirements with data not utilized during development. It must encompass sufficient information and data to ensure the appropriate behavior of the DL constituent within the expected Operational Design Domain (ODD) and operational scenarios. Furthermore, it should gather data to handle corner case situations that pose safety risks and confirm the fulfillment of performance requirements.

¹ In order to ensure robustness, both the training and validation datasets should encompass corner cases while also guaranteeing their representativeness of the Operational Design Domain (ODD).

² The definitions of "validation" and "verification" can vary across different technology areas or domains. In the realm of AI, "validation" typically denotes a step in the process aimed at ensuring the convergence of the developing model to terminate the AI training process. This differs significantly from the "V&V" (Verification and Validation) concepts commonly used in the functional safety community.

As a result of this phase, the following documents will be generated:

1. [REF PhDMD0001 Data Requirements Specifications.docx](#). This document collects the data requirements specifications refined from the DL requirements specifications previously defined in phase 2.
2. [REF PhDMD0002 Data Requirements Specifications IR.xlsx](#). Internal review document to be checked after completing [REF PhDMD0001 Data Requirements Specifications.docx](#).
3. [REF PhDMD0003 Data Collection Log.docx](#). Document collecting the information relative to the process of gathering and acquiring data.
4. [REF PhDMD0004 Data Collection Log IR.xlsx](#). Internal review document to be checked after completing [REF PhDMD0003 Data Collection Log.xlsx](#).
5. [REF PhDMD0005 Data Preparation Log.docx](#). Document collecting the information relative to cleaning, processing and annotating the data.
6. [REF PhDMD0006 Data Preparation Log IR.xlsx](#). Internal review document to be checked after completing [REF PhDMD0005 Data Preparation Log.xlsx](#).
7. [REF PhDMD0007 Data Requirements Verification Tests.docx](#). Data requirement tests encompass a set of metrics to assess whether the Data requirement specifications have been fulfilled, the test definitions, and their corresponding outcomes.
8. [REF PhDMD0008 Data Requirements Verification Tests IR.xlsx](#). Internal review document to be checked after completing [REF PhDMD0007 Data Requirements Verification Tests.xlsx](#).

Additionally, the following data artifacts must be generated and stored:

1. Development (training and validation) and verification datasets (previously defined). These datasets are composed of:
 - i. Collected data (raw data files). Refers to all data gathered during the collection step, including data generated from datasets, sensors, and synthetically generated data³.
 - ii. Prepared data. Encompasses all data that has undergone a cleaning, processing, or annotation process.
2. Verified datasets. Correspond with Development (training and validation) and Verification datasets that meet the data requirements specifications after performing the data verification step.

Throughout the entire process, the following should be ensured:

1. The data is free from malicious modifications or alterations while being stored or processed. This may involve reviewing the credibility of data sources, reviewing the data collection step or the data preparation step.
2. Traceability of the data, being able to determine the source of the data through the complete pipeline of operations carried out during all the processes.
3. Since traceability and the source of the data must be guaranteed throughout the entire process, a filename policy should be defined in the [REF PhDMD0001 Data Requirements Specifications.docx](#) document.

Table 1 presents the inputs and outputs associated with each step of Data Management, which will be elaborated in the subsequent sections. Section 5 guides the development of the data requirements specifications. Sections 6 and 7 guide data collection and preparation, respectively. Section 8 conducts an analysis to ascertain whether the datasets meet the data requirements stated in Section 5. Finally, Section 9 collects the acronyms and abbreviations relevant to this document.

³ The use of synthetic data together with real world data can produce the AI model to get biased during training. The use of synthetic data is subject to demonstrate that this bias is not included.

Table 1. Inputs and outputs of the Data Management phase

Phase	Step	Inputs	Outputs	Corresponding templates
PhDM Data Management	Data Requirements Specifications	<u>REF Ph2D0001 DL Requirements Specifications</u> <u>REF Ph1D0001 DL Operational Design Domain</u> <u>REF Ph1D0003 DL Operational Scenarios</u>	<u>REF PhDMD0001 Data Requirements Specifications</u> <u>REF PhDMD0007 Data Requirements Verification Tests</u>	<i>PhDMT0001_Data_Requirements_Specifications_template</i> <i>Ph0T0009_Test_definition_and_results_template</i>
		<u>REF PhDMD0001 Data Requirements Specifications</u> <u>REF PhDMD0007 Data Requirements Verification Tests</u>	<u>REF PhDMD0002 Data Requirements Specifications IR</u> <u>REF PhDMD0008 Data Requirements Verification Tests IR</u>	<i>PhDMT0001_Data_Requirements_Specifications_template_IR</i> <i>Ph0T0009_Test_definition_and_results_template_IR</i>
	Data Collection	<u>REF PhDMD0001 Data Requirements Specifications</u>	<u>REF PhDMD0003 Data Collection Log</u> Collected data structured in datasets ^(*)	<i>PhDMT0002_Data_Collection_Log_template</i>
		<u>REF PhDMD0003 Data Collection Log</u>	<u>REF PhDMD0004 Data Collection Log IR</u>	<i>PhDMT0002_Data_Collection_Log_template_IR</i>
	Data Preparation	<u>REF PhDMD0001 Data Requirements Specifications</u> <u>REF PhDMD0003 Data Collection Log</u> Raw data files structured in datasets ^(*)	<u>REF PhDMD0005 Data Preparation Log</u> Prepared data structured in datasets ^(*)	<i>PhDMT0003_Data_Preparation_Log_template</i>
		<u>REF PhDMD0005 Data Preparation Log</u>	<u>REF PhDMD0006 Data Preparation Log IR</u>	<i>PhDMT0003_Data_Preparation_Log_template_IR</i>
	Data Verification	<u>REF PhDMD0001 Data Requirements Specifications</u> <u>REF PhDMD0007 Data Requirements Verification Tests</u> Datasets ^(*)	<u>REF PhDMD0007 Data Requirements Verification Tests</u> Verified datasets ^(*)	<i>Document previously generated in data requirements specifications step</i>

(*) Datasets: i) Development (training and validation), ii) verification datasets.

5 Data Requirements Specifications

While the Data Management phase entails the formulation of individual data requirement specifications, these specifications come from the DL requirements established in prior phases.

The [REF PhDMD0001 Data Requirements Specifications.docx](#) document stored in the "PhDM Data Management" folder must be completed in this step. The provided template includes extra information guiding the definition of the requirements. Furthermore, it provides an example of data notation policy, although the user can define and employ their own policy if properly defined.

Since there are three datasets, each serving a different purpose, the data requirements for each of them will not always be identical. Therefore, there are requirements associated with the dataset (dataset requirements specifications) and requirements associated with the data comprising the datasets (data requirement specifications). For example, each dataset can have a specified number of samples (or a range of values if the requirement is not very strict), and all the data involved in the datasets can share format, accuracy, and resolution requirements. Additionally, the dataset requirements should define the degree of differentiation between the datasets. For instance, there should be dataset requirements specifying the degree of difference between validation and training datasets, which will involve differences in the associated data requirements for these datasets.

After defining the data requirements specifications, the [REF PhDMD0009 Data Requirements Verification Tests.docx](#) document must be completed. This document collects and defines the mechanisms or tests that must be carried out to check that the data meets the data requirements specifications. It must explicitly define at least one verification activity related to each data requirement. Therefore, the tests defined in this document should incorporate metrics, key performance indicators (KPIs), or other significant indicators to authenticate requirement fulfillment.

After completing the definition of data requirements specifications, associated tests, and KPIs in the [REF PhDMD0001 Data Requirements Specifications.docx](#) and [REF PhDMD0009 Data Requirements Verification Tests.docx](#) documents, an internal review of both documents ([REF PhDMD0002 Data Requirements Specifications IR.xlsx](#) and [REF PhDMD0010 Data Requirements Verification Tests.xlsx](#), respectively) should be conducted.

6 Data Collection

Data collection refers to the process of gathering and acquiring information or data from various sources, such as open-source databases, surveys, sensors, or digital platforms. This data must meet the requirements previously defined in accordance with the data requirement specifications.

During this step, the utilization of data augmentation techniques and the incorporation of synthetic data may be necessary to ensure that the data meets requirements, such as completeness. Since the information required to ensure the reproducibility of the data differs from the data directly stored, this step is decomposed into two substeps: i) data gathering—referring to the data directly obtained from sensors and datasets (before being prepared in the subsequent step) and ii) data generation—referring to the data that is generated (synthetic data, data augmentation, etc.).

The [REF PhDMD0003 Data Collection Log.docx](#) document stored in the "Data Management" folder guides this step and must be fulfilled. All the decisions about the data source, the use of augmentation techniques, or synthetic data must be described in this document. This step can be iterative, generating a new [REF PhDMD0003 Data Collection Log.docx](#) document version each time new data is collected. A copy of the raw data files collected in each iteration shall be stored in the "collected data" folder.

After completing the data collection and fulfilling the [REF PhDMD0003 Data Collection Log.docx](#) document, an internal review of this document should be conducted ([REF PhDMD0004 Data Collection Log IR.xlsx](#)).

7 Data Preparation

When data preparation is necessary, the collected data in the previous step undergoes cleaning (i.e., removing anomalies), processing (including normalization, scaling, and/or feature selection), and/or annotation (such as labeling) to align with the specified input requirements of the model to be trained or verified. In those cases, the activities performed, and the decisions made should be defined and documented in the [REF PhDMD0005 Data Preparation Log.docx](#).

It's essential to emphasize the need for harmonizing data sources to mitigate the influence of erroneous learning, as improper data modifications can significantly impact the learning process. One way to achieve this is by employing statistical methods to evaluate and ensure harmonization. After modifying the raw data from the "collected data" folder, it should be stored in the "preparation folder" changing the identifier of the data prepared to distinguish the prepared data from the collected data clearly.

After completing the data preparation and fulfilling the [REF PhDMD0005 Data Preparation Log.docx](#) document, an internal review of this document should be conducted ([REF PhDMD0006 Data Preparation Log IR.xlsx](#)).

8 Data Verification

This step must verify that the datasets generated from the previous steps meet the [REF PhDMD0001 Data Requirements Specifications.docx](#) file. The tests previously defined in the [REF PhDMD0007 Data Requirements Verification Tests.docx](#) document must be performed for that. The results and any deviation from the expected results must be reported and justified in the document.

Regarding the process by which the data is analysed, highlight that:

- The development team (collection and/or preparation) must be different from the teams performing the tests.
- For cases where the datasets are re-used from previous projects or external sources, data verification is required for the specific project.

Reminder:

- Update the state of [REF PhD0003 AI Document List.docx](#) when a document is generated or modified, including the last version generated.
- The status of the tests (Not done/Pass/Fail) must be updated in the [REF PhD0009 AI Log of Tests.docx](#).
- The tools and frameworks employed must be listed in [REF PhD0011 AI Tools Selection.docx](#).
- The traceability between DL and data requirements must be updated in [REF PhD0013 AI Traceability Matrix.docx](#)
- The tests results must be documented in [REF PhDMD0007 Data Requirements Verification Tests.docx](#)

9 Acronyms and Abbreviations

Below is a list of acronyms and abbreviations employed in this document:

- AI-FSM – Artificial Intelligence - Functional Safety Management
- FSM – Functional Safety Management
- DL – Deep Learning
- ODD – Operation Design Domain
- KPI – Key Performance Indicator
- V&V – Verification & Validation