PhDMT0001 Data Requirements Specification

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 Requirements

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

Version	Date	Description Change
V2.0	15/02/2024	Changes Applied as a result of TÜV Review 2024-01-19
V1.0	04/12/2023	First version after complete internal review
V0.3	04/12/2023	Modifications and improvements based on internal review
V0.2	30/08/2023	Modifications and improvements
V0.1	15/05/2023	First draft

*Note: The paragraphs/name of the project/Rev./Ref./history table in **blue** must be replaced with the information for the specific project. The paragraphs written in **red** are instructions that can be used as a guide, so they must be deleted.*

2 Objective

The objective of this document is to collect the safety and non-safety requirements associated with the data employed in the Artificial Intelligence-Functional Safety Management (AI-FSM).

3 Scope

This template applies to the data requirements related to the Data Management phase of the AI-FSM.

4 Definition of the Data Notation Policy

This section includes an example of naming nomenclature for the data, but users can define their own notation:

<Procedence_of_the_data>_<ID_number>_<Data_Format>. Where:

- <Procedence_of_the_data>:
  - SENS: when the raw data has been collected from sensors in the application field.
  - SYNT: when the data has been synthetically generated.
  - SIM: when the data has been generated through simulation.
  - NORM: when the data has been generated with normalization.
  - PUBL_<DATASET_NAME>: When the data has been downloaded from a public dataset. <DATASET_NAME> has to be replaced with the name of the public dataset.
  - PRIV: If the dataset is not publicly available.
- <ID_number>: Identifier starting from 0 in each of the previous groups. For example, group SYNT can have SYNT_0 to SYNT_120 and the normalization group from NORM_0 to NORM_34.
- <Data_format>: Include the data format (I.e., for images, it can refer to the resolution employed, such as 1920x1080).

5 Requirements Features and Conventions

This section provides a set of characteristics and attributes that are required for each requirement.

5.1 Description of the Requirements

All the requirements of a safety-related project must be specified following the characteristics below:

- Shall be described in such a way that they are*:*
  - **Clear.** The requirement must be easy to understand and not misleading. The requirements are written in a way that allows them to be understood by all stakeholders in the project.
  - **Concise.** The requirement must not be too long. It must be easy to read and understand, and it must not contain definitions, descriptions of its use, or reasons for its need.
  - **Unambiguous.** The requirements can be interpreted only one way.
  - **Verifiable.** The implementation of the requirement can be determined through one of four possible methods: inspection, analysis, demonstration, or test.
  - **Traceable.** A good requirement is traceable: it must be easily traced through to specifications, design and testing. Besides, it must have a unique identity or number.

- **Complete.** All conditions under which the requirement applies should be stated.
- **Feasible.** The requirement can be implemented within the constraints of the project.
    - Shall be written to aid comprehension by those who are likely to use the information at any phase of the E/E/PE safety-related system.
    - Shall be expressed in natural or formal language and/or logic, sequence or cause and effect diagrams that define the necessary safety functions.

5.2 Requirements Specifications Table

The attributes required for each requirement to ease its identification and description are defined in the following table.

Table 1. Table of attributes for each requirement

<Identifier>	<Title>
Description	
Source	
Phase of the lifecycle	
Reference	
Type	
Validation criteria	
Date	
Version	

A brief description of each field of the previous table has been done below:

- *<Identifier>.* Each Safety Requirement must be identified with a code like **REQ-YYY-OOO-ZZZ** where **REQ** refers to the *requirement* and **YYY** is a code that identifies the requirement category. In this case, the code selected is **DATRS** (*DATa Requirement Specification*). Depending on the project more subdivisions can be necessary to identify the requirements (The identifier "**OOO**" can be used to identify these subdivisions). Besides, the identifier "**ZZZ**" represents an incremental number used to identify the requirements in each sub-division. The last two identifiers are optional, and they can be replaced for another approach.
- *<Title>.* Title of the requirement. The requirements can be divided into some more specific groups that can be directly related to the application, such as components and devices of the system, the regulations, tools, programming languages, and others. In order to identify these specific sub-divisions, it can be useful to define first the name of the sub-division and then the title of the requirement.
- *<Description>.* Brief description of the requirement. The description must be short (max. 2-3 lines), aligning with the characteristics outlined in Subsection 5.1, and it must only include one requirement. Besides, it must be defined by the following keywords:
    - ***MUST.*** This word, or the adjective "required", means that the definition is mandatory.
    - ***MUST NOT.*** This phrase means that the definition is an absolute prohibition of the specification.
    - ***SHOULD.*** This word, or the adjective "recommended", means that there may exist valid reasons in particular circumstances to ignore this item, but the full implications must be understood and carefully considered before choosing a different course. A requirement expressed as "should" could become a mandatory requirement in a future version of the document.

- *<Source>.* Source of information relevant to the requirement, i.e., department, contact person, etc.
- *<Reference>.* References relevant to the requirement, i.e., documents, files, figures, emails, etc.
- *<Type>.* Mandatory/Desirable/Optional.
- <Validation Criteria> The requirements will be validated using one, or several of the following methods (at different process phases), depending on the attributes assigned to them:
  - ***Inspection***: This takes place when a direct examination of the design documentation is enough to show that the requirement has been implemented. The requirements must be traced to the proposed design for the system. This design should be defined in the documents.
  - ***Analysis***: This is a non-functional validation, which may include simulation, quantitative analysis, statistical analysis and comparison between the results obtained analytically and numerically. The requirements will be tested by a dedicated analysis.
  - ***Test***: Implies a functional validation, which can verify the suitability of the implemented requirement by direct measurement. The requirements will be verified using a test procedure. A test is defined by a set of input parameters, an environment, a procedure and a set of outcomes or outputs.
- *<Phase>:* Specification/Architecture/Detail. Phase in which the requirement is defined as specified in the V-model. For the data requirements case, Data Management.
- *<Date>:* Date of creation of the requirement version.
- *<Version>:* Requirement version. All versions are included in the report in a consecutive order.

6 Data Requirements Specifications

The data requirements specifications could be structured in the following way:

6.1 Data Requirements Specification

Data requirements associated with attributes:

1. Format. Specify the desired image file format and compression standards (e.g., JPEG, PNG) to ensure compatibility with the model and efficient data storage.
2. Data characteristics. Detail image properties such as resolution, colour depth, brightness, contrast, and any other relevant visual attributes crucial for perceptual tasks.

6.2 Datasets Requirements Specifications

1. Completeness. Completeness requirements relate to identifying and defining all object categories that should be covered with the data. Completeness also implies having sufficient variation within each class of object, and therefore this variability must be defined in the requirements. Example: in a visual perception task, specify that the “car” class must include various types of cars (trucks, SUVs,…) in several conditions and environments.
2. Representativeness. Requirements associated with ensuring that data are representative of the Operational Design Domain (ODD). I.e., the definition of visual scenarios, viewpoints, lighting conditions, and object variations. Furthermore, the data must maintain representativeness throughout the intended usage period. If there are modifications to the ODD post-system deployment, a reanalysis of the Data Management phase is necessary.
3. Balance. It refers to achieving a balanced distribution of instances across different object classes within the dataset. This balance is essential to ensure that the model is exposed to an equitable representation of each class, preventing biases and promoting fair learning. For instance, ensure an approximately equal distribution of instances across all relevant vehicle classes. Following the

same example, this means that the number of images containing sedans, trucks, motorcycles, and other vehicle types is roughly the same.

4. Volume. Define requirements associated with the total number of images required, considering both the diversity and quantity.
5. Data origin (collection). Specify the allowed sources of the data, private or public dataset, the need for real-world data, synthetic data, or a combination of both, explaining the rationale behind the choice (availability, cost…). Include all the data requirements associated with them.
6. Degree of differentiation between datasets. These requirements are designed to establish distinctions among the training, validation, and verification datasets. Examples of such requirements may include, but are not limited to:
    a. Ensuring that datasets are disjoint, with samples/data in each not included in the others.
    b. Specifying that the percentage of samples in the training dataset shall be higher than that in the validation and verification datasets and providing the numerical value for this percentage.
    c. Providing more extensive coverage for certain objects in the verification dataset compared to the validation dataset.

6.3 Datasets Preparation Requirements

Requirements associated with datasets preparation:

1. Labelling annotation. Define quality standards for image labelling, including precise object annotations, semantic segmentation, or any other relevant labelling schema, ensuring accuracy and consistency.
2. Data augmentation. Describe permissible augmentation techniques, such as rotation, translation, scaling, and colour transformations, essential for expanding the dataset while maintaining its relevance and diversity for training and validation.
3. Data cleaning: Removing anomalies using an anomaly detector, imputing missing values, correcting erroneous values or standardizing them (e.g., cropping to remove irrelevant information from an image), filling missing data (in general by inputting the mean value for the data in the dataset). etc.
4. Data pre-processing: Normalization of the input data to help prevent features with larger magnitudes from dominating those with smaller magnitudes (e.g., mi-max scaling, z-score normalization, robust scaling to reduce the sensibility to outliers…), Scaling, Feature Selection, Dimensionality Reduction, Data Balance, etc.

7 Data Requirements Overview

To obtain a comprehensive overview of all the requirements established in this phase, it is essential to collect them in the following table:

Table 2. Data requirements overview

<Identifier>	Description

***Reminder**:*

- Update the state of REF_Ph0D0003_AI_Document_List.docx when a document is generated or modified, including the last version generated.

- The tools and frameworks employed must be listed in REF_Ph0D0011_AI_Tools_Selection.docx.

- The traceability between the DL with respect to the data requirements specifications must be updated in REF_Ph0D0013_AI_Traceability_Matrix.docx

8 Acronyms and Abbreviations

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

- AI-FSM – Artificial Intelligence - Functional Safety Management
- ODD – Operational Design Domain

9 Bibliography

Add here the reference to used bibliography / references (if any).