



Safe and Explainable
Critical Embedded Systems based on AI

D7.2 Data Management Plan

Version 1.1

Documentation Information

Contract Number	101069595
Project Website	www.safexplain.eu
Contractual Deadline	31.03.2023
Dissemination Level	PU
Nature	DMP
Author	Jaume Abella (BSC)
Contributors	Francisco J. Cazorla (BSC)
Reviewer	Thanh Bui (RISE)
Keywords	Data Management Plan



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

Change Log

Version	Description Change
V0.1	First draft
V0.2	Comments from reviewers
V1.0	Final Version
V1.1	Final Version (for the resubmission at M24 following the Review recommendations)

Table of Contents

1. Introduction.....	3
2. Data Summary & Datasets	3
3. FAIR (Findable, Accessible, Interoperable and Re-usable) Data.....	6
3.1. Findable	6
3.2. Accessible.....	9
3.3. Interoperable.....	12
3.4. Re-usable.....	12
4. Responsibilities	13
5. Summary	14

Executive Summary

The Data Management Plan (DMP) has been developed in compliance with the latest EC DMP guidelines.

SAFEXPLAIN will produce 8 datasets, each one with different characteristics. This document identifies and describes these 8 datasets and provide details on FAIR (Findable, Accessible, Interoperable and Reusable) for each one individually. Finally, we indicate who is responsible for each dataset and what implications this brings.

The current version of SAFEXPLAIN DMP is the official deliverable, however this will be a living document and evolve during the project lifespan according to the progress and needs of project activities. Hence, updates of this document will be released at M12, M24 and M36, as described in the DoA.

1. Introduction

This document describes the Data Management Plan (DMP) of the SAFEXPLAIN project and corresponds to Deliverable 7.2 (D7.2). The DMP explains the generic datasets that the project will generate and how they will be managed.

This document is structured into 4 content sections. Section §2 describes the main datasets that the project will generate. Following that, the FAIR standards are specified in section §3. Section §4 describes who is responsible for the datasets. Finally, section §5 summarizes the DMP.

2. Data Summary & Datasets

The SAFEXPLAIN project analyses solutions and technologies that enable the use of AI-based software in the context of safety-critical systems. The datasets generated aim towards eight objectives. (1) Requirements and success criteria specification, (2) source code for the DL libraries, (3) explanations and drawings for the safety argumentation, (4) software implementation for the timing analysis and other related platform tools, (5) software implementation of the safety analysis commercial toolset, (6) technologies validation and evaluation in the context of some benchmarks and case studies, (7) DL model training, and (8) project progress tracking.

The SAFEXPLAIN project will generate the following datasets:

DATASET NAME	DESCRIPTION
DATASET-REQSC	Datasets obtained from the set of requirements and success criteria for the different developments of the project. A table lists and monitors all requirements, with information such as ID, owner, description, WPs involved, and the success criteria to be met by different key times in the project such as month 6, 18, 30 and 36. This table will be stored in the form of a spreadsheet. Its contents are project specific and are unlikely to be of the interest to external audiences. Nevertheless, the contents of such table are already part of D1.1, whose dissemination

	level is public and hence, this dataset has been made public already along with D1.1.
DATASET-DLSRC	<p>Datasets consisting of the source code for the DL libraries providing explainability and traceability. Whether these libraries will be implemented in C++, Python, or any other language is decided on a case-by-case basis. In the case of DLlib, since it is intended to be integrated along with DATASET-COMTOOLS, the language is constrained to make such integration possible.</p> <p>Those DL libraries will be used for the assessment of the case studies and will form a dataset of the project. Potentially, along with the source code itself, DL parameters (weights) for such DL models will be delivered if relevant. Such DL parameters will be obtained by training the DL models with publicly available datasets.</p>
DATASET-FUSA	Datasets with safety argumentation supporting files. Appropriate descriptions and explanations of how safety standards need to be updated to allow DL software to be certified, along with the associated safety argumentation and diagrams, form another key dataset of the project. The relevant test scenarios description is also part of this dataset.
DATASET-TIMETOOLS	Datasets for timing analysis tools and supporting libraries. A number of source files will conform the instrumentation libraries to collect timing measurements, and the libraries to interface the performance monitoring counters of the target platform. We jointly refer to those as the PMUlib. Some other source files, typically to run on R or Matlab, will provide capabilities for analysing the execution time measurements obtained and raise timing predictions. Finally, some post-processing scripts tailored to connect the aforementioned components with the middleware part of the DATASET-COMTOOLS will also be developed. All those source files together form another dataset.
DATASET-COMTOOLS	Datasets for the commercial toolset for safety analysis of DL libraries. Source code integrated in the corresponding toolset forms this dataset along with test descriptions for each scenario in AI-V&V-SCENARIOS, where those tests are referred to as AI-V&V-TESTSPECS. The overall dataset is not intended to be shared as open source for commercial reasons.
DATASET-CASESTUDY-OUTPUT	Datasets generated from the validation and evaluation of the case studies on the target platform. These datasets include the measurements, test-cases and proof to assert the validation of the developed technologies when run on top of the target platform, as well as data gathered from performance monitoring units and derived metrics. Files from this dataset aim to validate and evaluate the platform. Overall, this dataset will consist of a small set of values with no expected use other than its interpretation to assess the degree of

	<p>success of the SAFEXPLAIN technologies when applied to the specific case studies.</p>
DATASET-CASESTUDY-MODEL	<p>This dataset includes the DL models of each case study, as well as a shared DL model whose characteristics resemble as much as possible those of the actual case studies of the project, but with the difference that this DL model is intended to be distributed openly. This demo case study, which we often refer to as “toy model”, internally allows integrating technologies developed into a shared model without IP constraints, and externally allows providing an open model that can be used to assess open source technologies of the project as well as external technologies. Any other use will be allowed. Note that the data used for training and validation of all these models is not included in this dataset. Instead, it is part of the DATASET-TRAINING.</p>
DATASET-TRAINING	<p>Regarding the automotive and space DL models, proprietary training and validation datasets will be used to train and validate those models. Hence, those datasets will remain confidential.</p> <p>Regarding the railway DL model, public datasets will be used, and relevant documents and reports will properly identify those datasets. Note that, eventually, proprietary datasets may be used for training. If that is the case, then those proprietary datasets will remain confidential, as in the case of the automotive and space case studies.</p> <p>Regarding the demo case study, datasets for training and validation will be generated and distributed openly through public and perdurable repositories. Those datasets will be produced following similar methods to those used the space ones, but without imposing any confidentiality constraints so that they can be distributed.</p>
DATASET-PUBLIC-DEMOS	<p>This dataset aims at complement DATASET-CASESTUDY-MODEL and DATASET-TRAINING proprietary parts (i.e. those for the automotive, space and railway case studies) with excerpts, videos, presentations, and any other related material showcasing the proprietary case studies when building on top of the SAFEXPLAIN solutions and technologies. Note that this dataset does not include the so-called “toy model” since it is already fully open and anyone can fully use it and test it.</p>
DATASET-PROGRESS	<p>Datasets created to document the progress of the project. This dataset includes any document generated to track the project evolution and its results. Two types of documents are expected: (1) public deliverables, and (2) sensitive deliverables and progress reports. The second set of documents responds to project monitoring objectives only.</p>

3. FAIR (Findable, Accessible, Interoperable and Reusable) Data

3.1. Findable

Of all the data generated, only a limited amount will be findable and accessible. We proceed to describe the findability status of each data set:

DATASET NAME	DESCRIPTION
DATASET-REQSC	Requirements and success criteria are managed internally in the form of a spreadsheet for their use and monitoring during the execution of the project. Such use is exclusive to the project, and hence, the spreadsheet is not distributed. Instead, requirements and success criteria have been included as part of D1.1, as indicated before, which is already public. D1.1 is already available through the website of the project at https://www.safexplain.eu/ as well as through the CORDIS portal (https://cordis.europa.eu/). By being available through D1.1 in the project website and CORDIS portal, this dataset is easily findable during the project duration, when its contents have some use and meaning, and beyond the project duration through the CORDIS portal.
DATASET-DL SRC	<p>The project will produce two major DL libraries, namely Explib and DLlib, as well as other smaller libraries that will be developed opportunistically based on the research findings. Those DL libraries will be released as open source. Such libraries will be shared through a public and perdurable repository not yet decided on, and will include in their name and description key terms identifying their relation to DL, functional safety, and explainability, as well as to the name of the project, to make them findable. The candidate repositories considered so far include Zenodo, Github, Gitlab and the AI on Demand platform.</p> <p>In fact, a library developed opportunistically to realize semantic diverse redundancy for object detection with YOLO is already publicly available through Zenodo (https://zenodo.org/doi/10.5281/zenodo.12606744) and can be easily found. We refer to that specific lib such as DivRedLib.</p>
DATASET-FUSA	This dataset will be organized in the form of technical documents, processes descriptions and templates that will be released as open source packages for their use, and also described through deliverables, publications in conferences and journals, white papers, webinars, as well as presentations and tutorials that can also be recorded. All these supporting formats will be provided or linked through the project website to make them findable. Moreover, some key materials will be part of conference and journal proceedings with a DOI, and posted in perdurable repositories such as UPCommons, hence with a permanent link also. Those contents will be conveniently named and tagged with

	<p>terms such as functional safety, DL, certification, explainability and SAFEXPLAIN to make them findable.</p> <p>As part of this dataset, the AI-FSM, a set of processes and associated document templates devised to allow the use of AI software in the context of development processes with functional safety requirements, has already been released publicly in Zenodo, so that it can be easily found (https://doi.org/10.5281/zenodo.10964402). Such dataset describes the steps to follow and provides the templates to be filled in for the aforementioned development process. A paper further describes the overall AI-FSM rationale and use (https://safexplain.eu/publication/ai-fsm-towards-functional-safety-management-for-artificial-intelligence-based-critical-systems/). Such paper has been presented as part of the CARS'24 (https://conf.laas.fr/cars/cars2024.html) workshop.</p> <p>Analogous strategies will be used for the rest of the processes and templates composing this dataset. For instance, the AI-related verification and validation scenarios description (AI-V&V-SCENARIOS) has been publicly presented in the form of slides describing one scenario each and distributed along with D2.1 publicly.</p>
DATASET-TIMETOOLS	<p>This dataset will be released in the form of standalone tools, or packages extending popular tools such as R or Matlab, with accompanying documentation that will include descriptive key terms in the name and in the keywords provided along with the tools/packages, such as DL, timing analysis, functional safety and SAFEXPLAIN.</p> <p>The PMUlib will be kept proprietary. It will be announced through the IP catalogue of the BSC website so that it can be found and, eventually, licensed.</p> <p>Analysis tools and packages using the output of the PMUlib, but that could be connected to other tools, will occur through public repositories so that they are findable and accessible for a long time, such as the CRAN project for R files (https://cran.r-project.org/).</p> <p>Finally, interfaces for the middleware of the DATASET-COMTOOLS have no entity on their own, so they will be kept proprietary and distributed along with the middleware.</p>
DATASET-COMTOOLS	<p>This dataset, which includes the middleware on top of which the use cases run, as well as some other analysis tools and AI-V&V-TESTSPECS, is only intended to be used for commercial exploitation, either as standalone tools or integrated into a larger tool, and as a collection of test specifications. The tool and test specifications will be publicized as needed for commercial reasons, and the dataset may be referred to in such process if found pertinent but, in general, this dataset will be private, stored in internal servers of the partner(s) producing it, and not explicitly findable.</p>

<p>DATASET-CASESTUDY-OUTPUT</p>	<p>This dataset will be included as part of technical documents, where context for each of the values of the dataset is given (e.g., what model and inference dataset has been used for a particular accuracy value part of the DATASET-CASESTUDY-OUTPUT), which will be released through deliverables, publications in conferences and journals, white papers, as well as presentations and tutorials that can also be recorded. All those supporting formats will be provided or linked through the project website to make them findable. Moreover, some key materials will be part of conference and journal proceedings with a DOI, or posted in perdurable repositories with a permanent link. These contents will be conveniently named and tagged with terms such as case study, functional safety, DL, certification, explainability and SAFEXPLAIN, as well as the corresponding domain for the case study (automotive, space, or railway) to make them findable.</p>
<p>DATASET-CASESTUDY-MODEL</p>	<p>DL model for the actual case studies of the project, namely the automotive, space and railway case studies from NAV, AIKO and IKR respectively, are proprietary and they will remain confidential by default. Instead, the DL model of the demo case study will be shared openly through public and perdurable repositories and frameworks so that it is easily findable, such as Github, Gitlab, Zenodo and/or the AI on Demand platform. Appropriate tags will be used to describe its characteristics.</p>
<p>DATASET-TRAINING</p>	<p>Private datasets from the automotive and space case studies, as well as, potentially, some for the railway case study, will be stored in internal servers of the partner(s) producing them, and not findable.</p> <p>Regarding the railway case study, public datasets will be used for training and validation, and relevant documents and reports will properly identify those datasets and how they have been used, to allow finding them easily. However, since they are already public, they will not be replicated into any other repository.</p> <p>Regarding the demo case study, datasets will be distributed through open source repositories and frameworks, and tagged precisely to make them findable.</p>
<p>DATASET-PUBLIC-DEMOS</p>	<p>This dataset will be distributed in different forms potentially including binaries execution excerpts of the automotive, space and railway case studies, as well as video recordings of such execution and technology integration, presentations, pictures, among other types of material. Each individual format for each individual case study will be publicized through the project website and social media, and through the most appropriate channels for the target audience for each case study, such as domain-specific fora, conferences and events, and whenever appropriate through public repositories such as Github, Gitlab, or Zenodo.</p>

DATASET-PROGRESS	<p>This dataset will be organized in the form of documents that will be released through deliverables progress reports. As said before, it includes (1) public deliverables, and (2) sensitive deliverables and progress reports.</p> <p>Regarding (1), they will be distributed openly through the project website, as well as through the CORDIS portal. In fact, some deliverables are already public through these means. This way, those deliverables will be easily findable, and perdurable through, at least, the CORDIS portal.</p> <p>Regarding (2), these documents include information that is confidential for at least a subset of the partners, such as business-confidential information, financial information and, indirectly, aggregated information about personnel (e.g., information related to salaries and workforce). Hence, this dataset will stay confidential and only will be shared with the Project Officer and reviewers, as requested by the European Commission. The dataset will be stored in private repositories, such as OneDrive, with access control and without disclosing any personal data.</p>
------------------	--

3.2. Accessible

Datasets that qualify for public availability will be made accessible through the project’s website and through public repositories such as re3data.org and Zenodo. In the case of source code data sets, they will be shared internally through git repositories. Particularly, BSC’s internal gitlab will be used whenever IP confidentiality permits. Thus, all data will be stored in BSC’s servers if IP confidentiality permits.

DATASET NAME	DESCRIPTION
DATASET-REQSC	This dataset, as part of D1.1, is accessible in the form of a document (PDF) through the project website. As a spreadsheet, instead, it is used as a tool for the project execution and only accessible through a OneDrive repository with access control so that only project partners can modify its contents.
DATASET-DLSRC	This dataset will be managed through repositories such as Github, Gitlab and the like. The specific location of such repository and the access allowed will depend on whom the contributors are and whom the partners are that need to use it. In general, during development phases, only partners contributing to the corresponding DL libraries will be granted access to the corresponding repositories. Whenever those libraries are ready to be used, other members of the project, such as case study owners, will be granted access either to the binaries only, or to the source code and binaries. Eventually, whenever those libraries

	<p>are mature enough, the corresponding repositories will be open to the public, or cloned into public repositories or portals with public access such as Zenodo, Gitlab, Github, or AI on Demand.</p> <p>For instance, in the particular case of the DivRedLib library, it is accessible through Zenodo, as explained before (https://zenodo.org/doi/10.5281/zenodo.12606744).</p>
DATASET-FUSA	<p>This dataset will be accessible through public repositories (e.g., Gitlab, Github, Zenodo), and described in public deliverables shared through the project website. Whenever parts of it are distributed through publications in conferences and journals, such documents will also be placed in public repositories with open access so that they are accessible unconstrainedly.</p> <p>While publications will describe the purpose and use of the processes and templates composing this dataset, the processes and the corresponding templates to be used by end users will be made accessible through public repositories, as indicated before. In the particular case of the AI-FSM, part of this dataset, as explained before, it is already public in Zenodo. AI-V&V-SCENARIOS is already accessible along with D2.1, although further channels for public distribution will likely be explored.</p>
DATASET-TIMETOOLS	<p>The PMUlib will not be accessible openly, and access will only be provided under an appropriate license to be agreed.</p> <p>Tools and packages composing the analysis part of this dataset will be released publicly through public repositories such as the CRAN project so that they are accessible. These repositories are perdurable, thus easing accessibility. Other packages have been released by the owners of this dataset (BSC) through these means in the past, so they have the experience to do it.</p> <p>Regarding the scripts to interface with the middleware, they will not be accessible in general except for authorised users of the middleware.</p>
DATASET-COMTOOLS	<p>This dataset will be part of a commercial toolset. Hence, it will not be accessible to anyone but the owner of the dataset. Only the needed access for the execution of the project will be given to the other partners, which typically will occur in the form of a binary and textual test specifications, not source code. The owner will decide on the commercial use of this dataset, either to provide services to customers, or for internal use to produce results for the customer as part of a service.</p>
DATASET-CASESTUDY-OUTPUT	<p>This dataset will be made available in the form of documents through the project website, whether deliverables, technical papers or videos to offer tutorials. These documents put in the form of a technical paper, typically published in a conference or journal, or as a white paper, will be deployed in public repositories with public access and a DOI so that</p>

	<p>they are accessible and perdurable. While some values part of this dataset may be depicted in figures in those documents, they will be included in tables for any other analysis that anyone could want to do. In any case, those values forming this dataset are final results without relevant uses expected beyond the project itself, and typically will not exceed few tens of values. Hence, if, eventually, anyone wanted to use them for any unforeseen purpose, they will be easily accessible in the corresponding reports and documents from where it will be possible to retrieve them.</p>
DATASET-CASESTUDY-MODEL	<p>The DL models for the three case studies of the project, since they are proprietary, will not be accessible publicly in general, and will be used and/or licensed as needed by their respective owners in accordance with their exploitation strategies. The demo case study, instead, will be fully open and released with a permissive license, and distributed through open repositories to make sure it is fully accessible. It will be released in a standard format for DL models to be decided (e.g., ONNX).</p>
DATASET-TRAINING	<p>Proprietary datasets are, typically, business critical. Hence, they will not be accessible to anyone but the owner of the dataset. Only if strictly needed for the execution of the project, access will be given to the other partners. However, we foresee that only the trained models will need being shared across partners, but not the data used to train them.</p> <p>Regarding the public datasets used to train the railway case study, they are already public and accessible.</p> <p>Regarding the datasets used to train and validate the demo case study, since they will be distributed through open repositories and platforms, they will be easily accessible.</p>
DATASET-PUBLIC-DEMOS	<p>This dataset will be open and, at least a subset of format for each case study will be accessible through public repositories such as Github, Gitlab, or Zenodo. Other channels such as conference websites will also be used, yet backed up with perdurable repositories to ensure long-term accessibility.</p>
DATASET-PROGRESS	<p>Regarding public deliverables, they will be accessible through the project website and the CORDIS portal. In fact, some of them are already public.</p> <p>Regarding confidential deliverables and progress reports, since these datasets are intended to remain confidential within the consortium, they will only be accessible through the internal OneDrive repository with access control so that only partners can access them. Eventually, these documents will be sent to the EC when finalized for their review, but will not be offered to anyone outside of the consortium or the EC.</p>

3.3. Interoperable

Datasets in SAFEXPLAIN are generally isolated from each other, and interoperability is needed mostly within datasets across different subsets. For instance, as part of the DATASET-TIMETOOLS, there will be libraries instrumenting applications and producing execution time measurements (PMUlib) that will be consumed by some timing analysis tools also part of this dataset. Hence, both subsets of the dataset need being consistent so that they can exchange results (i.e., the outcome of one subset is the input of another subset). However, interfaces are generally trivial (e.g., files listing one execution time value per line) with the aim of making public components (e.g., the analysis tools) compatible with other sources of data to enable a wider use of the tools.

In terms of interoperability across different datasets, as said, there is not such need in general, with just few exceptions. Datasets are related semantically, but do not need to be consumed by other datasets. For instance, DATASET-FUSA provides arguments, templates to be filled in when applied to DATASET-DLSRC, and test scenarios descriptions, but does not use DATASET-DLSRC itself for the generation of those arguments or to fill in those templates. The main exceptions to this are (1) the DLLib, part of the DATASET-DLSRC, which is used by the middleware part of DATASET-COMTOOLS. However, the DLLib is simply encapsulated with an additional container so that it fits the interfaces used by the middleware. And (2), the DATASET-TRAINING, which includes training and validation data used by the DATASET-CASESTUDY-MODEL.

However, while intra-project interoperability is not needed, interoperability beyond project boundaries needs to be addressed conveniently. For that purpose, where applicable, datasets will be accompanied with appropriate metadata according to relevant EC standards such as DCAT-AP.

3.4. Re-usable

The datasets that qualify for public availability will be licensed under specific licenses. Scientific dissemination in the form of papers will be licensed under Creative Commons (BY-NC-SA). Datasets that include open source code will be licensed with permissive open source licenses such as MIT or Apache 2.0. The table below provides the current type of license of each dataset as planned today, which will determine the ease of reuse. Note that such licenses will be reassessed in a case-by-case basis depending on the exploitation opportunities that arise during project execution.

The datasets that qualify for public availability will be licensed under Creative Commons (BY-NC-SA, CC BY 4.0 or CC BY-NC-SA 4.0) to increase their scientific and utility dissemination, yet in some cases preserving rights for exploitation.

DATASET NAME	DESCRIPTION
DATASET-REQSC	Creative Commons Attribution 4.0 International (CC BY 4.0) licence
DATASET-DLSRC	Apache 2.0 or other non-restrictive open source licenses. For instance, DivRelLib has already been released publicly with Apache 2.0 license.
DATASET-FUSA	Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) has been used for AI-FSM. Creative Commons Attribution 4.0 International (CC BY 4.0) has been used for AI-V&V-SCENARIOS so far by being part of D2.1
DATASET-TIMETOOLS	Proprietary (PMUlib), GPLv3 for analysis tools due to their dependence on R, and proprietary for the interface scripts for the middleware
DATASET-COMTOOLS	(commercial use only)
DATASET-CASESTUDY-OUTPUT	Creative Commons (BY-NC-SA or CC BY 4.0), with the license being determined by the documents where they are released.
DATASET-CASESTUDY-MODEL	(commercial use only for automotive, space and railway models) Creative Commons (BY-NC-SA or CC BY 4.0) preferably, but the exact open license is not yet fully decided.
DATASET-TRAINING	(commercial use only for automotive and space datasets, and some proprietary datasets that may be used for the railway case study) Public datasets for the railway case study have their own licenses already. For the demo case study, creative Commons (BY-NC-SA or CC BY 4.0) preferably, but the exact open license is not yet fully decided.
DATASET-PUBLIC-DEMOS	Creative Commons (BY-NC-SA or CC BY 4.0), depending on the channel where they are distributed, for most of the formats, yet the specific open license to be used for each format will be decided in a case-by-case basis.
DATASET-PROGRESS	For public deliverables, Creative Commons (BY-NC-SA or CC BY 4.0), depending on the channel where they are distributed. (confidential for the rest of documents)

4. Responsibilities

Partners will be responsible of each dataset generated. Requests for sharing datasets will be managed internally by email explicitly asking the owner of the dataset for access. The owner of the dataset will deliver the dataset without unreasonable delay, if needed for the execution of the project.

5. Summary

Overall, SAFEXPLAIN will generate several datasets of a different nature. Some of these consist of public reports and documentation that will be made available through the project website so they are easily findable and accessible. A relevant subset will be published in conferences and journals and made publicly available in perdurable repositories. Other datasets include open source code and will be distributed through public repositories with permissive licenses. Finally, some datasets and subsets of datasets are intended to remain confidential for their commercial exploitation only.

Acronyms and Abbreviations

- D – deliverable
- DL – Deep Learning
- DMP – Data Management Plan
- DoA – Description of Action
- EC – European Commission
- MIT – Massachusetts Institute of Technology
- WP – Work Package