



HORIZON EUROPE PROGRAMME: HORIZON-CL4-2022-DIGITAL-EMERGING-02

SolIDAIR

**Solid, rapid and efficient adoption of Data, AI & Robotics
applications in production**

Deliverable D1.4: Data Management Plan Update

Primary Author(s)	Sophia Bastidas VIF
Deliverable Type	Report
Dissemination Level	Public
Due Date (Annex I)	31.03.2025 (M18)
Pages	25
Document Version	Final
GA Number	101120276
Project Coordinator	Andreas Frommknecht Fraunhofer IPA (FHG) (andreas.frommknecht@ipa.fraunhofer.de)

Contributors	
Name	Organisation
Sophia Bastidas	Virtual Vehicle Research
Hannes Allmaier	Virtual Vehicle Research
Tobias Weigand	Brose
Kerman Osoro	CIE
Hakan Muslu Cem	Bosch
Alexander Katzbeck	AUTFORCE
Panagiotis Giannikos	THL
Linghao Zhou	THL

Formal Reviewers	
Name	Organisation
Linghao Zhou	THL
Gerhard Baier	UGS

Version Log			
Rev #	Date	Author	Description
0.1	07.02.2025	Sophia Bastidas (VIF)	Release Candidate
1.0	19.03.2025	Linghao Zhou (THL) Gerhard Baier (UGS)	Quality review
2.0	26.03.2025	Anesa Begovic (i2m)	Formatting check
3.0	28.03.2025	Andreas Frommknecht (FHG) Emil Andreas (FHG)	Coordinator review and approval, deliverable ready for submission

Project Abstract

SoliDAIR aims to accelerate the uptake of Artificial Intelligence (AI) and Robotics in European manufacturing, using data as an enabler. It will co-develop and demonstrate tailored solutions to digitalise and automate visual inspection and physical testing, enable predictive quality control and process optimisation. The SoliDAIR project tackles the problem of AI & Robotics systems not being extensively used in the manufacturing industry, because it is not clear whether these systems are safe and when or why they will fail. By researching, developing and testing methods that are as solid and trustworthy as possible to be adopted by the European industry, while being cost-efficient to develop and replicate.

New methods and tools will be developed by research and technology providers, which leverage the current state of the art in visual AI, AI for process data, and smart & collaborative Robotics. The developed technologies will be applied and demonstrated in four industry use cases to prove their functionality and applicability in real and mature production environments. The objective is to improve production processes through digitalised and automated quality control for high volume, high rate and flexible manufacturing. The developed methods shall be efficiently and easily adaptable and replicable, so they can be easily applied to new use cases outside the consortium.

Table of Contents

1	Public Summary	4
2	Introduction	5
2.1	Rationale of this deliverable	5
3	Data Summary	6
3.1	Open data	13
4	Fair Data	14
4.1	Making Data Findable	14
4.2	Making Data Accessible	15
4.3	Making Data Interoperable	17
4.4	Making Data Reusable	18
5	Allocation of Resources	19
6	Data Security	21
7	Ethical Aspects	22
8	Conclusions	22
9	Bibliography	23
10	Acknowledgements and disclaimer	23
11	Abbreviations and Definitions	24
12	List of Tables	25

1 Public Summary

The Data Management Plan (DMP) presented in this deliverable is an updated version of the first DMP (deliverable D1.3). It describes the strategy for managing all research data generated, utilized, and shared throughout the project so far. The DMP emphasizes transparency, collaboration, and efficient data sharing among partners while promoting secure storage and adherence to data quality practices.

The purpose of this DMP is to support the SoliDAIR project by providing partners and interested parties with clear and concise information about the data generated and utilized in the project, where it can be found, the access restrictions and licenses, data formats and software requirements, among others. Additionally, it aims to enhance the project's visibility and impact by promoting data sharing, reuse, and open publication whenever possible. To help achieve this target and promote open science, partners generating data have proposed a few datasets that will be made available to the public after anonymization due to confidentiality restrictions. To share the open data, a community has been created in Zenodo (<https://zenodo.org/communities/solidair/>), which is an open-access repository supporting the FAIR (findable, accessible, interoperable, and reusable) data principles. Additionally, the links to the open datasets will be provided on the SoliDAIR webpage [1] as soon as they are available.

This latest version of the DMP has been developed in accordance with the European Commission's guidelines for FAIR data management in Horizon 2020 [2], which provides a questionnaire for each of the FAIR principles to promote their implementation as effectively as possible in the projects. Additional sections for resource allocation, data security, and ethical considerations are also included in the DMP, ensuring the long-term preservation of the data and, therefore, that it is accessible after the project is finished. All partners contributed to addressing these questions thoroughly, using the most up-to-date information available.

2 Introduction

2.1 Rationale of this deliverable

This deliverable is produced under the frame of work package 1 (WP1) - Project Management and Coordination- and it presents the updated version of the DMP for the SoliDAIR project. Building on the foundational principles established in Deliverable D1.3, this revised DMP aims to enhance the management of data generated and utilized throughout the project. It emphasizes the implementation of the FAIR data principles in accordance with the EU guidelines.

The information in this deliverable reflects the contributions of all SoliDAIR partners, especially the Use Case (UC) owners. To compile these contributions, a questionnaire was created based on the DMP template proposed by the EU [2], which included specific questions aimed at enhancing the DMP development, with a strong emphasis on the FAIR data principles. In this way, the DMP deliverable has been structured as follows:

- A data summary section where both confidential and open dataset information (metadata) has been provided. Additionally, there is a dedicated subsection for open data featuring a summary table that provides relevant details about the datasets that are planned for publication in the Zenodo SoliDAIR community.
- A FAIR data section where different questions are addressed for the implementation of each FAIR principle in the SoliDAIR project.
- An allocation of resources section to outline the costs related to data management.
- A data security section where the security measures implemented during data handling and management are described.
- A section to outline any ethical and legal aspects of data usage within the project.
- Conclusions.

Deliverable D1.4 is associated with the deliverables in work package 5 (WP5), which focuses on the dissemination and communication activities. This includes, among other things, the management of intellectual property rights (IPR), dissemination tools like the SoliDAIR website and social media channels, as well as journal publications and conference presentations.

Attainment of the objectives and explanation of deviations

The objectives of this deliverable have been achieved without deviations.

3 Data Summary

The datasets created, used, and shared within the SoliDAIR project are crucial for all scientific research conducted here and, in most cases, contain sensitive and confidential information. Despite the restrictions that confidentiality may impose, SoliDAIR encourages the use of the FAIR data principles, ensuring secure data sharing within the project while facilitating the publication and reuse of data whenever possible.

To provide a clearer overview of the datasets used in the project's development thus far, summary tables have been created for each dataset, shown in Table 2 to Table 13. In the last table, a non-UC-related dataset is presented, which is used in the project for methodology development. These summary tables include key information such as data format, size, location, access restrictions, their relevance to the project objectives and UCs and the data level defined within the SoliDAIR project and summarized in the following Table 1. The template used for the summary tables aligns with the Dublin Core guidelines for metadata creation [3].

Regarding open publications, a Zenodo community has been set up. Therefore, all the open datasets in the project will be made available on the same platform. Nonetheless, it is also encouraged that partners publish their data on platforms relevant to the specific topic of the data; additional information in this regard is given in the next section.

Table 1. Levels of SoliDAIR data.

Type of data	Findability	Accessibility/ Reusability	Interoperability	Curation/ storage
Level 1: Use case data such as sensor and process data, QC data etc.	Data will be held within the use case under use case owners information security / storage policies.	All consortium partners on need-to-know basis. Industrial standards (e.g. Functional Mock-up Units (FMU)) will be used for model exchanges between partners	Ensured through interaction with WP2 and WP5	Depending on WP (e.g., large scale data storage)
Level 2: Data from WP2 such as strategies, frameworks, source code, test data, tools	Managed via in project SharePoint with suitable access permissions and code repositories like GitHub with links to the project website and partners' repositories	For all consortium partners; available at request to the public, data & code will be shared with 3rd parties through suitable licencing	Will be ensured through quality management measures at project level	FHG, I2M and all partners involved in WP2
Level 3: Documents for publications e.g. research papers, guidelines, blueprints	Final analysis of data available through workshops, publications (also using ORE (Open Research Europe)) and project website; source code will be managed via repositories e.g. GitHub	For all consortium partners; available to the public, data & code will be shared with 3rd parties through suitable licencing like GPL-3.0 or creative common license model CC BY-SA	Ensured through interaction with WP2 and WP5, quality management WP1	All partners involved in WP5

Table 2. Dataset 1.

Topic	Description of Data collected/generated
Dataset name	Brose door modules - production dataset - V01
Contact information	Brose Bamberg, Berliner Ring Tel.: +49 951 7474 0 E-Mail: bamberg@brose.com
Source/origin	Brose Universal End-Of-Line Tester
Level (of SoliDAIR data)	Level 1
Modification date	From 2024-06-04 to 2024-08-26 (yyyy.mm.dd)
Scope (WP & Task)	WP2 and WP3, Task 3.2
Purpose	Dataset from running series production to train and test deep learning model and explainable AI features. Contains images from 8 End-Of-Line Testers and 6 images per product, with up to 8 error classes. Data is collected between May and September 2024.
Utility	BROSE, FHG, UGS
Relation to the project objectives	Objectives 1, 2 and 4
Type/Format	image data as .png, metadata as SQL-database
Volume/size	220 GB raw image data, 5.8 GB training data
Owner	BROSE
Repository	Original data FHG-Repo
Language/s	German
Confidentiality	Confidential
Link (only if public)	Not applicable
License (only if public)	Not applicable
Is it re-used (yes/no)	No
Backup	Original data UGS-Repo
Preservation after the project is finished	By Brose (see contact information) - raw data

Table 3. Dataset 2.

Topic	Description of Data collected/generated
Dataset name	Brose door modules - synthetic dataset - V01
Contact information	Fraunhofer-Institut für Produktionstechnik und Automatisierung IPA Andreas Frommknecht Tel.: +49 711 970 - 1818 E-Mail: andreas.frommknecht@ipa.fraunhofer.de
Source/origin	Brose door modules - production dataset - V01
Level (of SoliDAIR data)	Level 1
Modification date	From 2024-10-01 to 2025-02-28 (yyyy.mm.dd)
Scope (WP & Task)	WP2 and WP3, Task 3.2
Purpose	Synthetic images based on images taken from running series production to train and test deep learning models and explainable AI features. Used to enhance the original image dataset and based on original images from 8 End-Of-Line Testers and 6 images per product, with up to 8 error classes.
Utility	BROSE, FHG, UGS
Relation to the project objectives	Objectives 1, 2 and 4
Type/Format	.png

Volume/size	1.1 GB
Owner	BROSE and FHG
Repository	FHG Compute Cluster
Language/s	German
Confidentiality	Confidential
Link (only if public)	Not applicable
License (only if public)	Not applicable
Is it re-used (yes/no)	No
Backup	FHG-Repo
Preservation after the project is finished	By FHG - synthetic image data

Table 4. Dataset 3.

Topic	Description of Data collected/generated
Dataset name	Brose door modules - latent representation - V01
Contact information	Fraunhofer-Institut für Produktionstechnik und Automatisierung IPA Andreas Frommknecht Tel.: +49 711 970 - 1818 E-Mail: andreas.frommknecht@ipa.fraunhofer.de
Source/origin	Brose door modules - production dataset - V01
Level (of SoliDAIR data)	Level 2
Modification date	From 2025-01-07 to 2025-02-28 (yyyy.mm.dd)
Scope (WP & Task)	WP2 and WP3, Task 3.2
Purpose	Latent representation of the classification results based on the production dataset.
Utility	BROSE, FHG, UGS
Relation to the project objectives	Objectives 1, 2 and 4
Type/Format	.npz
Volume/size	approximately 200 MB
Owner	BROSE and FHG
Repository	Original data FHG-Repo
Language/s	German
Confidentiality	Public
Link (only if public)	The link will be provided once the latent representation of the data has been set up. Additional target platform probably Hugging Faces
License (only if public)	MIT
Is it re-used (yes/no)	No
Backup	Hugging Faces
Preservation after the project is finished	By FHG - latent representation of data and in Zenodo

Table 5. Dataset 4.

Topic	Description of Data collected/generated
Dataset name	HPDC production Dataset
Contact information	Fundación CIE
Source/origin	CIE Vilanova
Level (of SoliDAIR data)	Level 1

Modification date	Upgraded every week
Scope (WP & Task)	WP3
Purpose	Production and final quality data in order to train the machine learning models and optimize the production line.
Utility	Fundación CIE Automotive
Relation to the project objectives	OEE improvement
Type/Format	.xlsx
Volume/size	Each file comprises data related to one calendar week 304 KB
Owner	CIE Automotive
Repository	Original Production Data is confidentially stored at the CIE CV plant server. Post-processed dataset stored at FCIE
Language/s	Spanish
Confidentiality	Original data and post-processed data are confidential. Anonymized data will be public.
Link (only if public)	Not applicable
License (only if public)	Not applicable
Is it re-used (yes/no)	No
Backup	At CIE CV plant (everyday)
Preservation after the project is finished	By Fundación CIE Automotive

For Dataset 4, open publication is planned after data anonymization. The link will be provided once the data is prepared at EU-SoliDAIR-CIE-TWI exchange repository.

Table 6. Dataset 5.

Topic	Description of Data collected/generated
Dataset name	20240207_initial_raw_data_set
Contact information	BOSCH Türkiye
Source/origin	BOSCH injectors production line
Level (of SoliDAIR data)	Level 1
Modification date	2024.01.30
Scope (WP & Task)	WP3, Task 3.4
Purpose	Administrated production data set for data understanding, training and validation purposes
Utility	BOSCH, VIF
Relation to the project objectives	Objectives 1, 2
Type/Format	.zip, including Parquet files
Volume/size	144 MB
Owner	BOSCH Türkiye
Repository	SoliDAIR SharePoint @ VIF /Data/Raw data/20240207_initial_raw_data_set
Language/s	English, German
Confidentiality	Confidential
Link (only if public)	Not applicable
License (only if public)	Not applicable
Is it re-used (yes/no)	No
Backup	Yes. VIF storage spaces

Preservation after the project is finished	By BOSCH Türkiye
---	------------------

Table 7. Dataset 6.

Topic	Description of Data collected/generated
Dataset name	080424_245
Contact information	BOSCH Türkiye
Source/origin	BOSCH injectors production line
Level (of SoliDAIR data)	Level 1
Modification date	2024.04.08
Scope (WP & Task)	WP3, Task 3.4
Purpose	Administrated production data set for data understanding, training and validation purposes
Utility	BOSCH, VIF
Relation to the project objectives	Objectives 1, 2
Type/Format	.zip including Parquet files
Volume/size	434 MB
Owner	BOSCH Türkiye
Repository	SoliDAIR SharePoint @ VIF /Data/Raw data/20240408_
Language/s	English, German
Confidentiality	Confidential
Link (only if public)	Not applicable
License (only if public)	Not applicable
Is it re-used (yes/no)	No
Backup	Yes. VIF storage spaces
Preservation after the project is finished	By BOSCH Türkiye

Table 8. Dataset 7.

Topic	Description of Data collected/generated
Dataset name	EU_Horizon_Projesi_0445111245
Contact information	BOSCH Türkiye
Source/origin	BOSCH injectors production line
Level (of SoliDAIR data)	Level 1
Modification date	2024.09.09
Scope (WP & Task)	WP3, Task 3.4
Purpose	Production data set for data understanding, exploratory data analysis, training and validation purposes
Utility	BOSCH, VIF
Relation to the project objectives	Objectives 1, 2
Type/Format	.xlsx
Volume/size	927.6 MB
Owner	BOSCH Türkiye
Repository	SoliDAIR SharePoint @ VIF /Data/Raw data/20240909
Language/s	English, German
Confidentiality	Original data is confidential. Anonymized data: public

Link (only if public)	Not applicable
License (only if public)	Not applicable
Is it re-used (yes/no)	No
Backup	Yes. VIF storage spaces
Preservation after the project is finished	By BOSCH Türkiye

For Dataset 7, open publication is planned after data anonymization.

Table 9. Dataset 8.

Topic	Description of Data collected/generated
Dataset name	EU_Horizon_Projesi_0445111245_250206
Contact information	BOSCH Türkiye
Source/origin	BOSCH injectors production line
Level (of SoliDAIR data)	Level 1
Modification date	2025.02.06
Scope (WP & Task)	WP3, Task 3.4
Purpose	Administrated production data set for data understanding, training and validation purposes
Utility	BOSCH, VIF
Relation to the project objectives	Objectives 1, 2
Type/Format	.XLSX
Volume/size	741 MB
Owner	BOSCH Türkiye
Repository	SoliDAIR SharePoint @ VIF /Data/Raw data/20250206
Language/s	English, German
Confidentiality	Confidential
Link (only if public)	Not applicable
License (only if public)	Not applicable
Is it re-used (yes/no)	No
Backup	Yes. VIF storage spaces
Preservation after the project is finished	By BOSCH Türkiye

Table 10. Dataset 9.

Topic	Description of Data collected/generated
Dataset name	Process Parameters Gearbox Manufacturing RAW
Contact information	Autforce Automations GmbH
Source/origin	Customer production line
Level (of SoliDAIR data)	Level 1
Modification date	2025.01.13
Scope (WP & Task)	WP3, Task 3.4
Purpose	Administrated production data set for data understanding, training and validation purposes
Utility	Autforce Automations GmbH
Relation to the project objectives	Objectives 1,2
Type/Format	SQL database backup
Volume/size	190 GB

Owner	External customer
Repository	AUTFORCE/00998 AUTFACTORY/111 SolidAIR/data/raw
Language/s	Not applicable
Confidentiality	Confidential
Link (only if public)	Not applicable
License (only if public)	Not applicable
Is it re-used (yes/no)	No
Backup	Yes, AUT storage spaces
Preservation after the project is finished	by Autforce Automations GmbH

Table 11. Dataset 10.

Topic	Description of Data collected/generated
Dataset name	Results EOL Tester Gearbox Manufacturing RAW
Contact information	Autforce Automations GmbH
Source/origin	Customer production line
Level (of SoliDAIR data)	Level 1
Modification date	2025.01.13
Scope (WP & Task)	WP3, Task 3.4
Purpose	Administrated production data set for data understanding, training and validation purposes
Utility	Autforce Automations GmbH
Relation to the project objectives	Objectives 1,2
Type/Format	SQL database backup
Volume/size	20 GB
Owner	External customer
Repository	AUTFORCE/00998 AUTFACTORY/111 SolidAIR/data/raw
Language/s	Not applicable
Confidentiality	Confidential
Link (only if public)	Not applicable
License (only if public)	Not applicable
Is it re-used (yes/no)	No
Backup	Yes, AUT storage spaces
Preservation after the project is finished	by Autforce Automations GmbH

Table 12. Dataset 11.

Topic	Description of Data collected/generated
Dataset name	THL_Data
Contact information	Tech Hive Labs (THL)
Source/origin	THL Innovation hub in robotics and AI, based in Athens
Level (of SoliDAIR data)	Level 2
Modification date	2025.01.28
Scope (WP & Task)	WP2 (Task 2.4), WP3 (Task 3.3), WP4 (Task 4.2)
Purpose	To facilitate the research of AI-enabled robotic inspection on machined parts with complex surface

	geometries. The collected data is used for training AI models and developing inspection solutions.
Utility	CIE, THL, SISW
Relation to the project objectives	Objectives 1,2
Type/Format	original data: .png files, original labels: .txt files
Volume/size	11.2 GB
Owner	Tech Hive Labs
Repository	EU-SoliDAIR-CIE-TWI exchange/THL_Data
Language/s	English
Confidentiality	Non-confidential
Link (only if public)	Link is under preparation
License (only if public)	CC-BY-SA-4.0
Is it re-used (yes/no)	Yes
Backup	Yes, THL storage spaces
Preservation after the project is finished	by Tech Hive Labs and in Zenodo

Table 13. Dataset 12.

Topic	Description of Data collected/generated
Dataset name	VIF_Bearings_Data
Contact information	Virtual Vehicle Research GmbH (VIF) E-Mail: Hannes.Allmaier@v2c2.at
Source/origin	VIF experimental journal bearing tests and simulation
Level (of SoliDAIR data)	Level 2
Modification date	2015.06.03
Scope (WP & Task)	WP2 (Task 2.3)
Purpose	Non-UC-related dataset to support the methodological framework and generic modules development.
Utility	VIF
Relation to the project objectives	Objective 2
Type/Format	Simulation data:.gid. Experimental data: .csv
Volume/size	70 GB
Owner	Virtual Vehicle Research GmbH (VIF)
Repository	VIF internal storage
Language/s	English
Confidentiality	Non-confidential
Link (only if public)	Will be provided soon
License (only if public)	MIT
Is it re-used (yes/no)	No
Backup	Yes, VIF storage
Preservation after the project is finished	By VIF and in Zenodo

3.1 Open data

So far, twelve datasets have been utilized within the SoliDAIR project. While not all of them can be shared publicly in their original form due to confidentiality restrictions, SoliDAIR is participating in the open data initiative by providing several modified subsets from the UCs BRO, CIE, BOS, and AUT, as well as one non-UC-related dataset used for methodology

development. These modified datasets are planned to be made publicly available soon or by the end of the project. A summary of these datasets is presented in Table 14.

To make these datasets available to the public, a community has been created in Zenodo specifically for the SoliDAIR project: <https://zenodo.org/communities/solidair/>.

Since the datasets pertain to different topics, the project partners have proposed additional repositories to enhance the visibility and usability of the data by third parties. This information is also included in Table 14.

Links to the individual datasets, both for Zenodo and any additional repository, will be provided on the SoliDAIR webpage [1].

Table 14. Open datasets.

Dataset	Description	Already published	Additional repository
Brose door modules - latent representation - V01	Latent data representation	No	Hugging Faces
BOS UC (dataset name to be defined)	Anonymized injectors' production data	No	To be defined
AUT – gearbox production dataset	Anonymized production data	No	Company Homepage
FCIE-HPDC production dataset	Anonymized post-processed production dataset	No	EU-SoliDAIR-CIE-TWI exchange
THL_Data	Inspection images used for AI model training	No	EU-SoliDAIR-CIE-TWI exchange/THL_Data
VIF_Bearings_Data	Experimental and simulation data	No	No

4 Fair Data

Fair data management consists of making the research data findable, accessible, interoperable and reusable. To ensure that these principles are followed and implemented as effectively as possible in a project, the European Commission (EU) proposed a Data Management Plan (DMP) template [2], which includes a set of questions designed to help document all the datasets generated and used within the project. This template has been used as a basis for the development of the SoliDAIR DMP.

4.1 Making Data Findable

The findable principle aims to make data easier for both humans and machines to find. To accomplish this target, metadata can be created along with the dataset, and unique identifiers and keywords can be assigned to optimize the search for the data and its reuse.

Currently, in SoliDAIR, most of the used datasets contain confidential information and will, therefore, not be shared publicly. Despite this restriction, metadata was created for each dataset following the Dublin Core guidelines [3], as presented previously in section 3, Data Summary, and is shared and stored solely by the partners directly involved. These data summaries include fields for Link and License, which are currently missing or not updated for all the datasets but reflect the intention of making some of this data, or a subset, available after some modifications for anonymisation.

Furthermore, each UC has provided information for naming conventions used in the data files and any additional standards for metadata creation. The following Table 15 shows the complete questions and answers to the findable data principle.

Table 15. Making data findable: questions and answers.

UC1 – BROSE	<p><u>Metadata provision:</u></p> <ul style="list-style-type: none"> - Metadata is provided as a JSON text file. <p><u>Naming conventions:</u></p> <ul style="list-style-type: none"> - Unique identifier for data points (images), unique identifier links to metadata description. <p><u>Standards for metadata creation:</u></p> <ul style="list-style-type: none"> - JSON, Metadata fields of errors, Device under Test (DuT, including, amongst others, production numbers) and Machine Data by Brose Standards.
UC2 – CIE	<p><u>Metadata provision:</u></p> <ul style="list-style-type: none"> - Data summary available, stored and explained in Word files. <p><u>Naming conventions:</u></p> <ul style="list-style-type: none"> - Name of the project and the reference. - Date of creation written in the file name. <p><u>Standards for metadata creation:</u></p> <ul style="list-style-type: none"> - Metadata is created following the guidelines of the Data Analytics department.
UC3 – BOSCH	<p><u>Metadata provision:</u></p> <ul style="list-style-type: none"> - Data summary available. On the VIF side, the metadata is also stored in the Git project work repository. <p><u>Naming conventions:</u></p> <ul style="list-style-type: none"> - Name of the project. - Creation date at the end of the file name. - Each measurement is stored in a column with a pre-defined standard naming. - There is a data dictionary for the dataset. <p><u>Standards for metadata creation:</u></p> <ul style="list-style-type: none"> - Metadata is created following the guidelines of the data summary table.
UC4 – AUTFORCE	<p><u>Metadata provision:</u></p> <ul style="list-style-type: none"> - Data summary available, stored also in the project work repository. <p><u>Naming conventions:</u></p> <ul style="list-style-type: none"> - The columns of a data set always refer to a UniqueID (Part ID) e.g. "2314567". <p><u>Standards for metadata creation:</u></p> <ul style="list-style-type: none"> - Basic product information, such as product type, will be stored.
Non-UC-related VIF_Bearing_data	<p><u>Metadata provision:</u></p> <ul style="list-style-type: none"> - Metadata is stored as a text file. <p><u>Naming conventions:</u></p> <ul style="list-style-type: none"> - Unique identifiers for each test and simulation run. - Column names in English and measurement units. <p><u>Standards for metadata creation:</u></p> <ul style="list-style-type: none"> - Metadata is created following the guidelines of the data summary table.

4.2 Making Data Accessible

Regardless of whether the created/used dataset is publicly available, clearly defined storage, sharing restrictions and access mechanisms are provided as detailed as possible to enhance information accessibility. Each dataset information includes details about its openness to the public, and if confidentiality is necessary, thoughtful explanations are provided. Additionally, information about the file types, software needed to access the data, and the location of the data repositories, metadata and code are also available.

The following Table 16 shows the complete questions and answers to the accessible data principle for each UC case.

Table 16. Making data accessible: questions and answers.

UC1 – BROSE	<p><u>Is data openly available?</u></p> <ul style="list-style-type: none"> - The data set is split up. The original data and synthetic data will be treated confidentially by the industry partner for data protection reasons. - The latent representation of these datasets will be made public for data scientists to use. <p><u>Methods or tools to access the data:</u></p> <ul style="list-style-type: none"> - Standard tools for image, SQL, JSON and .npz files. - Image data (original and synthetic): .png files. - Metadata: SQL-database (original), JSON (provision). - Latent representation: in .npz. <p><u>Where are the data, metadata and code deposited?</u></p> <ul style="list-style-type: none"> - Original data: FHG Repo. - Latent representation: A link will be provided once the latent representation of the data has been set up; the additional target platform is probably Hugging Faces.
UC2 – CIE	<p><u>Is the data openly available?</u></p> <ul style="list-style-type: none"> - Not the raw data; only colleagues from the manufacturing process quality control and data analytics have access to it. - THL_Data containing collected inspection images using CIE parts are openly available. <p><u>Methods or tools to access the data:</u></p> <ul style="list-style-type: none"> - The data is in .xlsx format, KNIME Analytics Platform and Rapid Miner can be used for data analytics. - THL_data are stored in *.png and *.txt format <p><u>Where are the data, metadata and code repositied?</u></p> <ul style="list-style-type: none"> - Docing/solidairProject - (SharePoint) - (Microsoft Teams) - THL_data is stored in THL repository, SoliDAIR project SharePoint and Zenodo
UC3 – BOSCH	<p><u>Is data openly available?</u></p> <ul style="list-style-type: none"> - Data is confidential due to product know-how. The dataset cannot be shared with the public in its current status and format. Reduced data volume with anonymization is currently under consideration, and this needs to be discussed and aligned with internal procedures. <p><u>Methods or tools to access the data:</u></p> <ul style="list-style-type: none"> - The data is in .xlsx format; therefore, any data processing software (Excel, Python, Matlab). <p><u>Where are the data, metadata and code deposited?</u></p> <ul style="list-style-type: none"> - SoliDAIR SharePoint @ VIF /Data/Raw data/20240909 - GIT repository
UC4 – AUTFORCE	<p><u>Is data openly available?</u></p> <ul style="list-style-type: none"> - As the data comes from a direct customer, it cannot be made available 1:1 and must be anonymised (this has not yet been done). <p><u>Methods or tools to access the data:</u></p> <ul style="list-style-type: none"> - The data is available in an SQL backup and can be processed using Microsoft SQL Server Management Studio. <p><u>Where are the data, metadata and code deposited?</u></p> <ul style="list-style-type: none"> - In the project work repository: AUTFORCE/00998 AUTFACTORY/111 SolidAIR/data/raw
Non-UC-related VIF_Bearing_data	<p><u>Is data openly available?</u></p> <ul style="list-style-type: none"> - Yes, it will be made public soon through Zenodo. <p><u>Methods or tools to access the data:</u></p>

	<ul style="list-style-type: none"> - The simulation data is available in .gid format, and the experimental data in .csv. Both formats are text-based and can be opened with any data processing software (Excel, Python, Matlab). <p><u>Where are the data, metadata and code deposited?</u></p> <ul style="list-style-type: none"> - VIF internal storage services.
--	--

4.3 Making Data Interoperable

This FAIR principle seeks to ensure that data is formatted and structured in a way that it can be used across partners within the project and by other interested parties. To implement this principle, the data should, as best as possible, comply with basic provisions for standard formatting and file types that ensure its use and integration without significant reformatting and conversions. The same interoperable principle applies to the metadata, which should adhere to standard guidelines facilitating the understanding, interpretability and processing of the dataset. In SoliDAIR, in addition to the Dublin Core guidelines for metadata, data used in the UCs follow industry-specific standards for naming conventions, vocabularies, storage, etc., as detailed in the complete questions and answers to the interoperable data principle in Table 17.

Table 17. Making data interoperable: questions and answers.

UC1 – BROSE	<p><u>What data and metadata vocabularies, standards, or methodologies are followed?</u></p> <ul style="list-style-type: none"> - The metadata is stored in a JSON File Format. - DS Number: Name of the machine taking the image. - MachineReadableName/MESTypeNumber: Name of the Type of DuT. - SequenceNumber/ProductionNumber: Serial number of the DuT. - LineIdentifier: Name of the production line producing the DuT. - BroseErrorCodes: <ul style="list-style-type: none"> - BowdenCrossingNok: Bowden does cross each other. - BowdenCrossing_SuM: Bowden beneath glider.
UC2 – CIE	<p><u>What data, metadata vocabularies, standards or methodologies are followed?</u></p> <ul style="list-style-type: none"> - The Cross-industry standard process for data mining (CRISP DM) is followed. - The ETL (extract, transform and load process) guidelines are followed. - THL_data were stored in widely interoperable format *.png and *.txt. The metadata and version tracking functions are provided by the open data platform such as Zenodo
UC3 – BOSCH	<p><u>What data and metadata vocabularies, standards, or methodologies are followed?</u></p> <ul style="list-style-type: none"> - Definitions of abbreviations, column names, codes, etc., used in the data were provided. The created metadata follows the Dublin Core guidelines. - General conventions, among others: <ul style="list-style-type: none"> o ASCII standard (128 characters) for data names. Reserved words from different databases are not used as names. o To avoid confusion between numbers and identifiers, logical numbers and integers must be described with a "No" suffix. - For Metadata management: <ul style="list-style-type: none"> o Data must be catalogued to the scope specified by the Data Domain Owner or Data Owner. o The cataloguing is done in a Data Catalog solution. o The data to be catalogued must be maintained in a Data Catalog with at least the following attributes: <ul style="list-style-type: none"> o Accountable Data Owner o Protection requirements according to protection classes o Technical Metadata to uniquely identify the data in the source system

	<ul style="list-style-type: none"> ○ Semantics of the data, with reference to glossaries or semantic Data Models of the Data Domains. ○ Glossaries of a Data Catalog must be synchronized with the Data Domain glossaries. ○ Data Domain Owners or Data Owners can define further binding attributes within the scope of their responsibilities.
UC4 – AUTFORCE	<u>What data and metadata vocabularies, standards, or methodologies are followed?</u> <ul style="list-style-type: none"> - Definitions of GUID, abbreviations, column names, codes, etc., used in the data were provided. - A data set is always saved based on a UniqueID (traceability) - Data records of add-on parts are merged with the production data.
Non-UC-related VIF_Bearing_data	<u>What data and metadata vocabularies, standards, or methodologies are followed?</u> <ul style="list-style-type: none"> - Measurement units and definitions of abbreviations and column names used in the data were provided. - Each experimental data file and simulation result is stored with a unique file name. - The filename is constructed from the experiment/simulation settings.

4.4 Making Data Reusable

In the SoliDAIR project, provisions are underway to allow for sharing some of the generated and used data within the project or a subset of them while ensuring that confidentiality restrictions are upheld. Once these datasets have been prepared for public access, they will be published in the Zenodo SoliDAIR community, where further details, such as licensing, will be provided. While there are currently no published open datasets, the SoliDAIR project has embraced the reusable FAIR principle as effectively as possible. This has been achieved by creating data summaries with comprehensive information, following standard guidelines for metadata, and implementing a data quality assurance process. These measures ensure that other researchers can accurately understand the data, its context, where to find it, and how to reuse it.

The following Table 18 shows the complete questions and answers to the reusable data principle.

Table 18. Making data reusable: questions and answers.

UC1 – BROSE	<u>Is the data produced and used in the project usable by third parties?</u> <ul style="list-style-type: none"> - The original data is not available for reuse by third parties due to IP rights and commercial restrictions. - A different representation of the data, such as latent activation maps, will be shared and can be reused by third parties. This latent representation of the original and synthetic image data can be used by other data scientists outside of the project consortium for testing and evaluation of their own developed and trained models. <u>Describe the data quality assurance processes:</u> <ul style="list-style-type: none"> - An initial labelling procedure by UGS was done on the data corpus. - Through data cleaning, falsely labelled samples were relabelled. <u>Open data availability, license and usage:</u> <ul style="list-style-type: none"> - Easy usage through MIT license. - Main usage by other data scientists for model evaluation. - No data embargo is needed. - Data will be made available until the middle of 2025. - Remaining time for date reusability: permanent availability through Hugging Faces.
UC2 – CIE	<u>Is the data produced and used in the project usable by third parties?</u>

	<ul style="list-style-type: none"> - No, the original production data, as well as the processed data, are confidential. CIE Vilanova only allows the CIE Data Analytics team to use it. Anonymized data will be shared with third parties - THL_data generated from CIE machine parts are reusable by third parties. <p><u>Describe the data quality assurance processes:</u></p> <ul style="list-style-type: none"> - After accessing the production data, the most relevant variables are filtered (data cleaning through the removal of irrelevant production machine data) - Combining the data from different sources in one dataset using unitary traceability and synthetic data - Different techniques were used to ensure the Machine Learning models were trained properly (e.g. cross-validation) <p><u>Open data availability, license and usage:</u></p> <ul style="list-style-type: none"> - To be defined. - THL_data is public under CC-BY-4.0 license.
UC3 – BOSCH	<p><u>Is the data produced and used in the project usable by third parties?</u></p> <ul style="list-style-type: none"> - The raw data is not available for reuse by third parties due to data protection. Discussion and provisions are underway to prepare an anonymised open dataset. <p><u>Describe the data quality assurance processes:</u></p> <ul style="list-style-type: none"> - Through data cleaning, removing irrelevant manufacturing data. - Summarizing the data from different sources in one dataset using unique identifiers for mapping. <p><u>Open data availability, license and usage:</u></p> <ul style="list-style-type: none"> - To be defined once the subset of anonymized data is created.
UC4 – AUTFORCE	<p><u>Is the data produced and used in the project usable by third parties?</u></p> <ul style="list-style-type: none"> - As the data comes from a direct customer, it cannot be made available 1:1 and must be anonymised. This anonymisation will take place at the end of the project. <p><u>Describe the data quality assurance processes:</u></p> <ul style="list-style-type: none"> - Through data cleaning, removing irrelevant manufacturing data. - Summarizing the data from different sources in one dataset using unique identifiers for mapping. <p><u>Open data availability, license and usage:</u></p> <ul style="list-style-type: none"> - The data will only be made available after the end of the funding project. - The download of the data is only available to authorised users; a licence fee must be paid to gain access. The data may not be passed on to third parties. - As this is a one-off licence fee, the data will be made available indefinitely.
Non-UC-related VIF_Bearing_data	<p><u>Is the data produced and used in the project usable by third parties?</u></p> <ul style="list-style-type: none"> - Yes, it can be used by third parties. <p><u>Describe the data quality assurance processes:</u></p> <ul style="list-style-type: none"> - Before publishing, data will be cleaned from any irrelevant information. <p><u>Open data availability, license and usage:</u></p> <ul style="list-style-type: none"> - The data will be made publicly available soon through the Zenodo repository. - MIT license. - Expected main usage by other data scientists and researchers. - No data embargo is needed.

5 Allocation of Resources

All data from the SoliDAIR project, except confidential information and datasets from the UCs, is stored and shared in the SharePoint space provided by Fraunhofer at <https://fraunhofer.sharepoint.com/sites/EU-SoliDAIR> based on Microsoft SharePoint hosted in

Europe with restricted access. All the stored information will be available for 12 months after the project is finished.

Non-public datasets are stored in the project's internal data storage. Each UC has defined its own dedicated and secure storage to share the necessary data with the UC partners. All costs related to the maintenance of these storage spaces are covered by the partner providing the service.

For the open datasets, Zenodo is a free open-access repository with no fees for uploading and storing the data, providing long-term digital preservation.

Additional information on the allocation of resources is presented in the following Table 19.

Table 19. Additional information on the allocation of resources.

UC1 – BROSE	<p><u>Cost/effort for making data FAIR:</u></p> <ul style="list-style-type: none"> - 5,500 € for Hardware (covered funding of SoliDAIR). <p><u>Responsibilities for data management:</u></p> <ul style="list-style-type: none"> - Data provision: Brose - Data preprocessing: UGS - Data evaluation: IPA - Public data provision: IPA <p><u>Describe the costs and potential value of long-term preservation:</u></p> <ul style="list-style-type: none"> - Costs for long-term preservation: 500 € maintenance per 10 years for non-public available data (own Brose budget), no costs for open data sets by using Hugging Faces. - Value of long-term preservation: <ul style="list-style-type: none"> o Non-public data: for in-house development. o Open data sets: for the scientific community.
UC2 – CIE	<p><u>Cost/effort for making data FAIR:</u></p> <ul style="list-style-type: none"> - The estimated effort for making the data FAIR, once the dataset is anonymized, is about 3 weeks. <p><u>Responsibilities for data management:</u></p> <ul style="list-style-type: none"> - CIE <p><u>Describe the costs and potential value of long-term preservation:</u></p> <ul style="list-style-type: none"> - The cost cannot be estimated as the dataset will be shared through the space provided by the SoliDAIR project.
UC3 – BOSCH	<p><u>Cost/effort for making data FAIR:</u></p> <ul style="list-style-type: none"> - The estimated effort for making the data FAIR, is about 2-3 weeks. <p><u>Responsibilities for data management:</u></p> <ul style="list-style-type: none"> - Bosch and VIF are responsible for the data management of their corresponding data records. - The space for data sharing is provided by VIF at https://v2c2.sharepoint.com/sites/solidair-site. Access is granted through user and password. <p><u>Describe the costs and potential value of long-term preservation:</u></p> <ul style="list-style-type: none"> - The costs and potential value of long-term archiving cannot yet be estimated.
UC4 – AUTFORCE	<p><u>Cost/effort for making data FAIR:</u></p> <ul style="list-style-type: none"> - The estimated effort for making the data FAIR, is about 2 weeks. <p><u>Responsibilities for data management:</u></p> <ul style="list-style-type: none"> - AUT <p><u>Describe the costs and potential value of long-term preservation:</u></p> <ul style="list-style-type: none"> - The costs and potential value of long-term archiving cannot yet be estimated.
Non-UC-related VIF_Bearing_data	<p><u>Cost/effort for making data FAIR:</u></p> <ul style="list-style-type: none"> - The estimated effort for making the data FAIR, is about 2 weeks. <p><u>Responsibilities for data management:</u></p> <ul style="list-style-type: none"> - VIF

	<p><u>Describe the costs and potential value of long-term preservation:</u></p> <ul style="list-style-type: none"> - Costs for long-term preservation: no costs for open data sets using Zenodo. - Value of long-term preservation: <ul style="list-style-type: none"> o In-house research and development. o Open data sets: for the scientific community.
--	--

6 Data Security

Provisions are in place for data security, including secure storage, data recovery and secure transfer of data between the project partners. Detailed information is provided in the following Table 20.

For publicly open datasets, Zenodo provides long-term storage for data based on the CERN's (European Organization for Nuclear Research) robust and redundant storage systems while adhering to the FAIR principles. This also includes assigning a unique Digital Object Identifier (DOI) to ensure the data is accessible and citable.

Table 20. Additional information on data security.

UC1 – BROSE	<p><u>Secure data storage:</u></p> <ul style="list-style-type: none"> - Data recovery/secure storage is done by mirroring data between Brose, UGS and IPA; backups for each entity corresponding to entity rules are in place. - Transfer of sensitive data / sharing data: <ul style="list-style-type: none"> - Hard copy express from Brose to UGS - Transfer UGS to IPA via Microsoft SharePoint - Not encrypted, but access right restrictions in place - Data storage/ access restrictions: <ul style="list-style-type: none"> - Original data: mirroring of data between Brose, UGS and IPA, backups for each entity corresponding to entity rules are in place, and restricted access to SoliDAIR partners. - Public data set: usage of standard data sharing platforms (target platform Hugging Faces), no access restrictions. <p><u>Data encryption:</u></p> <ul style="list-style-type: none"> - Datasets encrypted: not necessary
UC2 – CIE	<p><u>Secure data storage:</u></p> <ul style="list-style-type: none"> - For storage and backup of datasets: The data is stored in BOR-NAS05, which is a 10TB raid hybrid SHR (Synology Hybrid RAID). This information is synchronized with BOR_NAS04 (103TB raid5), which allows the recovery of data in a fast and easy way. - There is a mirrored copy in BOR-NAS03 (87TB raid5). In ARCSERVE-2 (72TB), there is an incremental copy. - There is end-to-end encryption when files are shared outside the plant (to FCIE (Fundacion CIE)) - Data access is restricted. - TISAX (Trusted Information Security Assessment Exchange) standard for data security. <p><u>Data encryption:</u></p> <ul style="list-style-type: none"> - Datasets encrypted: not necessary when data is kept within CIE domain.
UC3 – BOSCH	<p><u>Secure data storage:</u></p> <ul style="list-style-type: none"> - Data is stored on the VIF SharePoint Site, which has restricted access unless permissions are granted. The original data is also saved using VIF's storage services with Active Directory, and there is a daily backup of the data. - Datasets have access restrictions. Contributors have access to the datasets only within the UC.

	<u>Data encryption:</u> - Datasets encrypted: not necessary
UC4 – AUTFORCE	<u>Secure data storage:</u> - Data is stored on our local project storage, which has restricted access unless permissions are granted. Also, there is a daily backup available. <u>Data encryption:</u> - Dataset encryption is not necessary.
Non-UC-related VIF_Bearing_data	<u>Secure data storage:</u> - Data is stored on the VIF internal storage services with Active Directory; there is a daily backup of the data. <u>Data encryption:</u> - Dataset encryption is not necessary.

7 Ethical Aspects

This section addresses both ethical and legal considerations for data usage within the SoliDAIR project.

In the SoliDAIR project, all datasets created and utilized consist solely of manufacturing or production data; no personal data is stored or shared. As a result, privacy and confidentiality regulations concerning personal data do not apply.

Data ownership and intellectual property have been clearly defined for all datasets generated and utilized in the project. This information is detailed in the data summaries found in section 3.

The project partners are actively exploring various options for sharing data while ensuring that confidentiality restrictions are upheld. Currently, some modified versions of the datasets will be made publicly available after they have been anonymized. Once these datasets are ready, they will be shared in the SoliDAIR Zenodo community. Direct links to these repositories will also be provided on the SoliDAIR webpage.

All legal aspects for privacy and sharing of data are handled in the SoliDAIR Project Consortium Agreement.

8 Conclusions

This deliverable D1.4 presents the updated version of the Data Management Plan for the SoliDAIR project. It outlines the strategy for managing the research data used in the project with a focus on collaboration and efficient data sharing. It ensures clear documentation of the datasets, access restrictions, formats, licenses, and software requirements while promoting public data sharing and reuse whenever possible. The SoliDAIR project actively supports the FAIR data principles following the European Commission's Horizon 2020 guidelines for DMP development. In addition to addressing each FAIR principle, this DMP includes sections on resource allocation, data security, and ethics related to the management and handling of data.

While most datasets used in the project contain sensitive or confidential information, with the aim of promoting open science and data sharing, efforts are being made to create anonymized subsets that can be shared publicly. These datasets will be made available in the SoliDAIR Zenodo community, which offers long-term storage, proper metadata documentation and DOI assignment. These datasets will also be promoted through the SoliDAIR project website.

The DMP supports all the research work developed in the SoliDAIR project, and therefore, it has a strong interconnection with all the WPs, particularly with WP2 -Methodological framework & generic modules and WP3 -Use case implementation and deployment. Data management is essential to all tasks within these work packages, as it encompasses the collection, storage, sharing, and handling of data, thereby facilitating effective collaboration among the project partners.

9 Bibliography

- [1] SoliDAIR. [Online]. Available: <https://www.solidair-project.eu/>. [Accessed 26 03 2025].
- [2] European Commission, "Guidelines on FAIR Data Management in Horizon 2020," 26 July 2016. [Online]. Available: https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf. [Accessed 26 03 2025].
- [3] "Using Dublin Core," [Online]. Available: <https://www.dublincore.org/specifications/dublin-core/usageguide/2001-04-12/generic/>. [Accessed 26 03 2025].

10 Acknowledgements and disclaimer

The author(s) would like to thank the partners in the project for their valuable comments on previous drafts and for performing the review.

#	Partner	Partner full name
1	FHG	FRAUNHOFER GESELLSCHAFT ZUR FORDERUNG DER ANGEWANDTEN FORSCHUNG EV
2	BRO-B	BROSE FAHRZEUGTEILE SE & CO. KOMMANDITGESELLSCHAFT
3	CIE	FUNDACION CIE I+D+I
4	BOS	BOSCH SANAYI VE TICARET AS
5	AUT	AUTFORCE AUTOMATIONS-GMBH
6	SISW	AUTFORCE AUTOMATIONS-GMBH
7	UGS	UG SYSTEMS GMBH & CO. KG
8	THL	Tech Hive Labs
9	VIF	VIRTUAL VEHICLE RESEARCH GMBH
10	I2M	I2M UNTERNEHMENSENTWICKLUNG GMBH

LEGAL DISCLAIMER

Copyright ©, all rights reserved. No part of this report may be used, reproduced and/or disclosed, in any form or by any means without the prior written permission of SoliDAIR and the SoliDAIR Consortium. Persons wishing to use the contents of this study (in whole or in part) for purposes other than their personal use are invited to submit a written request to the project coordinator.

The authors of this document have taken any available measure in order for its content to be accurate, consistent and lawful. However, neither the project consortium as a whole nor the individual partners that implicitly or explicitly participated in the creation and publication of this document shall be liable or responsible, in negligence or otherwise, for any loss, damage or expense whatever sustained by any person as a result of the use, in any manner or form, of any knowledge, information or data contained in this document, or due to any inaccuracy, omission or error therein contained.



**Funded by
the European Union**

11 Abbreviations and Definitions

Term	Definition
CERN	European Organization for Nuclear Research
DMP	Data Management Plan
DOI	Digital Object Identifier
DuT	Device under Test
EU	European Commission
FAIR	Findable, Accessible, Interoperable, Reusable
FMU	Functional Mock-up Units
ORE	Open Research Europe
UC	Use Case
WP	Work Package

12 List of Tables

Table 1. Levels of SoliDAIR data.....	6
Table 2. Dataset 1.....	7
Table 3. Dataset 2.....	7
Table 4. Dataset 3.....	8
Table 5. Dataset 4.....	8
Table 6. Dataset 5.....	9
Table 7. Dataset 6.....	10
Table 8. Dataset 7.....	10
Table 9. Dataset 8.....	11
Table 10. Dataset 9.....	11
Table 11. Dataset 10.....	12
Table 12. Dataset 11.....	12
Table 13. Dataset 12.....	13
Table 14. Open datasets.....	14
Table 15. Making data findable: questions and answers.....	15
Table 16. Making data accessible: questions and answers.....	16
Table 17. Making data interoperable: questions and answers.....	17
Table 18. Making data reusable: questions and answers.....	18
Table 19. Additional information on the allocation of resources.....	20
Table 20. Additional information on data security.....	21