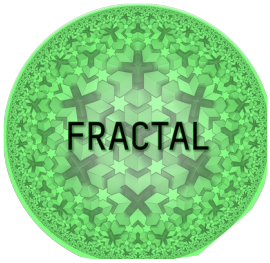

Deliverable D9.8

D9.8 Data Management Plan

Deliverable Id:	**D9.8**
Deliverable Name:	**Data Management Plan**
Status:	**Delivered**
Dissemination Level:	**Public**
Due date of deliverable:	**2021-02-28 (M6)**
Actual submission date:	**2021-02-28**
Work Package:	**9**
Organization name of lead contractor for this deliverable:	**ZYLK**
Author(s):	**Alfonso González (ZYLK), Iñigo Angulo (ZYLK), Ester Sola (ZYLK), Ander Galisteo (IKERLAN), Juan Manuel Besga (IKERLAN), Iñaki Paz (LKS), Juan José Rodríguez (LKS)**
Partner(s) contributing:	**ALL PARTNERS Thales, Siegen (reviewers)**

Abstract: This document presents the updated (and last) version of the Data Management Plan for the FRACTAL project at M32. It provides a detailed description of the data used by all partners during the project. Management of the complete data life-cycle is considered, from the collection and generation of data and metadata to the stages of storage, deletion and reuse. To ensure a correct data management, we base the DMP on the FAIR principles promoted by HORIZON 2020, providing considerations about having findable, accessible, interoperable and re-usable data. This document is the last of three versions, the first one at the beginning of the project, first update in M20 and final version in M32.

Co-funded by the Horizon 2020 Programme of the European Union under grant agreement No 877056.

This project has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 877056. The JU receives support from the European Union's Horizon 2020 research and innovation programme and Spain, Italy, Austria, Germany, Finland and Switzerland.

Contents

1 History

Version	Date	Modification reason	Modified by
V0.1	2020/11/18	Document creation	zylk
V0.2	2021/02/15	Document ready for review	zylk
V1.0	2021/02/28	Final document delivered in M6	zylk
V1.1	2022/03/23	1st update delivered in M20	zylk
V1.2	2023/04/30	2nd and final update delivered in M32	zylk

2 Summary

In line with promoting "Open Science", Horizon 2020 projects are required to develop a Data Management Plan (DMP) to specify how research data will be handled both during and after the project.

This report presents the latest version of the DMP for the FRACTAL project. Since the DMP is intended to be a living document, updates on Data Management have been provided as the implementation of the project progresses and when significant changes have occured. V1.0 of this document (see Section 1, History) corresponded to the first delivered version at M6, v1.1 corresponded to the first update of this document at M20 for the second reporting period, and v1.2 corresponded to the final version of the document, delivered in M32, reaching the final stages of the project.

The main goal of this document is to provide the DMP of the FRACTAL project, specifying all data used in the project, both collected and generated during the research, which data will have an open access, how they will be preserved and how they will be made accessible and reusable. Guidelines provided by the European Commission on FAIR Data Management in Horizon 2020 manual have been followed for the generation of this document (available on the Participant Portal[1]).

All partners involved in the Use Cases and research activities have contributed in the development of this deliverable by providing information about the data collected and/or generated throughout the project.

In the first version (v1.0), we reported an initial analysis on how the amount of data produced in the project is intended to be managed. With this aim, we elaborated a template to identify first the data sets that will be produced by each partner, and then, to define the life-cycle of those data sets. Some questions regarding FAIR data, costs, or IPR will be answered in following versions.

As a starting point, we presented the following elements of the Data Management Plan:

1 Introduction of the document in the context of the FRACTAL project

2 Guidelines for the generation of the DMP. A data summary is presented: Type of data generated, collected and stored, and relevant data features description.

3 Data security, data sharing and specifications on security aspects.

4 Ethical and legal aspects regarding data.

An update on the document has been done during M20 (v1.1), where a revisit of these elements and any new information related to data generation, aggregation, processing, security and storage is provided by the partners.

A second (and final) update on the document has been presented during M32 (v1.2), specifying any new information related to data generation, aggregation, processing, security and storage is provided by the partners, and providing the final details and aspects of the project data.

1 https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm

2.1 Achievements

2.1.1 Highlights

- *We have got an overview of all data used during the project, in a clear and useful way.*
- *A complete description of the datasets and their current and future treatment has been provided.*
- *The DMP contributes to the correct use and treatment of all data by all partners, using the FAIR principles promoted by HORIZON 2020*

2.1.2 Lowlights

- *Gathering all the information from so many partners has been a challenge.*
- *The fact that the partners have been collaborating with each other through the whole project made it difficult to specify which data aspects were related to which partner.*
- *The type of data, volume and other aspects were totally undefined at the beginning, but regular updates have made it possible to complete the information.*

2.1.3 Results or novelties

This document is not expected to present any novelties as it is not a research related document.

Instead, the main result of this document has been to provide a complete description of the datasets used by each partner, defining the data generation, storage, and usage processes in each of the Use Cases.

Information about the methodologies to be followed, common repositories, and shared collaboration spaces (IKERLAN's Sharepoint, GitHub repository) has been provided, as well as the details of what actions to take in the future to make the relevant data of the project findable, accessible, interoperable and re-usable.

3 Introduction

This document corresponds to the deliverable D9.8, belonging to the framework of WP9 (Exploitation, Training, Dissemination & Standardization), and describes the Data Management Plan (DMP) of the FRACTAL project. DMP is a plan that outlines what data will be collected and generated and how to ensure having well-managed data and FAIR data: Findable (access, storage, backup, …), Accessible, Interoperable and Reusable Data.

The purpose of this DMP is to set standard procedures and formats to be followed by the partners involved in each of the use cases, while defining platforms and repositories for data to be shared and/or stored. Moreover, the living aspect of the document makes it possible for partners to update any relevant or missing information that was not accessible at the time the document was created, resulting in a more flexible and complete Data Management Plan.

3.1 FRACTAL Project Overview

An overview of the FRACTAL project is presented in this section, so that this document can be aligned with the project's objectives and evolve regarding the necessities that may emerge related to data and their management.

The FRACTAL project's main objective is to develop a computing node based on novel Edge computing technologies. The computing node will be comprised by Internet of Things (IoT) devices that will be able to interoperate with each other, giving rise to independent and larger nodes.

The whole architecture must be designed such that the performance of every single device (or, once scaled, computing node) is supervised. Each of the nodes may be able to learn from other nodes, combine within each other to form new independent nodes (fractality), and ultimately improve the overall performance of the system.

From this fractal network architecture, a whole new range of enhanced capabilities emerge, supported by the edge computing paradigm, which allows lower latencies, faster prevision, premature alarm-tracking, and scalability, while also providing a more secure environment.

3.2 FRACTAL Data Life-cycle

Characteristics and features of all the data generated, collected, processed, analyzed, stored, and/or eliminated during any of the data life-cycle stages must be defined. Herein, a workflow to be followed to make data FAIR (Findable, Accessible, Interoperable and Reusable) and specifications about data preservation (or elimination) are defined.

The management of research data is a fundamental part of any research process in order to provide a clear overview of the data life cycle and to facilitate accessing and reusing data, not only during the project, but also in the future.

The data lifecycle of a project includes six phases: creation of data, data storage (repositories), use of data, sharing (between users, partners, or external researchers), archive (whenever data are stored for future activities) and destruction of data (if necessary). See Figure 1.

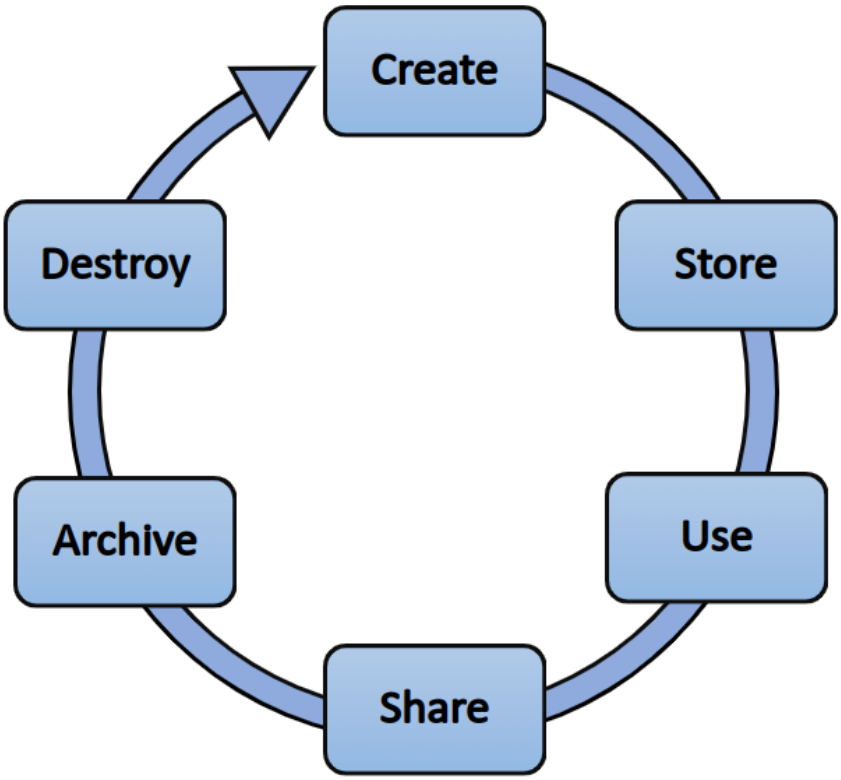

Figure 1: Data Life-cycle

In the case of FRACTAL, we must address data management during every life-cycle stage: (1) from the collecting data point on the edge (See Use Cases), (2) the generation, modification, or cleaning of data during its processing (FRACTAL platform), and (3) the storage of relevant data and metadata to promote sharing and reusing data for the aim of collective intelligence.

As far as it concerns use cases, different data types must be dealt with depending on the individual requirements of each use case. This fact makes it necessary for different processing and storing tools to be available and defined, as well as what file formats to be used.

An efficient Data Management must go deeper in this context and also define other data characteristics, *i.e.*, container architecture (if necessary), usage of standard file

formats, data loss risks, among others, to ensure that full comprehension and management of data are accomplished during the project.

4 FRACTAL Data Management Plan Generation

Each of the partners involved in the FRACTAL project has collaborated in order to fulfill and update the DMP document along the project. Keeping the document updated has been a key aspect in this regard, as some relevant information about data were not available nor well defined at the start, but in further stages of the project. The final update of the DMP has been delivered in M32.

Data sharing and collaboration between partners has also been of key importance to achieve the objectives of the project. For this reason, research data and metadata obtained as outputs from the use cases must be stored and shared among researchers in order to ensure that a more robust and agile research is performed.

Due to the number of partners participating in the project, some guidelines are defined so the DMP completion procedure is standardized. Large data sets will be gathered and stored according to national and European legislation frameworks and standards. The guiding principle must be the Horizon 2020 effort to create re-usable datasets for sustainable, comparable, and growingly valid and reliable research outcomes.

Given the variety of data types involved in each of the Use Cases, a common table format has been prepared that is filled by all partners, including information about all data collected and generated throughout the project (Context Data, Research Data and Aggregated Data), as well as the FAIR Data issues. These tables are included in Annex I.

For each partner, a set of questions are presented (See Annex I), related to some general aspects of the data to be used. After completion of the project, and as a guide for partners, the following questions must have been answered:

- What is the purpose of the data collection/generation and its relation to the objectives of the project?
- What types and formats of data will the project generate/collect?
- Will you re-use any existing data and how?
- What is the origin of the data?
- What is the expected size of the data?
- To whom might it be useful ('data utility')?

Additionally, a second table has been filled by each partner. The purpose of this second table is to specify the steps followed to make the data FAIR. More information about FAIR data can be found in section 2.2.

These tables are meant to ensure an efficient management and governance of the data, and set common standards, formats and requirements for all partners to use. The partners are asked to try to use common, open, and widely known data formats, as well as facilitate interoperability between external datasets from other projects. Notice that other data aspects like the ownership of the data generated during the project has been updated during the advancement of the project.

Lastly, the partners may recall that FRACTAL operates in the open data pilot[2]. This means that data generated and collected should be shared and openly accessible by researchers, through open access repositories.

In this regard, a common repository for data to be shared between partners during the project is available, using Teams and GitHub platforms (see Section 5 - Allocation of Resources). Partners can freely decide what data, code or information to share, including the possibility to keep data for themselves.

4.1 Data Collection and Generation

Each of the Use Cases in the FRACTAL project has a common pathway for data treatment. The data workflow is based on data generation and collection from a source at the edge, these data are later processed and stored in the edge nodes comprising the FRACTAL structure, and analyzed by means of AI algorithms, from whom output a response is made (the response nature depends on the Use Case scenario). On the other part, the cloud can be accessed and used for storage purposes, but most of the data life-cycle occurs at the edge.

There is still a very wide spectrum of data types involved and the purpose of each data type may vary from one to another. For simplicity, they have been classified in three types: Context Data, Research Data and Aggregated Data.

- **Context Data:** These refer to any data obtained during measurements in the Use Case environment, i.e., environment specifications (limitations, necessities, presence of humans in the scenario, obstacles, etc), and they are used to obtain any kind of information from the environment. Context data can be collected, generated or retrieved from already existing data. The available dataset is specified for each partner in Annex I.

- **Research Data:** Any data that is collected during the course of the study, for research purposes. They also include any derived data that may result from the direct transformation of the original data (by analysis, fusion, filtering…).

- **Aggregated Data:** Data created in order to answer research questions. They are generated from data reduction processes, which mainly summarize the most important aspects in the raw data into relevant parameters that

2ttps://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

can be used for analysis. These reduction processes can be validation, curation, conversion, etc.

4.2 FAIR Data

This document should ensure that the partners follow the "FAIR Guiding Principles for scientific data management and stewardship"[3]. These principles set up the standards for data to be easily findable, accessible, interoperable and reusable not only by humans, but also for computational systems.

These FAIR Guiding Principles are presented below in Table 1 , as they were published in the original source:

The FAIR Data Guiding Principles
To be Findable
F1. (meta)data are assigned a globally unique and persistent identifier F2. data are described with rich metadata (defined by R1 below) F3. metadata clearly and explicitly include the identifier of the data it describes F4. (meta)data are registered or indexed in a searchable resource
To be Accessible
A1. (meta)data are retrievable by their identifier using a standardized communications protocol A1.1 the protocol is open, free, and universally implementable A1.2 the protocol allows for an authentication and authorization procedure,where necessary A2. metadata are accessible, even when the data are no longer available
To be Interoperable
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation I2. (meta)data use vocabularies that follow FAIR principles I3. (meta)data include qualified references to other (meta)data
To be reusable
R1. (meta)data are richly described with a plurality of accurate and relevant attributes

3Wilkinson, M.D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., … Bourne, P.E. (2016). *The FAIR Guiding Principles for scientific data management and stewardship.* Scientific Data, 3, 160018, doi: 10.2038/sdata.2016.18

R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

Table 1 - FAIR principles

These FAIR principles provide a definition of the practices and methods that data resources and infrastructures should follow in order to support discovery and research.

As said before, in this final version of the DMP the information regarding FAIR data is provided.

In Annex I, information from each partner regarding the FAIR data specifications can be found.

Upon completion of the tables in Annex I, the following questions must have been answered:

Findability: 'Are the data produced and/or used in the project discoverable with metadata, identifiable and locatable by means of a standard identification mechanism?', and 'What metadata will be created?'

Accessibility: 'Which data produced and/or used in the project will be made openly available as the default? ', and 'How will the data be made accessible?'

Interoperability: 'Are the data produced in the project interoperable?'

Re-usability: 'How will the data be licensed to permit the widest re-use possible?'

Update on FAIR principles questions to be answered:

By the end of the project, Data which have been generated and collected from the Use Cases must have been treated in a way that the FAIR principles are fulfilled. To this extent, the following questions have been answered after revision of the information provided in Annex I:

Findability:

'What are your naming conventions?'

'Are search keywords available to optimize possibilities for re-use?'

'Are dataset version numbers available?'

Accessibility:

'What methods or software tools are needed to access the data?'

'Is it possible to include the relevant software (e.g., in open-source code)?'

'Where will the data and associated metadata, documentation and code be deposited? Preference should be given to certified repositories which support open access where possible.'

'If there are restrictions on use, how will access be provided?'

'Are there well described conditions for access (i.e., a machine-readable license)?'

'How will the identity of the person accessing the data be ascertained?'

Interoperability:

'What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?'

Re-usability:

'When will the data be made available for re-use?'

'Are data quality assurance processes described?'

5 Allocation of resources

In the FRACTAL Project we distinguish two types of information, the project management information, e. g. documents, meeting minutes, deliverables, templates, dissemination material, etc.; and the software developed, use case data and test data. Given the characteristics of each type of information, each of them has been allocated in a different repository.

The IKERLAN's Microsoft Teams and Sharepoint corporate repository, is the tool that is being used as a private area for preparing and sharing documents and knowledge among the members of the FRACTAL Project. IKERLAN provided guest access, to the individual members of FRACTAL partners. WP leaders and use case leaders are free to create specific working spaces to deal with their day-by-day activities. This repository is also used for backups or for archiving of documents.

As mentioned before, it is an IKERLAN private repository, so it doesn't involve any cost to the project. It is managed by IKERLAN, and the costs derived from this activity are already considered.

For hosting the developed software and hardware (source code), use cases data and test data, GitHub a collaborative development platform based on Git, an open-source distributed version control system, has been used.

GitHub has been selected as the code repository for the project because it is the most widely used repository, offers secure cloud storage, and is free for teams, offering unlimited public and private repositories for unlimited collaborators.

Regarding hardware and software integration two possible scenarios are considered:

- Integration from partners that use their own infrastructure. In this case, the basic idea is that at the end of the development iteration, they upload their changes to the centralised GitHub, for further integration with other partners development.

- Integration from partners that do not user their own infrastructure. They can use the centralized GitHub to host their day-to-day Git repositories (and commit to the centralised Git repository).

As Git is free and GitHub is a free online service, it won't involve any cost to the project. Any costs related with the management of the repository are included in the efforts of the project. This repository has been managed and maintained by LKS.

The use of free platforms such as GitHub reduces the cost of long-term preservation to zero. Moreover, since Git is a distributed version control system each user that clone for their own use the data repository has a complete copy of the data set (history, changes, meta-data, etc.).

6 Data Security

As said in previous sections, the two main tools for sharing data are:

- Microsoft Teams and Sharepoint corporate repository for documentation related with FRACTAL.
- GitHub for the code required for some of the work packages.

One of the most important concepts in security is that there is no completely secure system. It is simply not possible to prevent any current and future attack. If a new way to attack a system is found in the future, a system designed in the past can do nothing against it. Nevertheless, what a system can do is to have a state-of-the-art security when made public and allow to be updated when new vulnerabilities are found before it affects its users. And this is what our solutions do.

According to Microsoft: “Microsoft Teams is designed and developed in compliance with the Microsoft Trustworthy Computing Security Development Lifecycle (SDL), which is described at Microsoft Security Development Lifecycle (SDL). The first step in creating a more secure unified communications system was to design threat models and test each feature as it was designed. Multiple security-related improvements were built into the coding process and practices. Build-time tools detect buffer overruns and other potential security threats before the code is checked in to the final product. “

For enterprise customers, Microsoft stores the data from their 365 services, like Teams or Sharepoint, in datacenters nearest to the business location provided when the company creates its tenant.

In case of tenants created with a billing address in any European Union country, the datacenters are located in the European Union under European laws.

GitHub also has extremely high security standards. Their solution is GDPR compliant, they encrypt all the data in transit and store the hash of all passwords using bcrypt, one of the most used hashing solutions in the industry. It also prevents bruteforce attacks by limiting the number of attempts before locking.

One of the advantages of using these solutions is that both are prepared to handle multiple teams geographically separated and concurrent working. For this reason, all the information is secured in a third-party storage, without requiring trust among partners. As said before, all the information is end-to-end encrypted, assuring privacy.

All the data in GitHub has redundancy and history, meaning that losing all the data at once is really unlikely and changes can be tracked among partners.

Also, Microsoft Teams corporate not only integrates Office365 in their solution, but also has redundancy in the backend, making the data accessible even if a given server fails.

Obviously, as the data is stored at a trusted third party, it has the required trusted certificates. Both solutions promise long term preservation and as the biggest players in each of their segments of the markets, they are trusted by millions.

7 Ethical Aspects

Data sharing and the Open Access implementation to the scientific research process is clearly enhancing scientific progress, while also benefiting transparency and reproducibility. However, this also questions the implications and impact that making specific kinds of data public and accessible may have. Several discussions have taken place on this respect, however a general consensus in the scientific community has not yet been reached.

This chapter describes what impact the data in the FRACTAL project will have on ethical aspects, specifically when these data are related to personal information. The partners involved in the project may be aware that neither during or after the project, any of the data shared/stored will have an impact on ethical, gender, or other personal circumstances.

To this regard, the partners must follow a set of practices while involved in data treatment:

- Ensuring that the citizens' rights on their own data are protected.
- Confidentiality and privacy must be managed so that shared datasets are anonymized and people cannot be identified from shared data.
- Consent from people whom data are being obtained from must be validated, keeping clear the objective and purpose of the data to be obtained from them.
- Analyze the potential impact and implications on further research, in a way that further results obtained from shared data can be validated and trustworthy.

As a rule of thumb, partners and participants shall follow the statement: "**All participants in the FRACTAL project conform to GDPR and the current legislation and regulations in the countries where the research is carried out**"

Partners must ensure that the ethical and societal aspects are incorporated into the design process from the beginning. They must also allow security and validation of operational efficiency and transparency, possible interpretative distortions, liability for possible damage, patentability, anonymization of user data and privacy in compliance with the guidelines developed by the European Commission Expert Panel[4] that has developed and continues to develop guidelines for the design of reliable artificial intelligence systems, respecting the centrality of the human being.

Additional information on each of the individual Use Cases can be found below. This information accounts for any sensitive data that may be involved in any of the UCs, either during research, development, or storage of data for further purposes, and is only available for those UCs that deal with data that may have any ethical impact:

4https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52011DC0882&from=EN

- UC1: Demonstrator 2 will use personal information on construction site workers. The data should be anonymized and associated only with the device worn on their body, but an important part of the analysis of the pilot results is the detection of workers who are unprepared for a job, either as a field worker or a machine operator. In addition, the device will record their position at all times within the site, so that it is possible to know when they are moving through a machinery run-over danger zone. This may be seen as an infringement of workers' rights, both legally and ethically, but it should be clear that the sole purpose of the pilot is to improve the site workers' own safety, not their control within the site.

- UC6: As described in D2.1, UC6 demonstrator will collect and use personal information.

  UC6 partners will ensure compliance of the UE 679/2016 of the European Parliament and of the Council which entered into force from May 2018 on the General Data Protection Regulation (GDPR) of individuals with regard to the processing of personal data and on the free movement of such data.

  Collected data at all stages will be treated with respect to personal anonymity. Only when necessary, participants' personal data will be legally obtained after an informed consent. Data will be securely accessed and privacy protection measures will be undertaken proportionally to the risks involved and the sensitivity of the data, such as password protection, encryption in all transmissions, etc. Identification data will be encrypted and strictly separated from sensitive data such as health data.

  Users' personal data will be collected for processing, and undergo such processing, only if they are adequate, relevant and not excessive in relation to the scope and the specified, explicit and legitimate purposes for which they will be obtained. Also, personal data subjected to processing will not be used for purposes incompatible with those for which they will be collected. Further processing of the data for historical, statistical or scientific purposes shall not be considered incompatible.

  In addition, personal data will be erased when they will have ceased to be necessary or relevant for the purpose for which they were obtained or recorded. They will not be kept in any form that permits identification of the data subject for longer than strictly necessary. On a regular basis, it will be decided to keep the entire set of particular data, in accordance with the specific legislation, because of their historical, statistical or scientific value.

  The researchers will ensure that outcomes will be reported and will not contravene the right to privacy and data protection. They also will carefully evaluate and report the personal privacy implications of the intended use or potential use of the research outcomes.

8 List of figures

9 List of tables

10 List of Abbreviations

AI	Artificial Intelligence
DMP	Data Management Plan
FAIR	Findable, Accessible, Interoperable and Re-usable
GDPR	General Data Protection Regulation
IPR	Intellectual Property Rights
SDL	Security Development Lifecycle
UC	Use Case
WP	Work Package

Annex I

1-IKER

Table 2 - IKER

1- IKER			
DATA SETS	**CONTEXT DATA**	**RESEARCH DATA**	**AGGREGATED DATA**
Description of data types	N. A.	Raw data-sets, processed datasets, images & videos data for Cloud Platform modules testing purposes and Use Cases Implementation	Possible mean, median, etc. calculation from raw measurements in Cloud Platform's Data Transfromation module testing and Use Cases Implementation
Purpose of the data collection	N. A.	Cloud Platform modules testing: Data Ingestion, Raw Data & Object repostories, Data workflow, etc.	Cloud Platform processed dataset upload, storage, workflow and deployment tests
Relation to the objectives of the project	N. A.	Test the developments in WP3, WP4, WP5 and WP6, UCs implementation	Test the developments in WP3, WP4, WP5 and WP6, UCs implementation
Formats of data generated/collected.	N. A.	json, RAW, csv, png, jpg, avi	json, csv
Data that will be re-used (if any)	N. A.	For Cloud Platform testing purposes, Use Cases test data and public data can be used. The public data will also be used for training purposes. In Use Cases implementation, data provided by the Use Cases will be used	For Cloud Platform testing purposes, Use Cases test data and public data can be used. The public data will also be used for training purposes. In Use Cases implementation, data provided by the Use Cases will be used
Origin of the data	N. A.	Use Cases data, public repositories	Use Cases data, public repositories
Volumetry: Expected size of the data (if known) and sampled frequency	N. A.	frequency, in Cloud Platform model inference will not be performed. In some cases the training of these models will be performed on demand or according to the strategy defined by the UC, and	frequency, in Cloud Platform model inference will not be performed. In some cases the training of these models will be performed on demand or according to the strategy defined by the UC, and
Data utility: to whom will it be useful	N. A.	The Cloud Platform testing data will be used internally in WP5 for testig and also will be used externally por training purposes	The Cloud Platform testing data will be used internally in WP5 for testig and also will be used externally por training purposes

FAIR DATA	
FINDABILITY	
Making data **findable**, including provisions for metadata	Data used for Cloud Platform modules testing will be stored in Cloud repositories and/or in Fractal's GitHub repository
Naming convention for the data	Not defined
Search keywords	Not defined
Dataset version control	For datasets used in Coud Platform testing, the version control will be provided by tools used in Cloud Platform repositories (MinIO, lakeFS, OVH Cloud Object Store) and by he Fractal GitHub Repository
ACCESSIBILITY	
Making data openly **accessible**	Data provided by Use Cases for Cloud Platform testing purposes will be used internally and won't be openly accesible. Public data samples used for Cloud Platform Modules testing and training purposes will be openly accesible as they will come from public repositories.
Software tools required to access data (database querying, web services...)	Browser, Services provided by FRACTAL's Cloud Platform
Are these accesibility tools open-source?	Yes
Use restrictions (if any) and access granting	No
Are the data or code licensed?	Depends on the Use Case
Authentication and authorization to repositories	No, when accesing to public repositories data. Yes, when accesing to testing and training data stored in Fractal Cloud Platform repositories
INTEROPERABILITY	
Making data **interoperable**	Data will be interoperable, through the use of standard formats
Vocabularies, standards and methodologies	
RE-USABILITY	
Increased data **re-use** (through clarifying licenses)	The data provided by the Use Cases for Cloud Platform testing purposes and new data created from these, through Data Transformation services, will be used internally. The public data used for Cloud Platform testing purposes would be used also for training purposes in Use Cases implementation
When will the data be available?	The public data to be used for testing the Cloud Platform modules and training will be data from public repositories that are already accessible. These data also will be available through Fractal Cloud Platform repositories, aftere the preparation of training material
How is data quality assured?	The public data to be used for testing the Cloud Platform modules and training will be data widely used by development community. In Use Cases implementation, the data quality will be monitored by the data suppliers

2- BSC

Table 3 - BSC

2- BSC			
DATA SETS	**CONTEXT DATA**	**RESEARCH DATA**	**AGGREGATED DATA**
Description of data types	N. A.	Raw measurements (e.g. execution cycles, cache misses, and the like)	Mean, median, etc. of raw measurements
Purpose of the data collection	N. A.	Validation and internal evaluation of BSC's technical contributions in WP3 and WP4	Validation and internal evaluation of BSC's technical contributions in WP3 and WP4
Relation to the objectives of the project	N. A.	Validate contributions before sharing them with partners that need to use them	Validate contributions before sharing them with partners that need to use them
Formats of data generated/collected.	N. A.	Raw data normally managed in spreadsheets and plain text files	Aggregated data normally managed in spreadsheets and as part of papers
Data that will be re-used (if any)	N. A.	None (for now)	None (for now)
Origin of the data	N. A.	Own experiments on a local FPGA at BSC	Own experiments on a local FPGA at BSC
Volumetry: Expected size of the data (if known) and sampled frequency	N. A.	Typically few KBs	Typically few KBs
Data utility: to whom will it be useful	N. A.	For UC7 owner as supporting evidence to adopt the corresponding technologies, and to publish scientific contributions	For UC7 owner as supporting evidence to adopt the corresponding technologies, and to publish scientific contributions

FAIR DATA	
FINDABILITY	
Making data **findable**, including provisions for metadata	Does not apply. Data only useful for internal validation. Relevant data to be shared is the output of the UC, which is not produced by BSC. At most, data will be included in conference/journal papers
Naming convention for the data	
Search keywords	
Dataset version control	
ACCESSIBILITY	
Making data openly **accessible**	Does not apply. Data only useful for internal validation. Relevant data to be shared is the output of the UC, which is not produced by BSC. At most, data will be included in conference/journal papers
Software tools required to access data (database querying, web services...)	
Are these accesibility tools open-source?	
Use restrictions (if any) and access granting	
Are the data or code licensed?	
Authentication and authorization to repositories	
INTEROPERABILITY	
Making data **interoperable**	Does not apply. Data only useful for internal validation. Relevant data to be shared is the output of the UC, which is not produced by BSC. At most, data will be included in conference/journal papers
Vocabularies, standards and methodologies	
RE-USABILITY	
Increased data **re-use** (through clarifying licenses)	Does not apply. Data only useful for internal validation. Relevant data to be shared is the output of the UC, which is not produced by BSC. At most, data will be included in conference/journal papers
When will the data be available?	
How is data quality assured?	

3-UPV

Table 4 - UPV

3- UPV			
DATA SETS	**CONTEXT DATA**	**RESEARCH DATA**	**AGGREGATED DATA**
Description of data types	N. A.	Collection of performance statistics of benchmarks and applications executing in the platform. In the context of AI applications and benchmarks, open data sets for input data to validate correct behaviour of the system.	Mean, median, average of raw data values
Purpose of the data collection	N. A.	Performance assessment, verification and validation.	Performance assessment, verification and validation.
Relation to the objectives of the project	N. A.	Performance assessment, verification and validation of the NOEL-V platform FRACTAL developments	Performance assessment, verification and validation of the NOEL-V platform FRACTAL developments
Formats of data generated/collected.	N. A.	Raw data	Spreadsheets
Data that will be re-used (if any)	N. A.	Open-source data bases for testing purposes (COCO, MNIST, ...)	N.A
Origin of the data	N. A.	Open-source for AI workloads and self-generated for the statistics and performance metrics	Process of the collected data in our own computer setups
Volumetry: Expected size of the data (if known) and sampled frequency	N. A.	No more than Megabytes. At the granularity of one execution period, ms frequency for control applications. Not to be sampled periodically within the context of the project.	Few KBs.
Data utility: to whom will it be useful	N. A.	Partners in WP3 and WP4 for validation purposes and UC7 in WP8.	Partners in WP3 and WP4 for validation purposes and UC7 in WP8.

FAIR DATA	
FINDABILITY	
Making data **findable**, including provisions for metadata	Data is useful for internal validation purposes. Scripts to generate validation data will be provided in the same public repositories the tools are released.
Naming convention for the data	Not applicable
Search keywords	Not applicable
Dataset version control	Not applicable
ACCESSIBILITY	
Making data openly **accessible**	Data is useful for internal validation purposes. Scripts to generate validation data will be provided in the same public repositories the tools are released.
Software tools required to access data (database querying, web services...)	Plain text editors and excel tables readers
Are these accesibility tools open-source?	Open-source tools are available for all documents generated
Use restrictions (if any) and access granting	N/A
Are the data or code licensed?	Open-source code is licensed with MIT
Authentication and authorization to repositories	Only for private tools, open-source ones (the majority) have unrestricted access
INTEROPERABILITY	
Making data **interoperable**	Not applicable
Vocabularies, standards and methodologies	Not applicable
RE-USABILITY	
Increased data **re-use** (through clarifying licenses)	Data is useful for internal validation purposes. Scripts to generate validation data will be provided in the same public repositories the tools are released.
When will the data be available?	It can already be generated
How is data quality assured?	Data is only used for internal validation. N/A

4- PROI

Table 5 - PROI

DATA SETS	CONTEXT DATA	RESEARCH DATA	AGGREGATED DATA
Description of data types	Data used for UC1, specifically: Demo 1: Documental information about the most common types of cracks and fissures in concrete Documental information about the depth of cracks and their relationship to safety Opensource archive images and others used in other projects. Demo 2: Construction site accident statistics Type of machinery most involved in accidents Machinery operating speeds	It will be obtained from the UC1 through the two demonstrators. In demonstrator 1, data will be collected through the images taken by the drone, in relation to the surface condition of a given structure, bridge or viaduct. In demonstrator 2, information will be obtained from the sensor network deployed in a real work environment. The sensor network data, consisting of (alphanumeric) data, PositionsXY, timestamp, EntryZone and Exitzone, among others	Demo 1: Join the two datasets + Image augmentation. Demo 2: Data will be augmented by dataset completion techniques, keeping the same format and fields.
Purpose of the data collection	To be able to respond to a problem widely spread among construction sites in relation to the safety of work in the field. To obtain the context information that helps to provide a solution through the technology developed at UC1. To create a state of the art that serves as a technological starting point for the use case. To obtain a dataset to make possible to train AI models	It is expected to create a group of datasets from the above-mentioned information sources, and taken in different periods of time, within the UC1. In this way, this information will be exploited and experimented to develop the fractal node modules.	NA (Demo 1) Train models with better accuracy and bigger datasets (Demo 2)

DATA SETS	CONTEXT DATA	RESEARCH DATA	AGGREGATED DATA
Relation to the objectives of the project	Developing real-time mechanisms and AI techniques based on sensor and image processing. (Demo 1) NA (Demo 2)	Developing real-time mechanisms and AI techniques based on sensor and image processing.	NA (Demo 1) Developing real-time mechanisms and AI techniques based on sensor and image processing.(Demo 2)
Formats of data generated/collected.	.doc and .docx - Microsoft Word file. .odt - OpenOffice Writer document file. .pdf - PDF file. .jpg – Images from the dataset and from the drone acquisition	json and .jpg (Demo 1) CVS (Demo 2)	.json & .jpg (Demo 1) CVS (Demo 2)
Data that will be re-used (if any)	NA (Demo 1)	NA (Demo 1)	NA (Demo 1)
Origin of the data	Information created by PROINTEC based on the experience of its employees. Open source data (Demo 1) NA (Demo 2)	Demo 1 data origins: Data collected during UC1 on site information campaigns. Several flights of drones. Demo 2 data origins: Sensors weared by workers and machinery	Demo 1: Data collected during UC1 on site information campaigns & open data Demo 2: Original dataset from Demo 2
Volumetry: Expected size of the data (if known) and sampled frequency	1 GB (Demo1) NA (Demo 2)	20 GB (Demo1) Less than 1Gb (Demo 2)	20 GB 1/2fps (Demo1) Less than 1Gb (Demo 2)
Data utility: to whom will it be useful	These data will support the work developed prior to the start of the demonstrators and it will help to establish the requirements of the UC1	These data could be usseful to other partners in order to design and test similar modules	Ideally, these datasets should allow other partners to desing and test similiar modules.

FAIR DATA	
FINDABILITY	
Making data **findable**, including provisions for metadata	Github/Gitlab repos
Naming convention for the data	Demo 2: Column names: SrcMac DestMac DestType InRedZone EntryZone ExitZone StartTime EndTime PositionX
Search keywords	No KeyWords
Dataset version control	Only 1 dataset, no further versions expected
ACCESSIBILITY	
Making data openly **accessible**	Data is only used for internal training and validation purposes. Some of the images taken by the drones could be shared under WP7. (Demo 1)
Software tools required to access data (database querying, web services...)	Data could be accesible through GitHub/Gitlab
Are these accesibility tools open-source?	Yes
Use restrictions (if any) and access granting	No
Are the data or code licensed?	NA (Demo 1)
Authentication and authorization to repositories	Personal or organizational accouns required, registration is free.
INTEROPERABILITY	
Making data **interoperable**	Yes, the idea is to make all the data interoperable in different platforms, through the use of standard formats
Vocabularies, standards and methodologies	"CSV format is chosen for the dataset format. For data exchange and API interoperability we chose JSON format
RE-USABILITY	
Increased data **re-use** (through clarifying licenses)	It will be used for internal use.
When will the data be available?	Before the end of the project (August 2023)
How is data quality assured?	A technical expert will assure the quality.

5- CAFS

Table 6 - CAFS

5- CAFS			
DATA SETS	**CONTEXT DATA**	**RESEARCH DATA**	**AGGREGATED DATA**
Description of data types	Recorded videos for AI model (deep learning based people detector) training and testing	Labelled data from recorded videos.	N.A.
Purpose of the data collection	Train and test CAFS's AI based people detector.	N.A.	N.A.
Relation to the objectives of the project	UC5's safe passenger transfers functionality core needs this AI model	Train and Test AI Core for UC5	N.A.
Formats of data generated/collected.	Recorded video (H264/mp4)	Yolo labelled data (JPG image + TXT labelf	N.A.
Data that will be re-used (if any)	None	N.A.	N.A.
Origin of the data	Own recordings in metro platforms	Own recordings in metro platforms	N.A.
Volumetry: Expected size of the data (if known) and sampled frequency	100GB 25fps left and right cameras	10GB, ~1000 images	N.A.
Data utility: to whom will it be useful	Any application based on people detection	Any application based on people detection	N.A.

FAIR DATA	
FINDABILITY	
Making data **findable**, including provisions for metadata	Data is not public and therefore will be stored in private servers.
Naming convention for the data	
Search keywords	
Dataset version control	
ACCESSIBILITY	
Making data openly **accessible**	Data is not openly accesible.
Software tools required to access data (database querying, web services...)	
Are these accesibility tools open-source?	
Use restrictions (if any) and access granting	
Are the data or code licensed?	
Authentication and authorization to repositories	
INTEROPERABILITY	
Making data **interoperable**	All the data will be interoperable in different platforms, through the use of standard formats.
Vocabularies, standards and methodologies	
RE-USABILITY	
Increased data **re-use** (through clarifying licenses)	Internal use.
When will the data be available?	
How is data quality assured?	Labelled data statistics and environment variability inside UC5 scope is analysed.

6- SML

Table 7 - SML

6- SML			
DATA SETS	**CONTEXT DATA**	**RESEARCH DATA**	**AGGREGATED DATA**
Description of data types	N.A.	Depending on the use cases. Mainly, images and structured data.	N.A.
Purpose of the data collection	N.A.	To define and implement how the data is fed into the LEDEL library. To validate the objetives results defined by the UC.	N.A.
Relation to the objectives of the project	N.A.	To validate the results defined by the objectives of the UC. The users of the LEDEL can execute some simple experiments to be able to adapt to their specific UC.	
Formats of data generated/collected.	N.A.	Image formats (png, jpg). Structured data (CSV). Model definition (JSON).	N.A.
Data that will be re-used (if any)	N.A.	N.A.	N.A.
Origin of the data	N.A.	UC partners, public web and repositories.	N.A.
Volumetry: Expected size of the data (if known) and sampled frequency	N.A.	depending on example or UC	N.A.
Data utility: to whom will it be useful	N.A.	N.A.	N.A.

FAIR DATA	
FINDABILITY	
Making data **findable**, including provisions for metadata	Public datasets: CIFAR-10, MNIST and COCO are a few examples easily findable online.
Naming convention for the data	N.A.
Search keywords	CIFAR, MNIST, COCO
Dataset version control	N.A.
ACCESSIBILITY	
Making data openly **accessible**	N.A.
Software tools required to access data (database querying, web services...)	Internet browser
Are these accesibility tools open-source?	Yes
Use restrictions (if any) and access granting	No
Are the data or code licensed?	Yes. Tipically, Creative Commons Attribution 4.0 License.
Authentication and authorization to repositories	Not necessary
INTEROPERABILITY	
Making data **interoperable**	N. A.
Vocabularies, standards and methodologies	N.A.
RE-USABILITY	
Increased data **re-use** (through clarifying licenses)	N. A.
When will the data be available?	It has been accessible for many years
How is data quality assured?	Widely used by development community

7- ZYLK

Table 8 - ZYLK

7- ZYLK			
DATA SETS	**CONTEXT DATA**	**RESEARCH DATA**	**AGGREGATED DATA**
Description of data types	N. A.	UC1's sensor network data, consisting of (alphanumeric) data, PositionsXY, timestamp, EntryZone and Exitzone, among others	Data will be augmented by dataset completion techniques, keeping the same format and fields
Purpose of the data collection	N. A.	Create datasets of the different data origins involved in the use case. This will allow to obtain test data for experimenting and developing the fractal node's modules	Train models with better accuracy and bigger datasets
Relation to the objectives of the project	N. A.	Developing real-time mechanisms and AI techniques based on sensor and image processing.	Same as Research Data
Formats of data generated/collected.	N. A.	CSV	CSV
Data that will be re-used (if any)	N. A.	The obtained datasets will be used to train several models	The obtained datasets will be used to train several models
Origin of the data	N. A.	UC1 Dem 2 data origins: Sensors weared by workers and machinery	Original dataset from UC1 Dem 2
Volumetry: Expected size of the data (if known) and sampled frequency	N. A.	Size: Less than 1Gb Sample frequency: Each time an alert is registered for 2 months	Size: Less than 1Gb
Data utility: to whom will it be useful	N. A.	Ideally, these datasets should allow other partners to desing and test similiar modules.	Ideally, these datasets should allow other partners to desing and test similiar modules.

FAIR DATA	
FINDABILITY	
Making data findable, including provisions for metadata	Github/Gitlab repos.
Naming convention for the data	Column names: SrcMac DestMac DestType InRedZone EntryZone ExitZone StartTime EndTime PositionX
Search keywords	No KeyWords
Dataset version control	Only 1 dataset, no further versions expected
ACCESSIBILITY	
Making data openly accessible	Normally, we are used to make data accesible thorugh repositories, such as Github/Gitlab repos for accesing code and resources. Also we tend to use other kind of repositories, such as Nexus, for artifact exchange if needed.
Software tools required to access data (database querying, web services...)	GitHub and Git CLI
Are these accesibility tools open-source?	Yes
Use restrictions (if any) and access granting	No. Access is not required to be granted
Are the data or code licensed?	Yes.
Authentication and authorization to repositories	Personal or organizational accouns required, registration is free.
INTEROPERABILITY	
Making data interoperable	In order to exchange data with other partners, some APIs could be created that permit access to the desired resources and data. These APIs could be designed following a “common predefined format” decided by the consortium. Also, APIs could be fully described for ease of use, using existing documentation tools, such as doc-swagger.
Vocabularies, standards and methodologies	CSV format is chosen for the dataset format. For data exchange and API interoperability we chose JSON format And every API built for the UC uses JSON as response object.
RE-USABILITY	
Increased data re-use (through clarifying licenses)	GNU Affero 3 License is used for all software generated throughout the project
When will the data be available?	When the repositories are made public (before the end of the project - Aug 2023)
How is data quality assured?	Data have been curated and expanded to ensure data quality

8- LKS

Table 9 - LKS

8- LKS			
DATA SETS	**CONTEXT DATA**	**RESEARCH DATA**	**AGGREGATED DATA**
Description of data types	We do not need data for our role in the project	FRACTAL Features from the project	We do not need data for our role in the project
Purpose of the data collection	N. A.	Define FRACTAL Features and global project alignment with FRACTAL concepts	N. A.
Relation to the objectives of the project	N. A.	Define a global view of what FRACTAL is about and use features as a traceability aspect from requirements to component selection (WP2)	N. A.
Formats of data generated/collected.	N. A.	XML / XLSX	N. A.
Data that will be re-used (if any)	N. A.	N. A.	N. A.
Origin of the data	N. A.	Project itself	N. A.
Volumetry: Expected size of the data (if known) and sampled frequency	N. A.	<1Mb	N. A.
Data utility: to whom will it be useful	N. A.	FRACTAL Builders	N. A.

FAIR DATA	
FINDABILITY	
Making data **findable**, including provisions for metadata	Project Sharepoint WP2 Folder for Features, FRACTAL's GitHub organization management (https://github.com/project-fractal/)
Naming convention for the data	
Search keywords	
Dataset version control	
ACCESSIBILITY	
Making data openly **accessible**	FRACTAL Features have been described in project deliverables
Software tools required to access data (database querying, web services...)	Excel, Eclipse Feature IDE
Are these accesibility tools open-source?	Feature IDE is Open Source
Use restrictions (if any) and access granting	None
Are the data or code licensed?	No
Authentication and authorization to repositories	
INTEROPERABILITY	
Making data **interoperable**	
Vocabularies, standards and methodologies	
RE-USABILITY	
Increased data **re-use** (through clarifying licenses)	
When will the data be available?	
How is data quality assured?	

9- RULEX

Table 10 - RULEX

9- RULEX			
DATA SETS	**CONTEXT DATA**	**RESEARCH DATA**	**AGGREGATED DATA**
Description of data types	Data deriving from cameras and other sources (user interaction etc...)	Simulated data on UC6 scenarios	Data obtained through (1) the fusion of different data sources (cameras, user inputs, ...), (2) the aggregation at a proper time resolution, (3) the elimination/correction of anomalous records (4) the definition of proper annotation
Purpose of the data collection	Enabling smart decisions based on data on the edge, in particular for UC6 scenarios	Generating realistic scenarios to test AI algorithms	Extracting data-driven models
Relation to the objectives of the project	Contribution to AI and Safe Autonomous Decisions and to UC6	Contribution to AI and Safe Autonomous Decisions and to UC6	Contribution to AI and Safe Autonomous Decisions and to UC6
Formats of data generated/collected.	JSON files	Txt files	JSON files
Data that will be re-used (if any)	None	None	None
Origin of the data	Cameras, other onboard sensors, publicly available sources	Simulation softwares	Data preparation and AI algortihms
Volumetry: Expected size of the data (if known) and sampled frequency	Few KB of data	Few KB of data	Few KB of data
Data utility: to whom will it be useful	For stakeholders in the UC	Internally for algortihm assessment	For stakeholders in UC6 and people working on AI methods

FAIR DATA	
FINDABILITY	
Making data **findable**, including provisions for metadata	Data collected will be stored in cloud environment.
Naming convention for the data	
Search keywords	
Dataset version control	
ACCESSIBILITY	
Making data openly **accessible**	Data are shared within UC6
Software tools required to access data (database querying, web services...)	
Are these accesibility tools open-source?	
Use restrictions (if any) and access granting	
Are the data or code licensed?	
Authentication and authorization to repositories	
INTEROPERABILITY	
Making data **interoperable**	Data are accessed only within UC6
Vocabularies, standards and methodologies	
RE-USABILITY	
Increased data **re-use** (through clarifying licenses)	Data will be usable for training with different machine learning algorithms
When will the data be available?	
How is data quality assured?	

10- AITEK

Table 11 - AITEK

10- AITEK			
DATA SETS	**CONTEXT DATA**	**RESEARCH DATA**	**AGGREGATED DATA**
Description of data types	Video flows collected by IP cameras	Information detected by processing the collected video flows (e.g. people detection, crowd analysis, age and gender recognition etc.)	Presence of a person in difference reference aeras
Purpose of the data collection	Apply Video Content Analysis algorithm to detect heterogeneous information, with particular reference to UC6	Evaluation of the Video Content Analysis block developed, improvement and refinement of this block	Performance evalutation (detection quality and response time/FPS supported)
Relation to the objectives of the project	It contributes to the development of UC6	It contributes to the development of UC6	
Formats of data generated/collected.	Video flows and metadata represented as JSON files	Video flows and metadata represented as JSON files	
Data that will be re-used (if any)	No	No	
Origin of the data	Cameras	Data are obtained from the VCA block under development	
Volumetry: Expected size of the data (if known) and sampled frequency	Continuos video stream	Continuos video stream	
Data utility: to whom will it be useful	Internally (for our own contribution) and maybe with other interested partners involved in UC6	Internally (for our own contribution) and maybe with other interested partners involved in UC6	

FAIR DATA	
FINDABILITY	
Making data **findable**, including provisions for metadata	Data collected will be stored in private cloud environment.
Naming convention for the data	
Search keywords	
Dataset version control	
ACCESSIBILITY	
Making data openly **accessible**	Data could be shared with other partners establishing some sort of cooperation regarding the usage
Software tools required to access data (database querying, web services...)	
Are these accesibility tools open-source?	
Use restrictions (if any) and access granting	
Are the data or code licensed?	
Authentication and authorization to repositories	
INTEROPERABILITY	
Making data **interoperable**	Ad-hoc interface will be instatiated to assure access to some subsets of collected data
Vocabularies, standards and methodologies	
RE-USABILITY	
Increased data **re-use** (through clarifying licenses)	Data collected will be available for reuse (e.g. to train neural network for video content analysis application)
When will the data be available?	
How is data quality assured?	

11- UNIVAQ

Table 12 - UNIVAQ

11- UNIVAQ			
DATA SETS	**CONTEXT DATA**	**RESEARCH DATA**	**AGGREGATED DATA**
Description of data types	Images of people faces taken by cameras, from which age and gender are extracted (no people identity)	Information detected by processing the collected images	Age of person in the picture Gender of person in the picture
Purpose of the data collection	To estimate age and gender of the people face close to camera, with the purpose of advertising customization	Evaluating the performance of the age and gender recognition	Accuracy of age estimation Accuracy of gender identification Response time of neural network execution
Relation to the objectives of the project	The objectives are mapped in the UC6, in particular: O1: Mobile intelligent totem for customer support and personalized advertising O2: Surrounding perception processing O3: Adaptive advertise and customers support	It contributes to the development of UC6	
Formats of data generated/collected.	jpeg	JSON	
Data that will be re-used (if any)			
Origin of the data	Cameras and MORPH Dataset		
Volumetry: Expected size of the data (if known) and sampled frequency	240x200 RGB 8 bit images Sampled every time there is a face in front of the totem (UC6)		
Data utility: to whom will it be useful	Internally (for our own contribution) and optionally with other interested partners involved in UC6	Internally (for our own contribution) and maybe with other interested partners involved in UC6	

FAIR DATA	
FINDABILITY	
Making data **findable**, including provisions for metadata	The types of data will be made availalble so to allow the replicability of the results. Repositories will be employed. No metadata will be provided for the purposes excluded metadata related to possible pubblications (created by the individuated editor such as IEEE, Elsevier etc.)
Naming convention for the data	not defined
Search keywords	not defined
Dataset version control	not defined
ACCESSIBILITY	
Making data openly **accessible**	dataset used to train the neural network can be shared
Software tools required to access data (database querying, web services...)	browser
Are these accesibility tools open-source?	yes
Use restrictions (if any) and access granting	no
Are the data or code licensed?	
Authentication and authorization to repositories	authentication required to access dataset used to train the neural network.
INTEROPERABILITY	
Making data **interoperable**	Ad-hoc interface will be instatiated to assure access to some subsets of collected data
Vocabularies, standards and methodologies	
RE-USABILITY	
Increased data **re-use** (through clarifying licenses)	Data will not be collected, because they are destroyed once extracted useful information.
When will the data be available?	
How is data quality assured?	The adopted dataset is widely used by development community

12- MODIS

Table 13 - MODIS

12- MODIS			
DATA SETS	**CONTEXT DATA**	**RESEARCH DATA**	**AGGREGATED DATA**
Description of data types	Diabetic Retinopathy Detection: Scientific articles that propose data sets and AI methodologies for the automatic detection of diabetic retinopathy.	Diabetic Retinopathy Detection: We used 2 publicly available datasets, MESSIDOR-2 (Decenciere et al. 2014) and IDRiD (Porwal et al. 2018). These datasets consists of high resolution images of the eye fundus with attached the diagnosis of diabetic retinopathy.	Diabetic Retinopathy Detection: Data have been imported as Pytorch tensors, resized and cropped for later processing by deep learning methods.
Purpose of the data collection	Diabetic Retinopathy Detection: The goal is to use these articles to identify suitable datasets to train and evaluate deep learning models at the task of identifying diabetic rethinopathy from images.	Diabetic Retinopathy Detection: Train and evaluate deep learning models at the task of diagnosis diabetic retinopathy.	Diabetic Retinopathy Detection: Train and evaluate deep learning models at the task of diagnosis diabetic retinopathy.
Relation to the objectives of the project	Diabetic Retinopathy Detection: Methodologial studies.	Diabetic Retinopathy Detection: Methodologial studies.	Diabetic Retinopathy Detection: Methodologial studies.

DATA SETS	CONTEXT DATA	RESEARCH DATA	AGGREGATED DATA
Formats of data generated/collected.	Diabetic Retinopathy Detection: pdf files.	Diabetic Retinopathy Detection: high-resolution RGB eye fundus images.	Diabetic Retinopathy Detection: Pytorch tensors.
Data that will be re-used (if any)	Diabetic Retinopathy Detection: pdf files.	Diabetic Retinopathy Detection: All.	Diabetic Retinopathy Detection: All.
Origin of the data	Diabetic Retinopathy Detection: scientific journals and conference proceedings.	Diabetic Retinopathy Detection: scientific clinical studies.	Diabetic Retinopathy Detection: Processing of the raw data.
Volumetry: Expected size of the data (if known) and sampled frequency	Diabetic Retinopathy Detection: about 20 scientific articles.	Diabetic Retinopathy Detection: MESSIDOR-2 - 1744 1440x960 images, IDRiD - 516 4288x2848 images.	Diabetic Retinopathy Detection: MESSIDOR-2 - 1744 456x456 images, IDRiD - 516 456x456 images. Numpy
Data utility: to whom will it be useful	Diabetic Retinopathy Detection: other researcher in the area.	Diabetic Retinopathy Detection: Internal.	Diabetic Retinopathy Detection: Internal.

FAIR DATA	
FINDABILITY	
Making data **findable**, including provisions for metadata	a database for diabetic retinopathy screening research. Data, 3(3): 25.
Naming convention for the data	Diabetic Retinopathy Detection: MESSIDOR-2, IDRiD.
Search keywords	N.A.
Dataset version control	N.A.
ACCESSIBILITY	
Making data openly **accessible**	**Diabetic Retinopathy Detection: Both datasets are publicly available on the web..**
Software tools required to access data (database querying, web services…)	**Diabetic Retinopathy Detection: Browser web.**
Are these accesibility tools open-source?	**Diabetic Retinopathy Detection: Yes.**
Use restrictions (if any) and access granting	Diabetic Retinopathy Detection: We will keep code and the aggregated data for interl use.
Are the data or code licensed?	Diabetic Retinopathy Detection: We will keep code and the aggregated data for interl use.
Authentication and authorization to repositories	Diabetic Retinopathy Detection: We will keep code and the aggregated data for interl use.
INTEROPERABILITY	
Making data **interoperable**	N.A.
Vocabularies, standards and methodologies	N.A.
RE-USABILITY	
Increased data **re-use** (through clarifying licenses)	N.A.
When will the data be available?	N.A.
How is data quality assured?	Diabetic Retinopathy Detection: By scientific clinical studies.

[1] S. Malik et al., Data Driven Approach for Eye Disease Classification with Machine Learning, Applied Sciences vol. 9, 2019.

[2] P.S.J. Kumar et al., Glaucoma Detection and Image Processing Approaches: A Review, Journal of Current Glaucoma Practice vol. 8, 2014.

[3] Y. Tong et al., Application of machine learning in ophthalmic imaging modalities, Eye and Vision vol. 7, 2020.

[4] L. Jain et al., Retinal Eye Disease Detection Using Deep Learning, In proceedings of the Fourthteenth Conference on Information Processing, 2018.

[5] I. Qureshi, Glaucoma Detection in Retinal Images Using Image Processing Techniques: A Survey, International Journal of Advanced Networking and Applications vol. 7, 2015.

[6] M. Madhusudhan et al., Image Processing Techniques for Glaucoma Detection, In proceedings of the First International Conference on Advances in Computing and Communications, 2011.

[7] M.A.U. Patwari et al., Detection, Categorization, and Assessment of Eye Cataracts Using Digital Image Processing, In proceedings of the First International Conference on Interdisciplinary Research and Development, 2011.

[8] I. Shaheen et al., Survey Analysis of Automatic Detection and Grading of Cataract Using Different Imaging Modalities, Applications of Intelligent Technologies in Healthcare, 2019.

[9] Z. Chen et al., Strabismus Recognition Using Eye-Tracking Data and Convolutional Neural Networks, Journal of healthcare engineering, 2018.

[10] J. Lu et al., Automated Strabismus Detection for Telemedicine Applications, arXiv, 2018.

[11] T.W. Yu. Iris Imaging for Health Diagnostics, PhD thesis, 2017.

[12] R. Azarmehr et al., Real-time embedded age and gender classification in unconstrained video. In the proceedings of the Twentyeighth Conference on Computer Vision and Pattern Recognition, 2015.

[13] A.T.Y. Chen et al., Hardware/ software co-design for a gender recognition embedded system. In the proceedings of Twentynineth International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, 2016.

[14] J. Xu et al. Deep Learning Application Based on Embedded GPU, In the proceedings of the First International Conference on Electronics Instrumentation and Information Systems, 2017.

[15] M.H. Gauswami and K.R. Trivedi, Implementation of Machine Learning for Gender Detection using CNN on Raspberry Pi Platform, In the proceedings of the Second International Conference on Inventive Systems and Control, 2018.

[16] M. Hacibeyoglu and M.H. Ibrahim, Human Gender Prediction on Facial Images Taken by Mobile Phone using Convolutional Neural Networks, arXiv, 2018.

[17] J.H. Lee et al., Joint Estimation of Age and Gender from Unconstrained Face Images using Lightweight Multi-task CNN for Mobile Applications, In the proceedings of the Conference on Multimedia Information Processing and Retrieval, 2018.

[18] P. Foggia et al., A system for Gender Recognition on Mobile Robots, In the proceedings of the Second International Conference on Applications of Intelligent Systems, 2019.

[19] C.E. Kim et al., A Comparison of Embedded Deep Learning Methods for Person Detection, arXiv, 2019.

[20] A. Salihbašić and T. Orehovački T, Development of Android Application for Gender, Age and Face Recognition Using OpenCV, In the proceedings of the Fortytwoth International Convention on Information and Communication Technology, Electronics and Microelectronics, 2019.

[21] H.T.Q. Bao and C. Sun Tae, A light-weight Gender/Age Estimation model based on Multi-taking Deep Learning for an Embedded System, In the proceedings of the Korea Information Processing Society Conference, 2020.

[22] A. Greco et al., A Convolutional Neural Network for Gender Recognition Optimizing the Accuracy/Speed Tradeoff, IEEE Access vol. 8, 2020.

[23] Y. Nagnath et al., Realtime Customer Merchandise Engagement Detection and Customer Attribute Estimation with Edge Device, In the proceedings of the International Conference on Consumer Electronics, 2020.

[24] G. Xu et al., Facial Expression Recognition Based on Convolutional Neural Networks and Edge Computing, In the proceedings of the Conference on Telecommunications, Optics and Computer Science, 2020.

[25] A. Ndikumana et al. Deep Learning Based Caching for Self-Driving Cars in Multi-Access Edge Computing, IEEE Transactions on Intelligent Transportation Systems vol. 22, 2021.

[26] X. Wang et al. Convergence of Edge Computing and Deep Learning: A Comprehensive Survey, IEEE Communications Surveys & Tutorials vol. 22, 2020.

[27] - S. Malik et al., Data Driven Approach for Eye Disease Classification with Machine Learning, Applied Sciences vol. 9, 2019.

[28] - P.S.J. Kumar et al., Glaucoma Detection and Image Processing Approaches: A Review, Journal of Current Glaucoma Practice vol. 8, 2014.

13- UNIMORE

Table 14 - UNIMORE

13- UNIMORE			
DATA SETS	**CONTEXT DATA**	**RESEARCH DATA**	**AGGREGATED DATA**
Description of data types	Image of people retrieved from a camera. People density estimation and face detection performed on the input images.	Software specification; hardware platform specification	Mean, min and max density estimation
Purpose of the data collection	The purpose of the Density Estimator (DE) is to estimate the density of the people in front of the smart totem. While the aim of the Face Detector (FD) is to crop the faces of the people in front of the totem for further neural network processing by the other UC6 components.	Assessment of the RT properties	Assessment of the RT properties
Relation to the objectives of the project	The objectives are mapped in the UC6, in particular: O1: Intelligent totem for customer support and personalized advertising O2: Surrounding perception processing O3: Adaptive advertise and customers support	The objectives are mapped in the UC6, in particular: O1: Intelligent totem for customer support and personalized advertising O2: Surrounding perception processing O3: Adaptive advertise and customers support	The objectives are mapped in the UC6, in particular: O1: Intelligent totem for customer support and personalized advertising O2: Surrounding perception processing O3: Adaptive advertise and customers support
Formats of data generated/collected.	Images (.jpg, .png) for the input/output of the components. Raw text files or spreadsheets to collect the outputs.	Raw text files, or spreadsheets	Raw text files, or spreadsheets
Data that will be re-used (if any)	Not planned on UC6	Performance and accuracy information can be reuse to build similar systems using the same components.	Performance and accuracy information can be reuse to build similar systems using the same components.
Origin of the data	Input images written on a shared memory pool.	Input images written on a shared memory pool.	Input images written on a shared memory pool.
Volumetry: Expected size of the data (if known) and sampled frequency	Few KBs of data	Few KBs of data	Few KBs of data
Data utility: to whom will it be useful	Internally (for our own contribution) and optionally with other interested partners involved in UC6	researchers that are interested to build AI-based components using FPGA accelerators	System designers

FAIR DATA	
FINDABILITY	
Making data **findable**, including provisions for metadata	Data will be published in scientific papers or shared under NDA.
Naming convention for the data	
Search keywords	
Dataset version control	
ACCESSIBILITY	
Making data openly **accessible**	Data will be published in scientific papers or shared under NDA.
Software tools required to access data (database querying, web services...)	
Are these accesibility tools open-source?	
Use restrictions (if any) and access granting	
Are the data or code licensed?	
Authentication and authorization to repositories	
INTEROPERABILITY	
Making data **interoperable**	Data will be published in scientific papers or shared under NDA.
Vocabularies, standards and methodologies	
RE-USABILITY	
Increased data **re-use** (through clarifying licenses)	Data will be published in scientific papers or shared under NDA.
When will the data be available?	
How is data quality assured?	

14- UNIGE

Table 15 - UNIGE

14- UNIGE			
DATA SETS	**CONTEXT DATA**	**RESEARCH DATA**	**AGGREGATED DATA**
Description of data types	Audio signal recordings	Feature extracted from the audio recording suited to be used by the audio processing methods implemented as well validation results of the developed solutions.	Language of the speaker recognized by using the audio signal processing function implemented
Purpose of the data collection	To obtain the input signal for the audio processing functions of the UC6	Evaluationg the recognition performance of the audio processing functions of the UC6	Testing the UC6 performance from the audio recognition functions viewpoint
Relation to the objectives of the project	Developing the audio processing solutions to recognize, using voice signals, the language of the speakers in the area of the totem.	Developing and Validating the audio processing solutions to recognize, using voice signals, the language of the speakers in the area of the totem.	Implementation of UC6 demonstrator
Formats of data generated/collected.	Typical audio format, preferably uncompressed, such as WAV, AIFF, AU or raw header-less PCM.	Typical text files formats such as arff, csv, dat, json	Typical text files formats such as arff, csv, dat, json
Data that will be re-used (if any)	Collected data can be reused to train/test similar systems or UCs employing the same component.	Performance data can be reused to build similar systems or UCs employing the same component	Aggregated data can be reused to build similar systems or UCs employing the same component
Origin of the data	Open source data, self-recorded data and recordings from the totem of UC6	Obtained from the processing of the audio recordings by applying features extraction procedures implemented during the project developement.	Obtained from the testing campaigns of the UC6 demonstrator
Volumetry: Expected size of the data (if known) and sampled frequency	Few MB of data.	Few MB of data.	Few MB of data.
Data utility: to whom will it be useful	These data will support the work developed prior to the start of the demonstrators and it will help to establish the requirements of the UC6 and, obviously, for the developement of the aformentiond audio processing functions.	These data will be used for the developement of the aformentiond audio processing functions of UC6.	These data will be used for the evaluation of the aformentiond audio processing functions of UC6

FAIR DATA	
FINDABILITY	
Making data **findable**, including provisions for metadata	The aformentiond types of data will be made availalble so to allow the replicability of the results. Online repositories will be employed. No metadata will be provided for the aformentioned purposes excluded metadata related to possible pubblications (created by the individuated editor such as IEEE, Elsevier etc.)
Naming convention for the data	
Search keywords	
Dataset version control	
ACCESSIBILITY	
Making data openly **accessible**	Simple online repositories will be employed. Data will be availalble and downloadable from individuated cloud by all partners and, for what concern research data and publications, public access will be provided.
Software tools required to access data (database querying, web services...)	
Are these accesibility tools open-source?	
Use restrictions (if any) and access granting	
Are the data or code licensed?	
Authentication and authorization to repositories	
INTEROPERABILITY	
Making data **interoperable**	The selected formats, reported above, are standards. As a concequance the employed data are natively interoperable.
Vocabularies, standards and methodologies	
RE-USABILITY	
Increased data **re-use** (through clarifying licenses)	Context and research data can be easily reused as above said. Concerning aggregated data, pubblication in particular, open access journals will be preferred.
When will the data be available?	
How is data quality assured?	

15- ROT

Table 16 - ROT

DATA SETS	CONTEXT DATA	RESEARCH DATA	AGGREGATED DATA
Description of data types	Text, images, metadata	Data from totems communication block on UC6 scenarios	Output of data processing
Purpose of the data collection	Load balancing between totems	Experimenting and developing UC6 communication block	Totems and communications performance
Relation to the objectives of the project	Processing of data from totems in UC6	Analysis and experiments on totems in UC6	To better energy consumption
Formats of data generated/collected.	Any data format (JSON, CSV, others)	Any data format (JSON, CSV, others)	Any data format (JSON, CSV, others)
Data that will be re-used (if any)	To be decided	To be decided	To be decided
Origin of the data	Data acquired from totems in UC6	Data acquired from totems in UC6	Data acquired from totems in UC6
Volumetry: Expected size of the data (if known) and sampled frequency	Indicatively MBs/month	Indicatively MBs/month	Indicatively MBs/month
Data utility: to whom will it be useful	This data will be useful for the development and the processing functionalities in UC6	This data will be useful for all the partners involved in UC6	This data will be useful for all the partners involved in UC6

FAIR DATA	
FINDABILITY	
Making data **findable**, including provisions for metadata	n/a
Naming convention for the data	n/a
Search keywords	n/a
Dataset version control	n/a
ACCESSIBILITY	
Making data openly **accessible**	The data will be accessible internally to the UC6
Software tools required to access data (database querying, web services...)	n/a
Are these accesibility tools open-source?	n/a
Use restrictions (if any) and access granting	n/a
Are the data or code licensed?	n/a
Authentication and authorization to repositories	n/a
INTEROPERABILITY	
Making data **interoperable**	We will use standard data format, thus data will be fully interoperable.
Vocabularies, standards and methodologies	JSON, wav, mp3
RE-USABILITY	
Increased data **re-use** (through clarifying licenses)	The data will be re-usable internally to the UC6
When will the data be available?	n/a
How is data quality assured?	n/a

16- AVL

Table 17 - AVL

AVL			
DATA SETS	**CONTEXT DATA**	**RESEARCH DATA**	**AGGREGATED DATA**
Description of data types	a.) Measurement data from the testbed or from the field, b.) Environmental data , c) Data augumented by simulating traffic over the real environmenta data	a.) Simulation data	a.) Curated simulation data
Purpose of the data collection	a.) Training (development) and validation (robustness) of the machine learning models b.) Input for the simulation model	a.) Training (development) and validation (robustness) of the machine learning models	a.) Training (development) and validation (robustness) of the machine learning models
Relation to the objectives of the project	a.) Development of the data-driven controller b.) development of predictive diagnostics fuction	a.) Development of the data-driven controller b.) development of predictive diagnostics fuction	a.) Development of the data-driven controller b.) development of predictive diagnostics fuction
Formats of data generated/collected.	Excel (.xls(x)), CSV, Parquet	Excel (.xls(x)), CSV, Parquet	Excel (.xls(x)), CSV, Parquet
Data that will be re-used (if any)	a.) Measurement data from the testbed or from the field will be re-used from previous projects.	a.) Engine simulation model developed from a previous project.	a.) Engine simulation model developed from a previous project, cross validation
Origin of the data	a.) Measurement data from AVL previous projects b.) Environmental data from an open source data base. c) Simulation model. d) Additional drive cycles based on open source cycles.	a.) Simulation model	a.) Simulation model
Volumetry: Expected size of the data (if known) and sampled frequency	a.) Size: > 100MB; b.) Sample rate: 10Hz	a.) Size: > 10GB; b.) Sample rate: 10Hz	a.) Size: > 10GB; b.) Sample rate: 10Hz
Data utility: to whom will it be useful	This data is usefull for in-simulation training of RL algorithms in WP7 UC2. Open source data is useful for WP5TC3. Open source data is also useful for simulation of offline training of RL algorithms in WP7UC2.	This data is usefull for valuation of plant model as well as of trained RL policies in WP7 UC2.	This data is usefull for valuation of plant model as well as of trained RL policies in WP7 UC2.

FAIR DATA	
FINDABILITY	
Making data **findable**, including provisions for metadata	Data is not openly accessible.
Naming convention for the data	
Search keywords	
Dataset version control	
ACCESSIBILITY	
Making data openly **accessible**	Data is not openly accessible.
Software tools required to access data (database querying, web services...)	
Are these accesibility tools open-source?	
Use restrictions (if any) and access granting	
Are the data or code licensed?	
Authentication and authorization to repositories	
INTEROPERABILITY	
Making data **interoperable**	Yes (with some efforts, relying on stantad formats and conventions).
Vocabularies, standards and methodologies	Csv, Json, Parquet
RE-USABILITY	
Increased data **re-use** (through clarifying licenses)	Will not be licensed, only for internal use.
When will the data be available?	
How is data quality assured?	

17- SIEM

Table 18 - SIEM

17- SIEMENS			
DATA SETS	**CONTEXT DATA**	**RESEARCH DATA**	**AGGREGATED DATA**
Description of data types	JPEG Photos, MPEG Video stream	Measurement (execution cycles, frame/s)	n/a
Purpose of the data collection	Identification and classification of the objects in the photos and video stream	Performance evaluation of AI hardware accelerators	n/a
Relation to the objectives of the project	Frame rate improvement	Frame rate improvement	n/a
Formats of data generated/collected.	MPEG Video and JPEG photo Weights filter in form of array Training data set	Text file or graphical representation	n/a
Data that will be re-used (if any)	n/a	n/a	n/a
Origin of the data	Camera Github repository	Measurement in real world demostration	n/a
Volumetry: Expected size of the data (if known) and sampled frequency	Few MB	5 Photos, 50MB MPEG video format	n/a
Data utility: to whom will it be useful	Use-case 4	Use-case 4	n/a

FAIR DATA	
FINDABILITY	
Making data **findable**, including provisions for metadata	Data are not openly accessable
Naming convention for the data	n/a
Search keywords	n/a
Dataset version control	n/a
ACCESSIBILITY	
Making data openly **accessible**	Data is not openly accessable
Software tools required to access data (database querying, web services...)	n/a
Are these accesibility tools open-source?	n/a
Use restrictions (if any) and access granting	n/a
Are the data or code licensed?	n/a
Authentication and authorization to repositories	n/a
INTEROPERABILITY	
Making data **interoperable**	Standard format for data representation is used
Vocabularies, standards and methodologies	n/a
RE-USABILITY	
Increased data **re-use** (through clarifying licenses)	Data is not openly accessable
When will the data be available?	n/a
How is data quality assured?	n/a

18- VIF

Table 19 - VIF

18- VIF			
DATA SETS	**CONTEXT DATA**	**RESEARCH DATA**	**AGGREGATED DATA**
Description of data types	Sensor data from lidar and GPS, vehicle information like velocity, acceleration, steering angle	Simulation data from a 3D simulator inluding simulated sensor and vehicle data	N.A.
Purpose of the data collection	Training, validation and testing of VAL_UC7 functions	Training, and testing of VAL_UC7 functions	N.A.
Relation to the objectives of the project	Development of a collision avoidance and path tracking function	Development of a collision avoidance and path tracking function	N.A.
Formats of data generated/collected.	rosbag (.bag), CSV	rosbag (.bag), CSV	N.A.
Data that will be re-used (if any)	Sensor and vehicle data from previous projects for development of functions	Simulation models	N.A.
Origin of the data	Measurements of sensors and actors	Simulation	N.A.
Volumetry: Expected size of the data (if known) and sampled frequency	1-10 GB/min, 1-50 Hz	1-10 GB/min, 1-50 Hz	N.A.
Data utility: to whom will it be useful	UC-7	UC-7	N.A.

FAIR DATA	
FINDABILITY	
Making data **findable**, including provisions for metadata	Data is ordered by time. Each data point contains a timestamp. The location is included in the recording and can be extracted for each time stamp. Additional metadata for adding scenario of the test.
Naming convention for the data	Naming of rosbag: spider_YYYY-MM-DD-hh-mm-ss.bag
Search keywords	-
Dataset version control	No version control is used.
ACCESSIBILITY	
Making data openly **accessible**	Sensor data from lidar can be made accessible on demand from project partners. No public data in general. The datasets are typically not usefull for public and are huge in size 1-50 GB.
Software tools required to access data (database querying, web services...)	Robot Operating System (ROS).
Are these accesibility tools open-source?	Yes, mainly in BSD license.
Use restrictions (if any) and access granting	
Are the data or code licensed?	Data can be made open on demand, the source code is closed license.
Authentication and authorization to repositories	
INTEROPERABILITY	
Making data **interoperable**	Storage of sensor data in standardizes ROS message types
Vocabularies, standards and methodologies	Use of common message definitions. E.g. ROS sensor_msgs::PointCloud2 for lidar data.
RE-USABILITY	
Increased data **re-use** (through clarifying licenses)	No license, only for internal use.
When will the data be available?	
How is data quality assured?	

19- SIEG

Table 20 - SIEG

19- SIEG			
DATA SETS	**CONTEXT DATA**	**RESEARCH DATA**	**AGGREGATED DATA**
Description of data types	Input and output scheduling problems.	Raw measurements (e.g Latencies e.t.c)	N.A
Purpose of the data collection	Training and testing AI models.	Validation of Implemented Intellectual properties, including AI based experiments	N.A
Relation to the objectives of the project	Adaptability at the system, node and core level.	The validation ensures the individual services of the FRACTAL node meets its specification	N.A
Formats of data generated/collected.	Files with extension ".json"	Files with extension ".dat" and ".mem" , ".data" or "*.dat"	N.A
Data that will be re-used (if any)	N.A	0	N.A
Origin of the data	Scenario Generator block, Genetic Algorithm Scheduler block and Metascheduler.	Local - Data originated from the University of Siegen	N.A
Volumetry: Expected size of the data (if known) and sampled frequency	Gigabytes projected	Gigabytes projected	N.A
Data utility: to whom will it be useful	Internally (for our own contribution)	Project partners, and researchers whose work are based on FPGA development	N.A

FAIR DATA	
FINDABILITY	
Making data **findable**, including provisions for metadata	The data used will be stored in Fractal's GitHub repository.
Naming convention for the data	Not defined
Search keywords	Not defined
Dataset version control	Not defined
ACCESSIBILITY	
Making data openly **accessible**	Does not apply. Data only useful for internal purposes.
Software tools required to access data (database querying, web services...)	N.A
Are these accesibility tools open-source?	N.A
Use restrictions (if any) and access granting	N.A
Are the data or code licensed?	N.A
Authentication and authorization to repositories	N.A
INTEROPERABILITY	
Making data **interoperable**	The data gathered will only be used for internal validation and will not be shared externally. The focus of data sharing will be on the output of the Use Case, which will not be generated by SIEG. If appropriate, the collected data may be utilized in conference or journal papers.
Vocabularies, standards and methodologies	N.A
RE-USABILITY	
Increased data **re-use** (through clarifying licenses)	Does not apply. Data only useful for internal purposes.
When will the data be available?	N.A
How is data quality assured?	N.A

20- QUA

Table 21 - QUA

20- QUA			
DATA SETS	**CONTEXT DATA**	**RESEARCH DATA**	**AGGREGATED DATA**
Description of data types	Connection data, e.g. latencies, bandwidths, data from real indoor localisation scenarios		
Purpose of the data collection	Data to drive optimisation and to enhance the system understanding. Training data for the implementation of machine learning routines		
Relation to the objectives of the project	Necessary for the implemantation.		
Formats of data generated/collected.	CSV		
Data that will be re-used (if any)	Connection data rather not. Training data yes.		
Origin of the data	Internal test systems, customers systems with access		
Volumetry: Expected size of the data (if known) and sampled frequency	Gigabytes projected		
Data utility: to whom will it be useful	Company internal for development and research		

21- BEE

Table 22 - BEE

21- BEE			
DATA SETS	**CONTEXT DATA**	**RESEARCH DATA**	**AGGREGATED DATA**
Description of data types	No data collection planned. Available data will be used for training and testing.	N.A.	N.A.
Purpose of the data collection	Train and test object recognition, regarding human bodies and other participants in the system.	N.A.	N.A.
Relation to the objectives of the project	UC8 model training and usage	N.A.	N.A.
Formats of data generated/collected.	The origin will be open source data sets and the inference will be done in the cameras.	N.A.	N.A.
Data that will be re-used (if any)	Not planned	N.A.	N.A.
Origin of the data	Open Source Data Sets	N.A.	N.A.
Volumetry: Expected size of the data (if known) and sampled frequency	several GB	N.A.	N.A.
Data utility: to whom will it be useful	BEE	N.A.	N.A.

FAIR DATA	
FINDABILITY	
Making data **findable**, including provisions for metadata	not planned
Naming convention for the data	
Search keywords	
Dataset version control	
ACCESSIBILITY	
Making data openly **accessible**	not planned
Software tools required to access data (database querying, web services...)	
Are these accesibility tools open-source?	
Use restrictions (if any) and access granting	
Are the data or code licensed?	
Authentication and authorization to repositories	
INTEROPERABILITY	
Making data **interoperable**	not planned
Vocabularies, standards and methodologies	
RE-USABILITY	
Increased data **re-use** (through clarifying licenses)	not planned
When will the data be available?	
How is data quality assured?	

22- THA

Table 23 - THA

22- THA			
DATA SETS	**CONTEXT DATA**	**RESEARCH DATA**	**AGGREGATED DATA**
Description of data types	Thales has not collected such data.	Thales has not collected such data.	Thales has not generated such data.
Purpose of the data collection	n.a.		
Relation to the objectives of the project	n.a.		
Formats of data generated/collected.	n.a.		
Data that will be re-used (if any)	n.a.		
Origin of the data	n.a.		
Volumetry: Expected size of the data (if known) and sampled frequency	n.a.		
Data utility: to whom will it be useful	n.a.		

FAIR DATA	
FINDABILITY	
Making data **findable**, including provisions for metadata	In FRACTAL, Thales plans has made open-source contributions related to RISC-V processor cores. These contributions are located in the public GitHub repository: - https://github.com/openhwgroup/cva6 - https://github.com/openhwgroup/cva6-sdk - https://github.com/thalesgroup/meta-openhw (likely to move to https://github.com/openhwgroup/)
Naming convention for the data	n.a.
Search keywords	n.a.
Dataset version control	GitHub version management
ACCESSIBILITY	
Making data openly **accessible**	Through the public repository (see above)
Software tools required to access data (database querying, web services...)	Web browser or GitHub client.
Are these accesibility tools open-source?	Yes
Use restrictions (if any) and access granting	None
Are the data or code licensed?	Yes, under Apache/Solderpad and MIT permissive licences.
Authentication and authorization to repositories	None needed to use the IP (read access) Contributors need a GitHub account, an Eclipse Foundation account and signing Eclipse Contributor Agreement.
INTEROPERABILITY	
Making data **interoperable**	For the RISC-V processor cores, we use a common hardware description language (SystemVerilog) and standardized interface, such as AXI. CVA6 mostly uses standard RISC-V extensions and is therefore supported by GCC compiler.
Vocabularies, standards and methodologies	RISC-V instruction set
RE-USABILITY	
Increased data **re-use** (through clarifying licenses)	The contributions through the OpenHW Group provide a high level of visibility. The contributions are be under Apache/Solderpad and MIT licences, which are in favor of reuse in many contexts, either proprietary or open-source.
When will the data be available?	A complete set is already available. We regularly update the repositories with new features, improvements...
How is data quality assured?	Non-regression tests. A follow-up in the subsequent TRISTAN KDT project will increase the verification level to reach 100% cove

23- ETH

Table 24 - ETH

23- ETH			
DATA SETS	**CONTEXT DATA**	**RESEARCH DATA**	**AGGREGATED DATA**
Description of data types	none - ETH develops an open source platform, but does not collect data for research purposes	none- ETH was active in digital design. The process creates source code files (Verilog) that can be turned into functional hardware, and uses simulation vectors to verify correct functionality. Open source datasets are sometimes used to provide the simulation vectors.	none
Purpose of the data collection	none - there is context data that is part of a usual digital design flow (i.e. listing operating conditions for a timing report for example), but none that fall into what is expected in the DMP	none directly - simulation data is used to verify functionality.	
Relation to the objectives of the project	n.A.	n.A.	
Formats of data generated/collected.			
Data that will be re-used (if any)			
Origin of the data		open source repositories	
Volumetry: Expected size of the data (if known) and sampled frequency			
Data utility: to whom will it be useful		verification of the developed platforms	

FAIR DATA	
FINDABILITY	
Making data **findable**, including provisions for metadata	ETH Zurich has released source code over GitHub, which is a well known platform for searching and locating open source
Naming convention for the data	
Search keywords	
Dataset version control	
ACCESSIBILITY	
Making data openly **accessible**	ETH has released all source code developed through FRACTAL over GitHub
Software tools required to access data (database querying, web services...)	
Are these accesibility tools open-source?	
Use restrictions (if any) and access granting	
Are the data or code licensed?	
Authentication and authorization to repositories	
INTEROPERABILITY	
Making data **interoperable**	There is no research data being produced by ETH Zurich in this project
Vocabularies, standards and methodologies	
RE-USABILITY	
Increased data **re-use** (through clarifying licenses)	ETH uses Apache/Solderpad for the SW/HW released as open source. Once again, there is no research data being generated/measured/collected/released
When will the data be available?	
How is data quality assured?	

24- ACP

Table 25 - ACP

24- ACP			
DATA SETS	**CONTEXT DATA**	**RESEARCH DATA**	**AGGREGATED DATA**
Description of data types	images of meters taken by the camera of the UC-3 prototype	extracted information by fractal node	n.a.
Purpose of the data collection	training set for CNN	verification of prototype	n.a.
Relation to the objectives of the project	meter "reading"	0	n.a.
Formats of data generated/collected.	raw image	raw	n.a.
Data that will be re-used (if any)	n.a.	none	n.a.
Origin of the data	camera of UC-3 prototype	camera of UC-3 prototype	n.a.
Volumetry: Expected size of the data (if known) and sampled frequency	in the range of kB	in the range of kB	n.a.
Data utility: to whom will it be useful	use case 3	use case 3	n.a.

FAIR DATA	
FINDABILITY	
Making data **findable**, including provisions for metadata	data will be stored in ACPs gitlab repository
Naming convention for the data	
Search keywords	
Dataset version control	
ACCESSIBILITY	
Making data openly **accessible**	not planned
Software tools required to access data (database querying, web services...)	
Are these accesibility tools open-source?	
Use restrictions (if any) and access granting	
Are the data or code licensed?	
Authentication and authorization to repositories	
INTEROPERABILITY	
Making data **interoperable**	
Vocabularies, standards and methodologies	
RE-USABILITY	
Increased data **re-use** (through clarifying licenses)	no license, only for internal use
When will the data be available?	
How is data quality assured?	

25- UOULU

Table 26 - UOULU

25- UOULU			
DATA SETS	**CONTEXT DATA**	**RESEARCH DATA**	**AGGREGATED DATA**
Description of data types	N.A.	PCB data, contain images of PCB boards with label of each component on board	direct use image data from downloaded dataset, just create .names, .data, train and validation files from .yaml file
Purpose of the data collection	N.A.	Use in auto training YOLO model via kubeflow	make yolov3 know where to train and evaluate the model
Relation to the objectives of the project	N.A.	UC4 - object detection	UC4 - object detection
Formats of data generated/collected.	N.A.	JPG	JPG
Data that will be re-used (if any)	N.A.	Data can be used for training different models or different versison of YOLO	Data can be used for training different models or different versison of YOLO
Origin of the data	N.A.	Provided by Siemens on LakeFS UC4	Provided by Siemens on LakeFS UC4
Volumetry: Expected size of the data (if known) and sampled frequency	N.A.	47.9MB with 1360 images	47.9MB with 1360 images
Data utility: to whom will it be useful	N.A.	UOULU FRACTAL project research team.	UOULU FRACTAL project research team.

FAIR DATA	
FINDABILITY	
Making data **findable**, including provisions for metadata	On Fractal LakeFS, UC4 folder or open source data at FISC-PCB site (available on Kaggle)
Naming convention for the data	No columns in data, just image pixels read as float arrays
Search keywords	No keyword
Dataset version control	No version, only one version available on fractal-uc4 folder on LakeFS
ACCESSIBILITY	
Making data openly **accessible**	Data can be public on Github or LakeFS
Software tools required to access data (database querying, web services...)	Git
Are these accesibility tools open-source?	Yes
Use restrictions (if any) and access granting	No
Are the data or code licensed?	On Fractal LakeFS, UC4 folder or open source data at FISC-PCB site (available on Kaggle)
Authentication and authorization to repositories	On LakeFS, may need to require registration from external people who want to download. On Git, make it as public
INTEROPERABILITY	
Making data **interoperable**	APIs could be used to exchange and permit access to the data among partners
Vocabularies, standards and methodologies	Original data should be in any image format extensions (PNG, JPG, JPEG, etc.,). For data exchange and API return, JSON format should be used
RE-USABILITY	
Increased data **re-use** (through clarifying licenses)	Self-collected or simuilated data can be public with licenses for everyone
When will the data be available?	Data available on LakeFS, code available on Github already.
How is data quality assured?	N.A.

26- HALTIAN

Table 27 - HALTIAN

26- HALTIAN			
DATA SETS	**CONTEXT DATA**	**RESEARCH DATA**	**AGGREGATED DATA**
Description of data types	Data on edge node locations, connectivity.	Simulated data, open data, UC data.	Not applicable
Purpose of the data collection	Research and development on A) intermediate edge platform B) distributed AI methods for FRACTAL nodes.	Research and development on A) intermediate edge platform B) distributed AI methods for FRACTAL nodes.	Not applicable
Relation to the objectives of the project	Task T5.2, deliverable D5.2. Task T5.1, deliverable D5.3.	Task T5.2, deliverable D5.2. Task T5.1, deliverable D5.3.	Not applicable
Formats of data generated/collected.	json and multimodal	json and multimodal	Not applicable
Data that will be re-used (if any)	Not applicable	Not applicable	Not applicable
Origin of the data	Open data sources	Simulators; open data sources; UC leaders.	Not applicable
Volumetry: Expected size of the data (if known) and sampled frequency			Not applicable
Data utility: to whom will it be useful	HALTIAN FRACTAL project research team.	HALTIAN FRACTAL project research team.	Not applicable

FAIR DATA	
FINDABILITY	
Making data **findable**, including provisions for metadata	(meta)data are assigned a globally unique and eternally persistent identifier.
Naming convention for the data	data are described with rich metadata.
Search keywords	(meta)data are registered or indexed in a searchable resource.
Dataset version control	metadata specify the data identifier.
ACCESSIBILITY	
Making data openly **accessible**	(meta)data are retrievable by their identifier using a standardized communications protocol.
Software tools required to access data (database querying, web services...)	(meta)data are retrievable by their identifier using a standardized tools for retrieving data
Are these accesibility tools open-source?	the protocol is open, free, and universally implementable.
Use restrictions (if any) and access granting	metadata are accessible, even when the data are no longer available
Are the data or code licensed?	Yes.
Authentication and authorization to repositories	the protocol allows for an authentication and authorization procedure, where necessary.
INTEROPERABILITY	
Making data **interoperable**	(meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
Vocabularies, standards and methodologies	(meta)data include qualified references to other (meta)data and (meta)data use vocabularies that follow FAIR principles.
RE-USABILITY	
Increased data **re-use** (through clarifying licenses)	(meta)data are released with a clear and accessible data usage license
When will the data be available?	Data will beavailable when there are accepltable needs which are fully compliant with data security and privacy rules and recu
How is data quality assured?	(meta)data are associated with their provenance and (meta)data meet domain-relevant community standards

27- OFFC

Table 28 - OFFC

27-OFFC			
DATA SETS	**CONTEXT DATA**	**RESEARCH DATA**	**AGGREGATED DATA**
Description of data types	N.A.	Research data will be obtained from various public locations and personal studies	none or discarded during processing
Purpose of the data collection	N.A.	Implement the new features	none or discarded during processing
Relation to the objectives of the project	N.A.	Developing new behaviour to the fractal platforms.	none or discarded during processing
Formats of data generated/collected.	N.A.	source code from public sources	source code to public sources
Data that will be re-used (if any)	N.A.	source code from public sources	source code to public sources
Origin of the data	N.A.	public sources	source code to public sources
Volumetry: Expected size of the data (if known) and sampled frequency	N.A.	Unknown	none or discarded during processing
Data utility: to whom will it be useful	N.A.	Ideally, these datasets should allow other partners to desing and test similiar modules.	none or discarded during processing

FAIR DATA	
FINDABILITY	
Making data **findable**, including provisions for metadata	tools such git-hub is used
Naming convention for the data	
Search keywords	fractal
Dataset version control	
ACCESSIBILITY	
Making data openly **accessible**	We use to make data accesible thorugh repositories, such as Github/Gitlab repos for accesing code and resources. Also we tend to use other kind of repositories, such as Nexus, in case that artifact exchange is needed.
Software tools required to access data (database querying, web services...)	web browser -i.e chrome
Are these accesibility tools open-source?	yes
Use restrictions (if any) and access granting	no
Are the data or code licensed?	no, copy rights may apply
Authentication and authorization to repositories	
INTEROPERABILITY	
Making data **interoperable**	In order to exchange data with other partners, some APIs could be created that permit access to the desired resources and data. These APIs could be designed following a “common predefined format” decided by the consortium. Also, APIs could be fully described for ease of use, using existing documentation tools, such as doc-swagger.
Vocabularies, standards and methodologies	
RE-USABILITY	
Increased data **re-use** (through clarifying licenses)	yes, open source licensing
When will the data be available?	
How is data quality assured?	

28- PLC2

Table 29 - PLC2

28- PLC2			
DATA SETS	**CONTEXT DATA**	**RESEARCH DATA**	**AGGREGATED DATA**
Description of data types	Base contribution on widely accepted reference data sets (mainly images) so no data specifically defined, created or collected.	Various Versal based project data and accompanying datasets for operation of all aspects of the platform and cognitive features	Provide aggregated Versal reference designs for the building blocks (WP3) with sets of validation data, a subset of research data.
Purpose of the data collection	n.a.	Provide a catalog of design approaches to validate particular features of the Versal platform	Allow proper reproduction of the features in the respective reference design
Relation to the objectives of the project	n.a.	Used in verification of specification and supports generation and validation of use cases	Used in verification of specification and supports generation and validation of use cases
Formats of data generated/collected.	n.a.	Various forms of full or partial Versal design data from basic project setups to complete deployable binary content	Various forms of full or partial Versal design data from basic project setups to complete deployable binary content
Data that will be re-used (if any)	n.a.	n.a.	n.a.
Origin of the data	n.a.	Generated by PLC2 or retrieved from partners to extend / enhance in PLC2	Generated by PLC2 or retrieved from partners to extend / enhance in PLC2
Volumetry: Expected size of the data (if known) and sampled frequency	n.a.	few GB per design unit, total not limited / unkown	few GB per design unit, total undefined.
Data utility: to whom will it be useful	n.a.	Partners working on Versal to help ramp with readily deployable (partial) solutions, foster insight.	Partners working on Versal to help ramp with readily deployable (partial) solutions, foster insight.

FAIR DATA	
FINDABILITY	
Making data **findable**, including provisions for metadata	Use versioning system (git) and provide indexed interface (like github), but to be inline with WP2 guidance on methods
Naming convention for the data	
Search keywords	
Dataset version control	
ACCESSIBILITY	
Making data openly **accessible**	Data will be provided openly within the consortium.
Software tools required to access data (database querying, web services...)	
Are these accesibility tools open-source?	
Use restrictions (if any) and access granting	
Are the data or code licensed?	
Authentication and authorization to repositories	
INTEROPERABILITY	
Making data **interoperable**	As most of the data under consideration is part of Versal design suite data or used in context of these projects, the interoperability results from clearly stating the version in use and adherence to the toolchain defintions.
Vocabularies, standards and methodologies	
RE-USABILITY	
Increased data **re-use** (through clarifying licenses)	n.a.
When will the data be available?	
How is data quality assured?	