DECICE

DEVICE-EDGE-CLOUD INTELLIGENT COLLABORATION FRAMEWORK

Grant Agreement: 101092582

D1.2 Data Management Plan

This project has received funding from the European Union's Horizon Europe Research and Innovation Programme under Grant Agreement No 101092582.

Document Information

Deliverable number:	D1.2
Deliverable title:	Data Management Plan
Deliverable version:	0.1
Work Package number:	WP1
Work Package title:	Project Management
Responsible partner	UGOE
Due Date of delivery:	2023-05-31
Actual date of delivery:	2023-05-31
Dissemination level:	PU
Type:	R
Editor(s):	Felix Stein (UGOE) Mirac Aydin (GWDG)
Contributor(s):	Julian Kunkel (GWDG/UGOE)
Reviewer(s):	Stefanie Mühlhausen (GWDG) Julian Kunkel (GWDG/UGOE)
Project name:	Device-Edge-Cloud Intelligent Collaboration framEwork
Project Acronym:	DECICE
Project starting date:	2022-12-01
Project duration:	36 months
Rights:	DECICE Consortium

Document History

Version	Date	Partner	Description
0.1	(2023-05-31)	UGOE/GWDG	First draft

Acknowledgement: This project has received funding from the European Union's Horizon Europe Research and Innovation Programme under Grant Agreement No 10192582.

Executive Summary

The document describes the Data Management policy that will be followed in the DECICE project. In more detail, the report lays out the DECICE project data management to specify what type of data will be used, collected or generated aiming to define the user requirements and use case scenarios. It also specifies FAIR principles to achieve for the methods and data produced by the DECICE project. More specifically, we can distinguish among the following types of research data that will be collected at different stages of the project:

- will collect use case requirements
- multiple datasets (existing or new ones) will be collected
- no data will be collected from the end users of the proposed AI solutions

The Data Management Plan will be a living document that will be regularly updated by the project partners in order to align with the needs and requirements of the project. As such updated versions of this document will be provided updating all the necessary sections in light of the needs and findings of the project.

Contents

1 Purpose and Scope of the Deliverable

The purpose of deliverable D1.2: Data Management Plan is to describe the data management life cycle for the data to be collected, processed and/or generated by the DECICE project. As part of making research data findable, accessible, interoperable and reusable (FAIR), the project's Data Management Plan (DMP) includes information on the handling of research data during and after the end of the project; what data will be accessed, collected, processed and/or generated; which methodology and standards will be applied; whether data will be shared and made open access; how data will be curated and preserved (including beyond the end of the project).

2 Abstract / publishable summary

A data management plan (DMP) is a document that outlines how research data will be collected, organized, stored, shared, and preserved throughout the course of a research project. It serves as a roadmap for managing data effectively and ensures that research data is handled in a way that promotes data integrity, accessibility, and long-term usability. As such the DECICE DMP follows the structure of the *Horizon Data Management Plan* [Eur16]. The DMP emphasizes the purpose of efficient data management and thus ensures that research data is collected, organized, stored, and shared in a structured and efficient manner. This procedure helps to manage data effectively during the project, improving data quality, accessibility, and usability. In accordance with the institutional requirements, the project proposal and the grant agreement, the DMP ensures compliance with the self imposed and predetermined requirements set by the funding agency, the scientific research initiative *Horizon Europe Research and Innovation Programme* [RI]. The DMP aims to ensure data integrity and reproducibility as well as reusability and long-term data preservation by archiving the research data beyond the project's duration and making it publicly available. As stated the report follows the DMP guidelines defined by the Horizon Data Management Plan and is based on the the FAIR guiding principles for scientific data management [Wil+16] and therefore will be structured as followed:

First, a short introduction to the research project will be given to gather a deeper understanding for the subsequent chapters and to get insights on what data will be collected, processed and what types and formats of data will the project generate during its entire life cycle. Second, a data documentation and data summary shall emphasize the importance of detailed data documentation, including the creation of metadata. This chapter focuses on the purpose and objective of the data collection/generation and its relation to the objectives of the project. With regard to the overarching project and its three use cases mentioned in the initial proposal the data summary focuses deeply on what types and formats of data will be collected, what is the origin of the data and to whom might it be useful. Afterwards, the deliverable will be aligned to the FAIR principles with regard to:

- Making data findable
- Making data openly accessible
- Making data interoperable
- Increase data reuse

- Allocation of resources
- Data security
- Ethical aspects

3 Project objectives

This deliverable contributes directly and indirectly to the achievement of all the macro-objectives and specific goals indicated in section 1.1.1 of the project plan:

Macro-objectives	Contribution of this deliverable
(O1) Develop a solution that allows to leverage a compute continuum ranging from cloud and HPC to edge and IoT.	A project that includes different working environments needs a proper management plan that provides necessary information to the partners in order to achieve their goals.
(O2) Develop a scheduler supporting dynamic load balancing for energy-efficient compute orchestration, improved use of green energy, and automated deployment.	A well-prepared data management plan contributes to collection, processing and storing of the data that the AI scheduler model will need during its training process.
(O3) Design and implement an API that increases control over network, computing and data resources.	Data management plan helps to create rules that regulate the data collection over API and the storage of this data.
(O4) Design and implement a Dynamic Digital Twin of the system with AI-based prediction capabilities as integral part of the solution.	A well-prepared data management plan contributes to collection, processing and storing of the training data that the AI scheduler model will need and the monitoring data that the Digital Twin will need when creating the current state of the system.
(O5) Demonstrate the usability and benefits of the DECICE solution for real-life use cases.	Data management plan helps to collect test cases that are similar to real-life use cases. Therefore, data management plan improves the accuracy of DECICE solution.
(O6) Design a solution that enables service deployment with a high level of trustworthiness and compliance with relevant security frameworks.	The use of data management plan allows the integration of FAIR principles, ensuring that the data management in DECICE project will be compatible with Europe Horizon and GDPR.

4 Changes made and/or difficulties encountered, if any

No significant changes to the project plan were made. No significant challenges were encountered during implementation

5 Sustainability

The data management plan is tightly coupled to multiple WPs such WP2, WP3 and WP4. It was the responsibility of WP1 to create standards for how to collect, process, analyze and store data and enforce these standards.

6 Dissemination, Engagement and Uptake of Results

6.1 Target audience

As indicated in the Description of the project, the audience for this deliverable is:

✓	The general public (PU)
	The project partners, including the Commission services (PP)
	A group specified by the consortium, including the Commission services (RE)
	This report is confidential, only for members of the consortium, including the Commission services (CO)

7 Detailed report on the deliverable

The document provides an overview of the data management plan and shows how it is aligned with the General Principle of FAIR data and its application in this deliverable. It begins with a brief summary of the project, highlighting its objectives and scope. The document then delves into the five data types utilized in the project, namely model data, scenarios for the digital twin, monitoring data from use cases, experimental data, and log data. Each data type is discussed in terms of its relevance and contribution to the project. The document further explores key topics such as the Allocation of Resources, emphasizing the efficient utilization of resources for data management and analysis. Data Security is another crucial aspect addressed, highlighting measures taken to safeguard data integrity, confidentiality, and privacy. Lastly, the document touches upon Ethical Aspects, emphasizing the adherence to ethical guidelines and regulations throughout the project to ensure responsible data handling and usage.

7.1 General Principle of FAIR data

The FAIR principles were developed to guide the management and sharing of research data in a way that promotes maximum accessibility, reusability, and interoperability. FAIR stands for Findable, Accessible, Interoperable, and Reusable, and these principles aim to address the challenges of data sharing and promote good data management practices [Wil+16].

1. **Findable**: Data should be easy to find and discover both by humans and machines. This involves assigning unique and persistent identifiers to data, using metadata to describe the data, and ensuring that data and metadata are indexed and searchable through appropriate repositories or catalogs.

2. **Accessible**: Data should be accessible to both humans and machines. This means providing open access to data, removing technical barriers to access, and ensuring that data can be retrieved and downloaded easily and without undue restrictions.
3. **Interoperable**: Data should be structured and formatted in a way that enables integration and interoperability with other data sources and tools. This involves using common data standards, formats, and vocabularies, as well as providing clear and well-documented data schemas and metadata.
4. **Reusable**: Data should be well-described, well-documented, and accompanied by relevant documentation and contextual information, enabling others to understand and reuse the data. This includes providing clear licensing information, specifying usage rights, and ensuring that data is in a usable and interpretable form.

7.1.1 Making data findable

In general, most of the data produced or used by the project is or will be identifiable and discoverable. Datasets that will be created within DECICE project will be made available to partners and will be uploaded to public or institutional data access repositories. We will use Aire [Res20] compliant sharing and make the data open along with the project publications, thus we will make this data both easily discoverable and identifiable.

7.1.2 Making data openly accessible

Balancing scientific interest and industrial confidentiality is crucial for data measurements. We will - wherever this is not prohibited by regulations - provide the raw and processed data. If this should not be possible in specific cases (despite anonymization and pseudonymisation), but data is needed scientifically, we will provide mock data. In general, we aim to share all project data (such as via AIRE)

7.1.3 Making data interoperable

DECICE uses data coming from diverse sources such as monitoring data of infrastructures. To be able to easily integrate, analyse, and share these diverse types of data, mechanisms for data integration will be adopted wherever possible aiming to ensure data interoperability.
In order to ensure interoperability and maximum re-use of DECICE data, project partners will try to collect existing and new data in standardized formats, following well-known data representation models. Effort shall be made so that all datasets use the same standards for data representation (to the extent possible) and metadata creation.

7.1.4 Increase data re-use

We will use a protocols and standards for data collection to ensure both quality and comparability (CSV files, ONNX models). All data will be structured in appropriate datasets. We will store data and its metadata in a way to adhere to community standards. If no such standards are applicable,

we will develop our own standards that will also be described in the data management plan. Doing so enables us to archive the data and ensure usage beyond the project duration.

8 Project Summary

The concept of device-edge-cloud collaboration refers to the integration and collaboration between devices (Internet of Things devices or sensors), edge computing (processing data near the source or on the edge of the network), and cloud computing (centralized processing and storage). This collaboration aims to optimize the performance, efficiency, and intelligence of systems by leveraging the capabilities of each component. Unifying such diverse systems into centrally controlled compute clusters and providing sophisticated scheduling decisions across them are two major challenges in this field. For this reason the Device-Edge-Cloud Intelligent Collaboration Framework (DECICE) will be developed. The optimal and efficient placement of workloads across heterogeneous hardware systems will be managed by the framework [Kun+23]. The projects objectives are depicted in Figure 1.

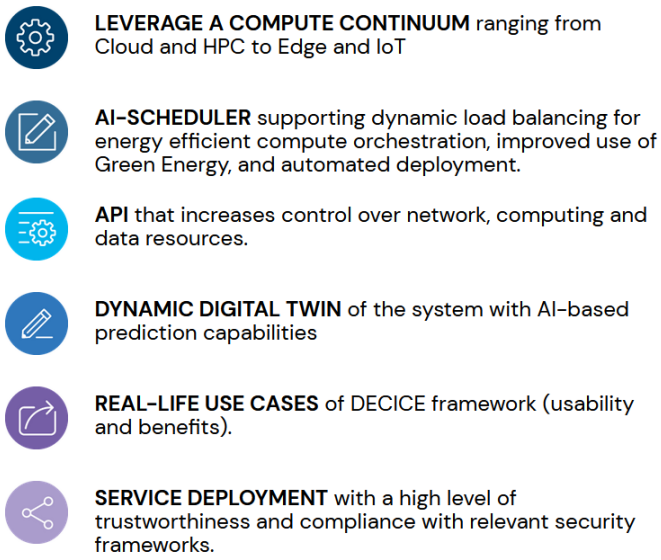

Figure 1: DECICE Objectives [DEC]

8.1 Core Components

At the core of DECICE objectives and ambition is the combination of a dynamic digital twin with domain-specific data and an AI scheduler service. A digital twin is a virtual representation or digital counterpart of a physical object. It is a digital replica that mirrors the characteristics, behavior, and attributes of its physical counterpart in real-time or near-real-time. Digital twins leverage technologies such as sensors, data analytics, and connectivity to capture, store, and analyze data from the physical object or system [AC23; Gri15]. Beyond these characteristics three use cases were defined to test, measure and evaluate the applicability of the system and to ensure that functionality and performance of the DECICE solutions do meet demands for real-life challenges:

1. Intelligent Intersection with VRU Detection (MARUN)
2. MRI Scans (GWDG)

3. In-the-Field Intelligence Supporting Emergency Response (UNIBO)

9 Data Summary

This chapter shall give a concise overview of the data generated, collected, or utilized within the research project. It provides a high-level description of the key characteristics, types, and volumes of the data, helping to provide an understanding of the project's data landscape. In general there are five areas in which data is being collected, analyzed and evaluated (digital twin, AI scheduler, use cases and their monitoring data, experimental data, log data).

9.1 AI Scheduler

9.1.1 Purpose of data collection

The data collection serves the purpose to map data and workflows to the most suitable resources. Information about data and its meta data need to be collected together with details on the workflows. In the context of an AI scheduler, the purpose of data collection is to gather relevant data that helps train and optimize the AI algorithms responsible for making intelligent scheduling decisions. The collected data serves as the foundation for building and improving the AI scheduler's predictive models, learning patterns, and making informed scheduling choices. The main goal is to enhance scheduling accuracy and to achieve optimal and efficient dynamic adaptation. Regarding the aforementioned topics, the collected data helps improve the accuracy of the AI scheduler's predictions and decision-making capabilities. By analyzing past scheduling data, the AI model can identify optimal scheduling strategies, predict resource demands, estimate task durations, and account for various factors that impact scheduling efficiency. The more comprehensive and diverse the collected data, the better the AI scheduler can learn and adapt to different scheduling scenarios. It also enables the AI scheduler to dynamically adapt and adjust scheduling decisions based on real-time or near real-time data. By continuously collecting and analyzing current scheduling data, the AI scheduler can respond to changing conditions, workload variations, resource availability, and user priorities. This adaptability helps optimize resource allocation, minimize delays, and improves overall scheduling performance.

9.1.2 Types and formats of the data

The foundation of the AI scheduler of the DECICE project can be categorized into two parts: training data and the ONNX file format. While the former falls into the category of reusing existing data to leverage the advantage of open and publicly available data for efficient training, the latter serves the purpose of providing a way to store and load the AI model. The ONNX file format is an open and interoperable format that allows for the exchange of machine learning models between different frameworks and tools [BLZ+19]. It provides a standardized way to represent deep learning models, enabling seamless integration and collaboration across various platforms. It supports a wide range of popular deep learning frameworks such as PyTorch or TensorFlow, allowing models to be trained in one framework and to be exported and used in another without the need for extensive reimplementation or conversion. The model is stored in a binary file format that contains the network architecture and model parameters.

Lastly, the trained AI scheduler models (source code, ONNX model) will be provided through a source code repository hosted on GitHub/Zenodo. Doing so enables us to archive the data and ensure usage beyond the projects duration. In addition to that a README file will also be generated and deposited into the repository and will include instructions for further usage and data analysis.

9.1.3 Reuse of existing data

In addition to generating new data, the DECICE project acknowledges the value of reusing existing data to augment and enhance the development of the framework. Reusing relevant and publicly available datasets can provide additional insights, facilitate comparative analysis, and contribute to broader research objectives. To support data reuse, we will explore reputable data repositories, scholarly databases, and domain-specific archives to identify suitable existing datasets that align with our research questions and objectives. We will thoroughly evaluate the quality, relevance, and reliability of the identified datasets before integrating them into our analyses. Proper attribution and citation practices will be followed to acknowledge the sources of reused data, ensuring compliance with applicable data usage agreements, licenses, and ethical considerations. Regarding the AI scheduler a plethora of datasets are available to which could be leveraged to our advantage and ease the need for finding/creating training data for the scheduler [FA22; TT20]. This procedure leads to two major advantages, initial training without the need to curate and create a new dataset and a fast first performance evaluation and model validation.

9.2 Digital Twin

9.2.1 Purpose of data collection

Collecting and gathering data enables the creation, simulation, and analysis of a virtual representation of a physical system. The collected data serves as the foundation for building and maintaining an accurate and functional digital twin. This endeavor provides the necessary input to develop accurate and detailed models that represent the physical asset. The collected data includes specifications, geometries, operational parameters, sensor readings, and other relevant information. These information enable the simulation and analysis of the digital twin to understand how the physical asset or system operates under different conditions. By feeding real-time or historical data into the digital twin, we can perform predictive modeling, scenario testing, and optimization. This allows for better decision-making, performance evaluation, and the identification of potential issues or improvements. Data collection also supports the monitoring and control of the real world object through its digital twin. By continuously collecting real-time data from sensors, devices, or other sources, the digital twin can provide up-to-date insights into the asset's behavior, performance, and health. This data can be used for condition monitoring, predictive maintenance, fault detection, and real-time control, improving operational efficiency and reducing downtime.

9.2.2 Types and formats of the data

The initial development of Digital twin will focus on discrete event simulations and stochastic simulations. This will be constantly improved based on real use cases and incidents. The forecast models depend on monitoring data and AI models that are based on historical data and from simulated

virtual environments. This monitoring data will be collected via Prometheus. It is an open-source monitoring system and alerting toolkit. It allows users to gather, store, visualize monitoring data as a time series data and receive alerts based on a variety of metrics. It supports multi dimensional data collection and querying using PromQL. This monitoring data will inform the digital twin and the AI models for optimisation and adaptation.

As mentioned above, the format of the data will be time series: streams of timestamped values belonging to the same metric and the same set of labeled dimensions. In addition, a glue code will be developed to parse and serialize the metric data such that it can be utilized by the digital twin.

9.2.3 Reuse of existing data

The metrics and system information collected on the nodes for the digital twin will be gathered and stored in a metrics storage such that they can be queried through the Telemetry API. Deliverable 3.3 "Final Implementation" includes the provisioning of persistent storage for the digital twin, enabling the reconstruction from disk. The GWDG will take care of the development of the glue code and the provisioning of storage solutions. BIGTRI and MARUN will contribute to the design of the data model, development of the metric storage solution. Therefore, the required data will be stored persistently, be formatted for digital twin and be available for reuse in the future.

9.3 Use Cases

9.3.1 Purpose of data collection

The overarching purpose of data collection for use cases in general and our use cases specifically is to ensure that relevant and high-quality data is systematically collected, organized, and managed to support the objectives of each individual use case. By taking our three use cases into consideration, which have been described in section 8.1, we need to assure that the data collection enables evidence based decision making and lets us execute a performance evaluation as well as validation and verification. Data and the collection of data is crucial for validating and verifying the outcomes of a use case. By continuously collecting data throughout the life cycle of the use case, we can compare the actual results against the intended objectives. This helps us in verifying whether the implemented solutions are achieving the desired outcomes and meeting the defined success criteria. Our use cases lay the foundation for an execution and evaluation of the digital twin and AI scheduler.

9.3.2 Types and formats of the data

The data required for the execution of the three defined use cases is not homogeneous. All use cases vary heavily in their nature and thus need different consideration when it comes to data collection. For the VRU detection and localization aspects Marmara University (MARUN) will collect vehicle and pedestrian data from the camera frames. The data includes the images, distance information relative to the camera, entity class, direction, and speed vectors from one or more intersections. These frames optionally will contain, motorcycles and stray animals and their relative 3D position to the camera.

The use case is planned to be first evaluated in a virtual environment specifically designed to reflect

the demo site at Marmara University Campus and will use artificially generated data. With that, the environment is generated in a 3D physics simulator and data is being collected from that environment by generating vehicle and pedestrian traffic. Collected images are going to be used for training the AI model. If needed, additional frames and 3D positioning information will be collected from the campus site. After the training and tests the model will be deployed to an edge device to realize intelligent intersection applications in real-time, incremental and federated learning will improve the model using the data collected during operation from the real or virtual environment. Anonymization techniques will be applied to be GDPR compliant.

University of Bologna (UNIBO) is responsible for the in-the-field intelligence supporting emergency response use case. The data is generated by photo realistic simulations of environments which are normally generated by drones in the real world. The data is being processed by computer vision algorithms, machine learning as well as deep learning models.

MRI scans as the third use case will be planned and executed by GWDG. In this case edge devices are deployed near MRI machines to interact with the generated data for further processing and analysis. As with the other use cases synthetic images will be created and used to test the functionality of the software architecture.

9.3.3 Reuse of existing data

Using existing data for our use cases can leverage the information and insights contained within the data to gather new approaches for decision-making, gain knowledge, and derive valuable outcomes. Existing data holds a wealth of information that can be analyzed and explored to uncover patterns, trends, correlations, and relationships. Furthermore machine learning models require a plethora and preferably diverse data to achieve a good performance and generalization when it comes to making predictions. Our use cases can leverage existing datasets as a foundation for the training process of our AI models. Regarding the MRI use case multiple datasets are available for public or scientific use [Zbo+18; Roy+22]. Among other datasets UNIBO will rely on data from Hilti SLAM [Hel+21].

9.4 Experimental Data

9.4.1 Purpose of data collection

Experimental data serves as the empirical foundation for our scientific research, allowing us to analyze and interpret the collected information to draw conclusions, establish patterns or relationships, and support or refute hypotheses. It provides evidence to support or modify our theories and contribute to the broader knowledge base. It data plays a crucial role in ensuring the reproducibility and transparency of research. By making experimental data available, other researchers can replicate or build upon previous work, validate findings, and contribute to the advancement of knowledge. The foundation for our gathering of experimental data will be our three predefined use cases.

9.4.2 Types and formats of the data

Experimental data can exist in various formats depending on the nature of the data. It can be in the form of structured, unstructured data or a combination of both. The overarching tool to collect

the data is Prometheus [RV15]. It supports multi dimensional data collection and querying using PromQL. DECICE will use this event monitoring tool for enabling monitoring telemetry data for new devices or new metrics that are not supported by Kubernetes. This monitoring data will inform the digital twin and the AI models for optimisation and adaptation and will be a crucial part for collecting experimental data during the execution of our use cases. The collected data might be several hundreds of gigabyte of data, providing these to the general public is almost impossible. But we are making CSV files of the performance data and results data publicly available.

9.4.3 Reuse of existing data

Existing Prometheus metrics can be exported as a JSON file. This exported file contains all the details of the query. Multiple queries can also be selected and exported in a .zip file. This exported Prometheus Metrics can also be imported for reuse. When we export Prometheus metrics, the exported JSON file contains all the details of the Prometheus metrics. Importing the JSON file and deploying it to the Prometheus Exporter will be enough for reuse. Multiple queries (JSON files) as a .zip file can also be imported.

9.5 Log Data

9.5.1 Purpose of data collection

Logging plays a crucial role in the DECICE project as it provides a mechanism for recording and tracking important events and activities related to data handling and processing. It enables the ability to trace and monitor data-related actions and allows for the monitoring of system performance during the execution of use cases. By capturing relevant metrics and events, such as response times, resource utilization, and throughput, logging helps identify potential performance bottlenecks or inefficiencies. It enables the analysis of system behavior and performance trends, facilitating optimizations and ensuring smooth use case execution. Logs also serve as a valuable tool for identifying and diagnosing issues or errors encountered during the execution of a test or real run. By capturing relevant contextual information, error messages, and stack traces, logging aids in the troubleshooting process. It helps pinpoint the root causes of issues, allowing developers or support teams to quickly identify and address any problems that may arise. Lastly generated logging files can also be used as input data for the AI scheduler to initiate an adaptation process.

9.5.2 Types and formats of the data

A log stores information about the state of the application (program) or the system. Logs can exist in the form of records in a regular text file, records in a database, records on a remote web service, and even emails to a specific address about certain states of application. Examples for types of log data can be found below:

- Structured Logs
- Performance & Results Data
- Event Logs/System Events
- Error Logs

- Network Traffic

In order to collect logs from the system, Prometheus will be used. Prometheus stores this data in its internal time-series database. In order to obtain this data, Prometheus provides a functional query language called PromQL (Prometheus Query Language) that lets the user select and aggregate time series data in real time. The result of an expression can either be shown as a graph, viewed as tabular data in Prometheus's expression browser, or consumed by external systems via the HTTP API.

9.5.3 Reuse of existing data

Existing Prometheus metrics can be exported as a JSON file. This exported file contains all the details of the query. Multiple queries can also be selected and exported in a .zip file.

The Prometheus Metrics which has already been exported can also be imported for reuse. When we export Prometheus metrics, the exported JSON file contains all the details of the Prometheus metrics. We have to import the JSON file and deploy (with or without modifications) to our Prometheus Exporter. We can also import multiple queries (JSON files) as a .zip file.

9.6 Data Utility

With this project, we hope to benefit the European community. This involves both using open source in deliverables, but, more importantly, contributing to open source communities. Moreover, this also encompasses contributing to standards (open or otherwise). All contributions to open source will be licensed under a framework suitable for commercialization, as recommended in the EU roadmap. Furthermore, contributors to the DECICE project will provide the research data under the FAIR principles to optimize the reuse of the collected and generated data.

10 Allocation of Resources

Costs for preservation and availability of the project data are covered by the institutions providing the repository infrastructure. To make the data long-term available even beyond the projects duration, we will publish our results on Zenodo. The data collected will never include information raising ethical issues (e.g. health related), but may include personal data related to expertise and skills. For this reason, the minimization of data will be guaranteed, and the sharing of data within the consortium will take place with non-disclosure of individual data (thanks to anonymization or pseudonymization). Anyways, data collection and management will be carried out, by each partner data protection officer, according to the related laws in place in the countries where data are collected. Additional cases will be regulated, during the project lifetime, according to the Grant agreement and the Consortium agreement.

11 Data Security

All software tools and data storage mechanisms developed within DECICE will be designed to safeguard collected data against unauthorized use and to comply with all national and EU regulations.

While the project solely utilizes experimental data and does not involve the use of real or personal data, it is important to maintain robust data security practices. The plan includes measures such as access controls, encryption, and secure storage to protect the experimental data from unauthorized access, loss, or alteration. DECICE datasets will be openly shared by uploading them in public or institutional data access repositories such as AIRE.

12 Ethical Aspects

It is not expected that any data collected and used during the project will be sensitive or personal data under the GDPR regulations. Although we are not handling any sensitive or personal data since the DECICE project is only a framework for others to use, we apply encryption, access controls, and other security practices to safeguard data integrity, confidentiality, and availability for others who might use the framework in the future. When a data cannot be shared, the reasons for this will be outlined (e.g. ethical restrictions, rules governing privacy and personal data protection, intellectual property, and commercial sensitivity). For example, imaging data - even though it is artificially generated - cannot be made publicly available due to the size of the data.

13 References

[AC23] Mohsen Attaran and Bilge Gokhan Celik. "Digital Twin: Benefits, use cases, challenges, and opportunities". In: *Decision Analytics Journal* 6 (2023), p. 100165. ISSN: 2772-6622. DOI: https://doi.org/10.1016/j.dajour.2023.100165. URL: https://www.sciencedirect.com/science/article/pii/S277266222300005X.

[BLZ+19] Junjie Bai, Fang Lu, Ke Zhang, et al. *ONNX: Open Neural Network Exchange*. 2019. URL: https://onnx.ai/ (visited on 05/24/2023).

[DEC] DECICE. *DEVICE - EDGE - CLOUD Intelligent Collaboration framEwork*. URL: https://www.decice.eu/ (visited on 05/24/2023).

[Eur16] Directorate-General for Research & Innovation European Commission. *H2020 Programme: Guidelines on FAIR Data Management in Horizon 2020 Version 3.0*. en. 2016. DOI: 10.25607/OBP-774. URL: https://www.oceanbestpractices.net/handle/11329/1259.

[FA22] Federica Filippini and Danilo Ardagna. *AI-SPRINT GPU STochastic Scheduler*. Zenodo, Dec. 2022. DOI: 10.5281/zenodo.7438625. URL: https://doi.org/10.5281/zenodo.7438625.

[Gri15] Michael Grieves. "Digital Twin: Manufacturing Excellence through Virtual Factory Replication". In: (Mar. 2015).

[Hel+21] Michael Helmberger et al. "The Hilti SLAM Challenge Dataset". In: (2021). DOI: 10.48550/ARXIV.2109.11316. URL: https://arxiv.org/abs/2109.11316.

[Kun+23] Julian Kunkel et al. *DECICE: Device-Edge-Cloud Intelligent Collaboration Framework*. 2023. DOI: 10.48550/ARXIV.2305.02697. URL: https://arxiv.org/abs/2305.02697.

[Res20] Open Access Infrastructure for Research in Europe - OpenAire. *OpenAIRE Guidelines*. 2020. URL: https://guidelines.openaire.eu/en/latest/ (visited on 05/24/2023).

[RI] Horizon Europe Research and Innovation. *Horizon Europe*. URL: https://research-and-innovation.ec.europa.eu/funding/funding-opportunities/funding-programmes-and-open-calls/horizon-europe_en (visited on 05/24/2023).

[Roy+22] Jessica Royer et al. "An Open MRI Dataset For Multiscale Neuroscience". In: *Scientific Data* 9.1 (Sept. 2022). DOI: 10.1038/s41597-022-01682-y. URL: https://doi.org/10.1038/s41597-022-01682-y.

[RV15] Bjorn Rabenstein and Julius Volz. "Prometheus: A Next-Generation Monitoring System (Talk)". In: Dublin: USENIX Association, May 2015.

[TT20] Shinichiro Takizawa and Ryousei Takano. *Effect of an Incentive Implementation for Specifying Accurate Walltime in Job Scheduling*. Proceedings of the International Conference on High Performance Computing in Asia-Pacific Region (HPCAsia2020), 2020.

[Wil+16] Mark D. Wilkinson et al. "The FAIR Guiding Principles for scientific data management and stewardship". In: *Scientific Data* 3.1 (Mar. 2016). DOI: 10.1038/sdata.2016.18. URL: https://doi.org/10.1038/sdata.2016.18.

[Zbo+18] Jure Zbontar et al. *fastMRI: An Open Dataset and Benchmarks for Accelerated MRI*. 2018. DOI: 10.48550/ARXIV.1811.08839. URL: https://arxiv.org/abs/1811.08839.

A Abbreviations and Acronyms

- AI - Artificial Intelligence
- AIRE - Access Infrastructure for Research in Europe
- API - Application Programming Interface
- CSV - Comma-Separated Values
- DECICE - Device-Edge-Cloud Intelligent Collaboration framEwork
- DMP - Data Management Plan
- FAIR - Findability, Accessibility, Interoperability, and Reuse
- GDPR - General Data Protection Regulation
- MRI - Magnetic Resonance Imaging
- ONNX - Open Neural Network Exchange
- VRU - Vulnerable Road User