AI for real-world network operation

WP6 – Project management and coordination

D6.2 – Data management plan V1

DOCUMENT INFORMATION

DOCUMENT	D6.2 – Data management plan V1
TYPE	Report
DISTRIBUTION LEVEL	PU - Public
DUE DELIVERY DATE	31/03/2024
DATE OF DELIVERY	20/03/2024
VERSION	V1.0
1DELIVERABLE RESPONSIBLE	INESC TEC
AUTHOR (S)	João Aguiar Castro
OFFICIAL REVIEWER/s	Daniel Boos (SBB), Marcello Restelli (POLIMI)

DOCUMENT HISTORY

VERSION	AUTHORS	DATE	CONTENT AND CHANGES
0.1	Ricardo Bessa	28/10/2023	Template creation
0.2	João Aguiar Castro	30/01/2024	Draft
0.3	Ricardo Bessa	30/01/2024	Revision. Two new Annex.
0.4	Samira Hamouche, Herke van Hoofe, Bruno Lemetayer, Ricardo Bessa, João Aguiar Castro	15/02/2024	Review by project partners. Overall update of the document.
0.5	João Aguiar Castro, Ricardo Bessa, Vasco Dias	15/03/2024	Comments from EDPC meeting and post-meeting feedback. Two new Annex.
0.6	Vasco Dias	18/03/2024	Additional changes in the ethics part
1.0	Daniel Boos, Marcello Restelli	20/03/2024	Formal review and changes

ACKNOWLEDGEMENTS

DISCLAIMER

SUMMARY

Sound management of research data is essential to ensure that AI4REALNET datasets are findable, accessible, interoperable, and reusable (FAIR) throughout the project life cycle and beyond. The Data Management Plan aims to describe the necessary practices and strategies to promote the FAIRness of AI4REALNET datasets. These deliverables outline recommended data management practices and information about the expected datasets (DS) the project will generate. Information on the datasets in the following list can be found in the different sections of the DMP and will be incrementally updated throughout their lifecycle.

DATASETS	TYPE OF DATA	PARTNERS INVOLVED
DS. 1.1 - Feedback from participants in the stakeholders' workshops	Textual data (comments)	RTE, TenneT, SBB, DB, NAV, TU Delft, INESC TEC
DS. 1.2 - Datasets for the use cases and digital environments	Typica data, real historical (confidential) data, public test networks, open data	All
DS. 1.3 - Interactive IA with order-agnostic architectures	Open and synthetic data	All
DS. 2.1 - Experimental and operational data collection for personalized biomarkers	Personal data	INESC TEC
DS. 2.2 - Socio-technical analysis with dispatcher and traffic controller - Interviews	Personal data	LiU, TU Delft, NAV
DS. 2.3 - Socio-technical analysis with dispatcher and traffic controller – Observation	Personal data	LiU, TU Delft, NAV
DS. 2.4 - Socio-technical analysis with dispatcher and traffic controller – Questionnaire	Personal data	LiU, TU Delft, NAV
DS. 2.5 - Knowledge-assisted AI performance	Open and synthetic data	UvA
DS. 2.6 - Hierarchical RL in relational domains	Open and synthetic data	POLIMI
DS. 3.1 - State-action space for AI agent interaction with the digital environment	Open and synthetic data	All

DATASETS	TYPE OF DATA	PARTNERS INVOLVED
DS. 4.1 – Test datasets for evaluation of AI technical performance	Open and synthetic data	All
DS. 4.2 - Test results for evaluation of AI technical performance	Open and synthetic data	All
DS. 4.3 - Test datasets for AI functional testing	Open and synthetic data	All
DS. 4.4 - Data of network operator interactions and feedback produced during Human-AI evaluation scenarios	Personal data	RTE, TenneT, SBB, DB, NAV, TU Delft, INESC TEC, LiU, TU Delft, FHNW
DS.4.5 - Feedback collected regarding the adoption of AI	Personal data	RTE, TenneT, SBB, DB, NAV, TU Delft, FHNW
DS. 4.6 - Validation data of air traffic control	Open and synthetic data	TU Delft, ZHAW
DS. 5.1 - Stakeholders map and contacts	Personal data	INESC TEC
DS. 5.2 - Dataset provided for AI competitions	Open and synthetic data	All

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS AND ACRONYMS

AI	Artificial Intelligence
ALTAI	Assessment List for Trustworthy Artificial Intelligence
CC	Creative Commons
DMP	Data Management Plan
DPO	Data Protection Officer
DOI	Digital Object Identifier
EDPC	Ethics and Data Protection Committee
FAIR	Findable, Accessible, Interoperable, Reusable
ISO	International Organization for Standardization
VPN	Virtual Private Network
WP	Work Package

TABLE 1 – ABBREVIATIONS AND ACRONYMS

INTRODUCTION

Under Horizon Europe, it is expected that beneficiaries make their data findable, accessible, interoperable, and reusable (FAIR) [1]. This Data Management Plan (DMP) aims to promote the adoption of good data management practices by the AI4REALNET project, with a principle of an approach as open as possible but as close as necessary to data sharing. This DMP provides an overview of such practices, considering FAIR requirements and recommendations. The development approach involved mapping these data management practices and aligning them with the structural components of the DMP, as recommended by the European Commission. This general approach promotes the necessary flexibility to address the different data types that the project will generate while also enabling the definition of more specific actions as the project evolves.

A template was developed and shared with project partners to obtain information about the datasets (see Annex 1). The gathered information is useful for completing the different components of the DMP and will be updated incrementally. In addition, a machine-actionable, lighter version of this DMP was created on the ARGOS platform to be made public once one of the project's datasets is ready to be openly shared.

The DMP is designed to be a living and working document, preferably updated every six months by the INESC TEC´s data steward, João Aguiar Castro, hereafter referred to as the DMP manager. The DMP manager will interact directly with the project coordinator regarding the DMP updates. Moreover, the DMP manager is a point of contact for the project partners for issues related to the DMP and is responsible for providing documents to be filled out by partners to support said updates.

Each dataset has a person responsible for its management and the preferred point of contact for updates to the datasets to be reported in the DMP throughout the project.

To raise awareness of the DMP benefits, the DMP manager was responsible for a presentation to introduce the project partners to the DMP components in the project kick-off meeting (12 October 2023) [2]. On 1 March 2024, the first meeting of the Ethics and Data Protection Committee (EDPC) took place, the Data Protection Officer (DPO) and Ethics Advisors of each partner analyzed a draft version of the DMP and provided their recommendations for improvement in the document, and also some general suggestions to the project [3]. The meeting minutes can be found in Annex 2.

The DMP is structured as follows:

- The Data Summary (Section 0) overviews the AI4REALNET datasets.
- The FAIR Data (Section 1) describes data sharing principles, the assessment of data repository services for the availability of the project datasets, recommendations for data documentation, citation information, Data Availability Statement, and file naming, among other practices.
- The Data Security (Section 2) describes high-level data storage and backup practices that can be adapted by project partners locally.
- The Legal and Ethics Compliance (Section 3) corresponds to general principles for handling data. Moreover, it explains how an ethics-by-design approach, via the Assessment List for Trustworthy AI (ALTAI) framework, is being followed to account for the possible implications of AI-based solutions to be developed or validated in the scope of this project.

DATA SUMMARY

AI4REALNET will create datasets with a variety of typologies and requirements. This section will be updated continuously to further detail datasets information throughout their lifecycle based on the DMP information template. Likewise, additional datasets will be added in the upcoming DMP updates.

The following tables provide an overview of the expected datasets by WP.

- The Data Summary for *WP1 - Design of a holistic framework for AI in critical network infrastructures* is detailed in Table 2.
- Table 3 outlines datasets from *WP2 - Fundamental AI building blocks*.
- *WP3 - AI augmented human decision-making*, *WP4 - Validation and impact assessment* and *WP5 - Dissemination, communication, and exploitation of results*, are summarized in Table 4-Table 6 respectively.

The description of the datasets corresponds to a preliminary assessment to establish the purpose of the data collection, how the datasets will be collected, and how they relate to the project´s objectives.

For completeness, a short description of the digital environments that will be used in the project, namely Grid2Op, Flatland, and BlueSky, is presented in Annex 3.

WP1 Datasets	Aim for Data Collection
DS. 1.1 - Feedback from participants in the stakeholders' workshops. Workshops and consultations will be held with external stakeholders to collect feedback on the AI4REALNET use cases.	**Aim:** To validate the proposed AI framework with realistic and relevant use cases. **How is collected:** Questionnaires, forms, free text, online meeting recording (collected by industrial partners that are leading the use cases definition). **Type of data:** Text, photos, video, and sound recording. **Software:** Online meeting recording tool and word processor software. **Expected size and format:** GB scale sound, text, and video.
DS. 1.2 - Datasets for the use cases and digital environments. Each use case needs to be reflected with a corresponding dataset. Synthetic datasets are pre-build when installing a digital environment (Grid2op, Flatland, BlueSky), but additional synthetic data will be needed and generated to match the use cases and corresponding scenarios. This data will be used to evaluate how the developed solutions deal with real-world conditions (e.g., missing, wrong, or delayed measurements) and understand how the step from simulated environments to real-world environments impacts the performance of the solutions.	**Aim:** This dataset is related to the development of novel variants of supervised and reinforcement learning, supported by a set of functions and human-machine interface for explainable, algorithmically transparent, and interpretable reinforcement learning. **How is collected:** Synthetic data will be generated by the digital environments' functions: 1) ChroniX2grid tool in Grid2Op, which may use real-world operational data of the TENNET grid such as historical load and generation time series; 2) Flatland provides functions to generate synthetic railway networks and demand for trains with structural similarity to real networks, and samples from the German or from the Swiss railway network and schedules in anonymized form will be used to create problems that resemble parts of real railway networks; 3) BlueSky contains open data aircraft performance models, and an open global navigation database including 14480 airports, performance and operating procedure coefficients for 295 different aircraft types. **Type of data:** Numerical **Software:** Digital environments Grid2op, Flatland, and BlueSky. **Expected size and format:** GB scale .csv format

WP1 Datasets	Aim for Data Collection
DS. 1.3 - Interactive AI with order-agnostic architectures. This dataset contains the interaction of one or more AI agents with virtual environments such as Grid2Op or Flatland, including the states of the virtual environment, the actions selected by the AI agent, and any rewards obtained. One AI agent could make an arbitrary subset of decisions, after which the AI agent of interest is tasked with 'filling in' the remaining decisions.	**Aim:** to provide a dataset for interactive (co-) learning. The dataset could be re-used within and outside of the project to evaluate the capacity of learning algorithms to make good decisions under various imposed decisions by an interaction partner. **How is it collected: via** API from the digital environment and AI agent and stored in a file or database. **Type of data:** Numerical **Software:** Log files will be compatible with freely available tools such as TensorBoard **Expected size and format:** Logfile in, e.g., TensorBoard format, MB scale. **Note:** This data will also be generated in WPs 2-3 using the environments from WP1.

TABLE 2 – WP1 DESIGN OF A HOLISTIC FRAMEWORK FOR AI IN CRITICAL NETWORK INFRASTRUCTURES, DATA SUMMARY

WP2 Datasets	Aim for Data Collection
DS. 2.1 - Experimental and operational data collection for operator personalized biomarkers. Make use of Psychophysiological data for human integration in digital environments. Experimental tests will be conducted using a validated experimental stress testing platform and a simple reaction time task procedure for cognitive performance analysis and personalization of biomarkers. Surveys and biosensors will be used to collect data from the human operators.	**Aim:** To understand how the user's psychophysiology changes can contribute to the digital environments' adaptation to the user during the various use cases. The dataset also aims to characterize each subject in an initial experimental test. **How is collected:** Making use of proprietary wearable devices, paper questionnaires, and smartphones that store all the data locally and share it with the INESC TEC database server using a REST API **Type of data:** Numerical (Discrete and continuous), categorical (nominal) **Software:** Smartphone for real-time computation of the data, offline tools based on Python, MATLAB, and SPSS
DS. 2.2 - Socio-technical analysis with dispatcher and traffic controller – Interviews. This dataset is generated via qualitative data collection, such as interviews. This dataset aims to model the "work-as-done" process from which requirements for the technical and social system can be derived.	**Aim:** to model the "work-as-done" process from which requirements for the technical and social system can be derived. **How is collected:** Interviews **Type of data:** Audio record transcripts **Software**: MAXQDA and audio recording Expected size and format: mx22, mp3
DS. 2.3 - Socio-technical analysis with dispatcher and traffic controller – Observation. This dataset is generated via qualitative data collection, such as observation, with quantitative elements, such as questionnaires. This data set aims to model the "work-as-done" process from which requirements for the technical and social system can be derived.	**Aim:** to collect information about the status quo concerning psychological criteria such as motivation to identify technical and social system requirements. **How is collected:** Observation during "normal" work activities, taking notes. **Type of data:** Text, description of observed subjects **Software:** MAXQDA, Word **Expected size and format:** mx22, docx
DS 2.4 - Socio-technical analysis with dispatcher and traffic controller – Questionnaire. This dataset is generated via quantitative data collection, such as questionnaires. This data set aims to collect information about the status quo concerning psychological criteria, such as	**Aim:** To collect information about the status quo concerning psychological criteria such as motivation to identify technical and social system requirements. **How is collected:** Questionnaire via Tivian (unipark). **Type of data:** text **Software:** Tivian, SPSS, Excel **Expected size and format:** to be defined

WP2 Datasets	Aim for Data Collection
motivation to identify technical and social system requirements.	
DS. 2.5 - Knowledge-assisted AI performance. This dataset contains 1) the interaction of an AI agent with digital environments, including the states of the environment, the actions selected by the AI agent, and any rewards obtained, and 2) a description of any prior knowledge bases used by the AI agent.	**Aim:** to provide essential reinforcement learning tools that extend current capabilities in the project domains, which the other WP can use. **How is it collected: via** API from the digital environment and AI agent and stored in a file or database. **Type of data:** Numerical, relational **Software:** Log files with freely available tools such as TensorBoard **Expected size and format:** Logfile in, e.g., TensorBoard format, MB scale.
DS. 2.6 - Hierarchical RL in relational domains. This dataset contains the interaction of an AI agent with virtual environments such as Grid2Op or Flatland, including the states of the virtual environment, the actions selected by the AI agent, and any rewards obtained. In particular, the study of a hierarchical AI agent in a relational domain.	**Aim:** to provide basic RL tools that extend current capabilities in the project domains, which the other work packages can use. **How is it collected:** via API from the digital environment and AI agent and stored in a file or database. **Type of data:** Numerical, relational **Software:** Log files will be compatible with freely available tools such as TensorBoard **Expected size and format:** Logfile in, e.g., TensorBoard format, MB scale.

TABLE 3 – WP2 FUNDAMENTAL AI BUILDING BLOCKS, DATA SUMMARY

WP3 Datasets	Aim for Data Collection
DS. 3.1 - State-action space for AI agent interaction with the digital environment. This dataset is generated via interaction between an AI agent (e.g., reinforcement learning agent) and the digital environment in a series of episodes. The digital environment generates the state data produced by considering synthetic or real data regarding network operating conditions combined with a physical network (which can be based on a real network). The AI agent produces the action data and implements it in the digital environment. Moreover, a reward value is computed with a given reward function for each interaction.	**Aim:** This dataset is related to the development of novel variants of supervised and reinforcement learning for large-scale complex networks, exploiting domain knowledge and hierarchical and distributed problem decomposition to increase scalability in realistic networks. **How is it collected:** via API from the digital environment and AI agent and stored in a file or database. **Type of data:** Numerical **Software:** Grid2Op, Flatland, BlueSky. Log files will be compatible with freely available tools such as TensorBoard. **Expected size and format:** MB scale. CSV or any file format, Logfile in, e.g., TensorBoard format, MB scale.

TABLE 4 – WP3 AI AUGMENTED HUMAN DECISION-MAKING, DATA SUMMARY

WP4 Datasets	Aim for Data Collection
DS. 4.1 - Test datasets for evaluation of AI technical performance. Datasets shall reflect realistic and representative operational scenarios in accordance with the predefined use cases. Datasets shall also be scaled so that AI solutions' scalability can be evaluated.	**Aim:** To validate the proposed AI framework with realistic and relevant use cases. **How is collected:** Synthetic data will be generated by the digital environments, using real data from industrial partners as the basic knowledge. **Type of data:** Numerical **Software:** Digital environment. **Expected size and format:** GB scale .csv format.
DS 4.2 - Test results from the evaluation of AI technical performance. This data will contain all results from tests carried out during Task 4.1, including the calculated KPIs.	**Aim:** To validate the proposed AI framework with realistic and relevant use cases. **How is it collected: Via API or log data from the digital environment during and after the test is run** and stored in a file or database. **Type of data:** Numerical **Software:** Any numerical data processor software. **Expected size and format**: GB scale .csv format
DS 4.3 - Data of network operator interactions and feedback produced during Human-AI evaluation scenarios. Experimental protocols will be conducted to evaluate the balance between AI and humans in the selected use case. This means that biosensors will be used to collect psychophysiological measures that provide information on human operator internal processes, user experience, and acceptability along several dimensions (e.g., attention, mental workload, stress).	**Aim:** To develop human-assistance and co-learning strategies between humans and AI that augment decision-making capabilities under risk and uncertainty. **How is collected:** Using a REST API, proprietary wearable devices, paper questionnaires, and smartphones that store all the data locally and share it with the INESC TEC database server. **Type of data:** Text, psychophysiological measures, video (e.g., eye position tracking). **Software**: Biosensors (for psychophysiological measures), cameras, any word processor software **Expected size and format:** GB scale text and video.
DS. 4.4 – Test datasets for AI functional testing. This data shall reflect realistic and representative operational scenarios in accordance with the predefined use cases and for functional testing of AI agents with adversarial datasets.	**Aim:** To validate the proposed AI framework with realistic and relevant use cases. **How is collected:** synthetic data will be generated **Type of data:** Numerical **Software:** Any numerical data processor software. **Expected size and format**: GB scale .csv format.
DS. 4.5 - Feedback collected regarding the adoption of AI. Interviews will be conducted with industry experts involved in the AI4REALNET project to collect reactions (e.g., qualms, worries, and positive aspects) towards adopting AI under employment law and workers' proper perspective.	**Aim:** To validate the proposed AI framework with realistic and relevant use cases. **How is collected:** Surveys, interviews, and questionnaires. **Type of data:** Text, Audio record and transcripts **Software:** Biosensors (for psychophysiological measures), cameras, any word processor software, MAXQDA and audio recording, and Tivian **Expected size and format:** MB scale text format; mx22, mp3; Tivian-format
DS. 4.6- Validation data of air traffic control. Datasets generated from human-in-the-loop evaluations with air traffic controllers and staff managers.	**Aim:** Related to the validation of the models created in the project **How is collected:** through a human-in-the-loop experiment **Type of data:** numerical and text (questionnaires) **Software:** BlueSky and Sector X **Expected size and format:** MB scale binary format. CSV.

TABLE 5 – WP4 VALIDATION AND IMPACT ASSESSMENT, DATA SUMMARY

WP5 Datasets	Aim for Data Collection
DS. 5.1 - Stakeholders map and contacts. Contact list of project stakeholders in different domains (AI, energy, mobility, etc.) and their feedback collected during meetings, workshops, and events. This includes the External Experts Advisory Board.	**Aim:** to develop a strategic ecosystem value map, including the protocol for identifying and analyzing the stakeholders. Organization of ecosystem engagement events to share knowledge and to foster synergies between the different initiatives. **How is it collected: Survey and questionnaires, newsletter subscription, and** feedback during online meetings (video). **Type of data:** E-mail, contact, text (name, company), video. **Software:** LimeSurvey, webpage, MS Teams. **Expected size and format:** MB scale, .xlsx format, .mp4 format.
DS 5.2 - Dataset provided for AI competitions. The dataset will include synthetic data, which are network models and time series to be used by competition participants. An example of an AI challenge run by RTE in 2023 for the power grid is the l2rpn_idf_2023 dataset, which is a 12GB CSV file created on the IEEE 118 grid. It includes time series with loads, productions, and next timestep forecasts for 500 years at 5-minutes resolution (network, generator, demand, wind, and solar signals).	**Aim:** To increase social, academic (AI community), and business awareness of AI potential. **How is collected:** Data will be either taken from a past AI challenge (e.g., l2rpn_idf_2023 dataset), created from real-world operational data, or entirely generated by the digital model considering the system's physical characteristics. Type of data: Numerical **Software: a** digital environment where the solutions are evaluated. **Expected size and format:** GB scale .csv format.

TABLE 6 – WP5 DISSEMINATION, COMMUNICATION, AND EXPLOITATION OF RESULTS, DATA SUMMARY

1. FAIR DATA

This section describes the guiding principles for promoting the FAIRness of AI4REALNET data by describing high-level practices to be adopted during the project. Although AI4REALNET will embrace an approach that is as open as possible, data availability must be considered on a case-by-case basis, according to the requirements that the distinct types of data may have. Further updates will report on the availability status of the different datasets.

AI4REALNET is committed to the following underlying principles:

- The data collected and generated during the project will be made available to project partners and made openly available concerning possible embargo periods through data repositories.
- Persistent identifiers are assigned to the data deposited in repositories and are included in the metadata.
- Metadata should be made accessible even when data is not publicly available.
- Metadata is based on standard vocabularies whenever possible, but customizable metadata may also be required.

1.1 DATA AVAILABILITY

As a general rule, AI4REALNET datasets described in the Data Summary section are accessible or can be shared between project partners with no specific conditions. However, some datasets need to be curated and anonymized before they can be shared. Table 7 shows which datasets have prior conditions to be shared within the project and those that must be prepared to be available to the different project partners.

Conditions for sharing within the project	Datasets
Data can be shared when anonymized	**DS. 4.4 – Data of network operator interactions and feedback produced during Human-AI evaluation scenarios** **DS. 4.5 - Feedback collected regarding the adoption of AI** **DS. 4.6- Validation data of air traffic control.**
Data can be shared with other partners after curation and anonymization. Operational data (real-time data) can be shared in real-time using a secure REST API	**DS. 2.1 -Experimental and operational data collection for operator personalized biomarkers**

TABLE 7 – CONDITIONS FOR DATA SHARING WITHIN AI4REALNET

AI4REALNET datasets, provided they do not convey personal or sensitive data, are to be made available as soon as possible, for instance, together with the correspondent deliverable or significant publications of project outputs. Table 8 informs us of the expected availability of each dataset.

Dataset	Expected availability
DS. 1.1 - Feedback from participants in the stakeholders' workshops	By the end of Task 1.2, it is not expected that this dataset will be made publicly available but only disclosed to the meeting´s participants. A summary

Dataset	Expected availability
	and anonymized feedback will be included in Deliverable D1.1.
DS. 1.2 - Datasets for the use cases and digital environments	The datasets will be part of the digital environments as pre-built data and/or embedded in the environment to generate additional synthetic data and scenarios for AI training and validation. Release in Deliverables D1.2-D1.4 (different software releases of the digital environment).
DS. 2.1 - Experimental and operational data collection for operational personalized biomarkers	After the experimental protocols and the operational tests in specific use cases, the data needs to be curated and fully anonymized to be publicly available. A dataset might be prepared to be shared within the INESC TEC open-source data repository by the end of WP2.
DS. 3.1 - State-action space for AI agent interaction with the digital environment	A curated set of interactions and results will be made available together with each WP3 deliverable, but earlier datasets can be published in the project GitHub.
DS. 4.3 - Test datasets for AI functional testing	The software code for the functional testing will also be released with the data. This should happen between months M30 and M42 in the project GitHub.
DS. 2.2, DS 2.3, DS. 2.4 - Socio-technical analysis with dispatcher and traffic controller	The results will be available with Deliverable 2.3, but the core results will be available in Report D4.3 of WP4.
• **DS. 2.5 - Knowledge-assisted AI performance** • **DS 2.6 - Hierarchical RL in relational domains** • **DS 1.3 - Interactive AI with order-agnostic architectures**	With the publication of the relevant algorithms in WP2 deliverables, namely in the software releases of D3.2 and D3.3.
• **DS 4.1 - Test datasets for evaluation of AI technical performance** • **DS 4.2 - Test results from the evaluation of AI technical performance**	They are published as part of the results of AI technical performance in deliverable D4.2. The code to generate the test datasets will be available in the project GitHub by month M42.
DS. 4.4 - Data of network operator interactions and feedback produced during Human-AI evaluation scenarios	Task 4.3, to be published as part of the work report from the project (anonymized).
DS. 4.5 - Feedback collected regarding the adoption of AI	Task 4.4, to be published as part of the work report (D4.3) from the project (anonymized).
DS. 5.2 - Dataset provided for AI competitions.	Task 5.2 is to be published for the competition.
DS. 5.1 - Stakeholders map and contacts	It is to be used only by the WP5 leader (INESC TEC).

TABLE 8 – DATASET EXPECTED AVAILABILITY

1.2 DATA REPOSITORY SELECTION

The AI4REALNET strategy for the publication of datasets will consider a set of different repositories to identify the repository solution that best fits the data requirements and targeted audiences. Therefore, AI4REALNET will adopt a flexible approach to enable the selection of the most appropriate data repository for each dataset, from generalist repositories to discipline or data-specific ones.

Regardless of their typology, repositories must meet the following conditions:

- Provide a persistent identifier.
- Apply for an open license.
- Ensure that the datasets are openly available with respect to the preservation needs of specific datasets.
- Enable the definition of access conditions if required.
- Are OpenAIRE compliant.

Moreover, AI4REALNET will also comply with existing data policies implemented by the journals where project results will be published, mainly if the submission to a specific repository is a condition for publishing.

In general, the catch-all repository Zenodo, a service hosted by CERN, will be adopted as a data repository for AI4REALNET, as it satisfies the aforementioned conditions if a better-suited repository is not identified for specific datasets. For this purpose, an AI4REALNET community was created on Zenodo from the start of the project. Moreover, Zenodo, as part of the OpenAIRE infrastructures, contributes to the open data movement in Europe.

NOTE: If AI4REALNET partners already have data repository instances in other services, like Figshare or B2Share, their use is encouraged to streamline data dissemination. However, project outputs deposited outside the Zenodo Community must be reported to the DMP manager so that they are registered in the DMP and added to the datasets in the online version of the DMP in the ARGOS service, with the necessary identifiers.

1.2.1 AI4REALNET ZENODO COMMUNITY

The AI4REALNET community policy specifies that it is for the exclusive use of sharing research outputs related to the project. All project partners with a Zenodo account can add their research outputs to the community.

The publication of content added by project partners requires the approval of the community curator, the DMP manager, who in turn will list the available outputs in the DMP updates.

The AI4REALNET Community can be accessed via: https://zenodo.org/communities/ai4realnet

The uploads to the AI4REALNET Community can be made via:
https://zenodo.org/uploads/new?community=ai4realnet

In each upload, AI4REALNET must be added to the funding awards section, as depicted in Figure 1.
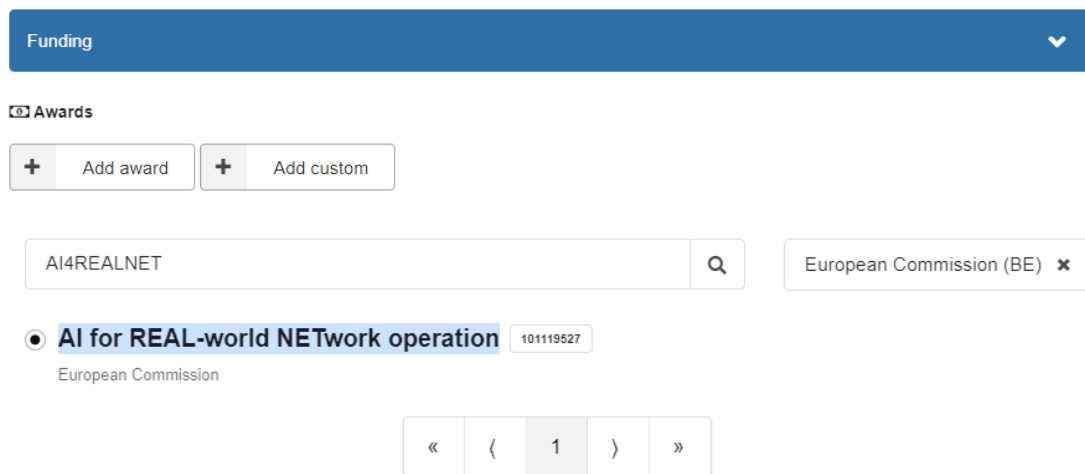
Figure 1 **– ADDING AI4REALNET FUNDING AWARD IN ZENODO**

1.2.2 GITHUB REPOSITORY SOFTWARE IDENTIFIER

AI4REALNET GitHub repositories will also be archived in Zenodo so that they are minted with a DOI if these repositories are open access and a license is clearly defined.

To preserve their GitHub repository in Zenodo, AI4REALNET partners can access this page: https://zenodo.org/account/settings/github/

In this process, it must be ensured that the registration is associated with the AI4REALNET Community.

1.2.3 INSTITUTIONAL DATA REPOSITORY

The INESC TEC institutional data repository can also support the publication of AI4REALNET datasets. Any user can access the INESC TEC data repository at any time and from any location. Therefore, the deposit data will be available for both project members and the general research communities for as long as the data is deposited in the INESC TEC data repository. The repository is registered in the repository's directory re3data.org.

Deposited datasets are described with a minimal set of Dublin Core metadata, including citation information and the DOI minted via the DataCite Fabrica service.

The deposit of data is the responsibility of the DMP manager and data steward at INESC TEC, and it must be coordinated with the responsible(s) for managing the datasets in each specific task or WP. The data backups are performed daily, while tape backups are performed weekly.

1.2.4 AI-ON-DEMAND PLATFORM

As part of the AI4REALNET strategy for data sharing, the project will be registered in the AI-on-demand platform. AI4REALNET will provide use-cases and AI assets (i.e., open-source software developed in WPs 2 and 3, digital environments from WP1) for strengthening the AI-on-demand platform catalog and sustaining the AI4REALNET concept/software beyond the lifetime.

The virtual lab can also host the digital environments and make them available to the AI community.

1.3 METADATA

Both the INESC TEC data repository and Zenodo already follow appropriate metadata standards, specifically Dublin Core, a highly adopted, domain-agnostic metadata standard published as ISO Standard 15836 :2017. Metadata records based on the Dublin Core standard promote findability and interoperability.

Therefore, a minimal metadata set for each dataset must consider the following categories:

- Administrative metadata for resource management (access rights, license).
- Descriptive metadata to enable the discovery, identification, and selection of datasets (keywords, authors, dates, file size, and format).
- Structural metadata describing hierarchical structures between resources.

Metadata elements for capturing the scientific production context of each dataset should be evaluated on a case-by-case basis. The use of standards for data description and encoding formats will depend on the requirements of each dataset and the setting in which they will be shared.

1.4 RECOMMENDED PRACTICES FOR AI4REALNET PARTNERS

1.4.1 DATA DOCUMENTATION

Data sharing within and outside the project must include the necessary metadata and documentation for others to understand and reuse the data.

Data documentation includes research protocols, codebooks, software syntax, equipment setting information, and instrument calibration.

AI4REALNET's data may need to be documented at various levels, and a Readme file, in the absence of another type of file, must be included as part of the dataset.

1.4.2 LICENSING

AI4REALNET datasets will be made available under Creative Commons (CC) by default. CC licenses are well-suitable for research data due to their conformity to both copyright law and database rights, and at the same time, they are easily readable by end-users.

The license Attribution-ShareAlike 4.0 International (CC BY-SA 4.0) specifies that:

- Credit must be given to the creator.
- Adaptations must be shared under the same terms.

CC Zero is also a license advocated for sharing research data, allowing end-users to use and reuse data without restriction.

In the case of code repositories to be registered in Zenodo (essential to be registered as part of the project products in OpenAIRE), a point of access will be created to the code repository by defining the most suitable relation between the records in *Related Works,* e.g. [*type of relation* – Software*: URL of the source code*], or in the *Software* section by completing the fields related to the Repository URL, Programming Language, and Development Status. Therefore, the relation to the original resource will

be clearly defined in the *External Resource* section in Zenodo [Available in], ensuring that users will have access to licensing information about the code repositories.

1.4.3 FILE NAMING

File names are an essential data identifier throughout the data life cycle. Not only does it make it easy to access data, but it also helps to understand what a data file is and its content. File names must remain meaningful and valuable beyond their original creation and storage location, particularly in shared environments.To allow ease of access and understanding of the data, some elements must be considered when implementing the file name strategy. The file name must provide metadata about its content.

Recognizing that the same strategy cannot be applied to the diversity of datasets that AI4REALNET will generate, the DMP does not define a policy for naming datasets, enabling the necessary flexibility for the data creators to specify the most suitable convention for their folders and file's structure. However, when adopting a file name strategy, elaborating a quick access guide is recommended, enabling partners to decode the elements that make up the filename quickly.

Despite the necessary flexibility, some conditions must be considered:

- Context information. The components of the file names must include specific or descriptive information (essential attributes).
- Consistency. The adopted convention must be followed systematically.
- Short and relevant names.
- The use of special characters is to avoid.
- The use of underscored or hyphens is recommended over full-stops or spaces.

1.4.4 FILE DATE FORMATTING

Dates included in the file name must follow the format Year-Month-Day to maintain the chronological order and simplify the process of sorting and browsing data files. This ensures compliance with the international standard for the representation of date and time, ISO 8601 (https://www.iso.org/iso-8601-date-and-time-format.html).

1.4.5 VERSIONING

For versioning purposes, it is common practice to use consecutive numbering for major version changes, with decimals used for minor changes (v1; v1.1; v2.1; v2.2). However, too many similar or related files can make it confusing to access the data as the number of files grows. The priority is to retain a copy, if possible, of the original "raw" data or definitive version, together with a well-documented record of changes.

1.4.6 DATA AVAILABILITY STATEMENT

The Data Availability Statement may be required for the manuscript submission workflow. When publishing project results, AI4REALNET partners will consider the following elements in their Data Availability Statement:

- **For open datasets:** data repository or service from which the dataset can be accessed, DOI, citation information, license, list of data items.

- **For datasets with conditions of access:** the ethical reason why the dataset cannot be shared openly, link to a permanent record detailing restrictions and conditions for access.
- **Third-party datasets:** citation information.

2. DATA SECURITY

This section outlines a preliminary set of recommendations regarding data storage and backup, to be refined locally, depending on each partner's infrastructure. AI4REALNET also recommends particular attention to the requirements stated in ISO 27001 for the purposes of information security management, which can be further detailed in the following updates of this DMP.

To prevent the risk of data loss, it is recommended that data be stored in institutional network drives, which must be routinely backed up, and account authentication systems must be provided to prevent unauthorized access, preferably by resorting to strong passwords. Moreover, if the data contains sensitive or personal information, the use of VPN is recommended.

To manage and share data securely, specific data storage platforms are not recommended for long-term storage:

- External portable storage devices, such as external hard drives and USB drives, given their longevity uncertainty and how easily these can be damaged or lost.
- Popular cloud storage platforms aimed at the general public.

Personal Computers and Laptops are very useful for daily work. However, the DMP will not specify or recommend how partners should manage their data in this context. However, project partners must ensure that data backup is regularly made through suitable networked drives.

There is no one-size-fits-all approach, and AI4REALNET acknowledges the necessary autonomy for project partners when adopting the most suitable approach to data security.

- Data must be stored in at least two different media (if possible, three).
- Daily backup of network drives as primary data storage device.
- Weekly backup for long-term off-site data preservation.

2.1 STORAGE AND BACKUP

Most datasets will be stored in the INESC TEC drive with daily backups and offline storage. INESC TEC drive is an instance of Nextcloud – an on-premise file access and sync platform with collaboration capabilities hosted on a secure server at INESC TEC. The accounts are created and managed at the INESC TEC Lightweight Directory Access Protocol, which enforces password changes and reset procedures to be done by the user exclusively, without disclosing passwords with anyone else, including IT staff. A more detailed description is presented in Annex 4.

Table 9 provides an overview of the storage and backup strategy for AI4REALNET datasets.

Dataset	Storage and backup strategy
DS. 1.1 - Feedback from participants in the stakeholders' workshops **DS. 2.1 - Experimental and operational data collection for operator personalized biomarkers** **DS. 4.3 - Test datasets for AI functional testing** **DS. 4.4 - Data of network operator interactions and feedback produced during Human-AI evaluation scenarios**	These datasets will be stored in the INESC TEC drive with daily backups and offline storage.

Dataset	Storage and backup strategy
• **DS. 4.5 - Feedback collected regarding the adoption of AI** • **DS. 4.6 - Validation data of air traffic control** • **DS. 5.1 - Stakeholders map and contacts** • **DS 5.2 - Dataset provided for AI competitions**	
DS. 3.1 - State-action space for AI agent interaction with the digital environment	The files can be stored in INESC TEC drive; however, individual instances that may differ slightly will be held independently (and locally) by each partner using the digital environment.
• **DS. 2.2, DS. 2.3, DS. 2.4 - Socio-technical analysis with dispatcher and traffic controller** • **DS. 4.1 - Test datasets for evaluation of AI technical performance** • **DS. 4.2 - Test results from the evaluation of AI technical performance**	Local PC and INESC TEC drive.
• **DS. 2.5 - Knowledge-assisted AI performance** • **DS. 2.6 - Hierarchical RL in relational domains** • **DS. 1.3 - Interactive AI with order-agnostic architectures**	Files are stored locally, and backups are created using Surfdrive cloud storage. Storage on safe and backed-up cloud storage (e.g., Surfdrive) until publication in a reliable repository (e.g., UvA instance of Figshare).
DS. 1.2 - Datasets for the use cases and digital environments	The files can be stored in the INESC TEC drive. When curated, the project's GitHub will include dataset and data generation functions.

TABLE 9 – STORAGE AND BACKUP

2.2 PRESERVATION

AI4REALNET datasets should be preserved because the results can be used as a baseline for future benchmarks. The exceptional cases are covered in Table 10.

Dataset	Preservation needs
DS. 1.1 - Feedback from participants in the stakeholders' workshops	There is no specific need since this dataset is directly used to elaborate on the use cases and deliverable D1.1.
DS. 2.1 - Experimental and operational data collection for operational personalized biomarkers	The data collected is intended to be fully anonymized after six months, at this point, all information that could directly or indirectly identify participants will be destroyed.
• **DS. 4.1 - Test datasets for evaluation of AI technical performance** • **DS. 4.2 - Test results from the evaluation of AI technical performance**	The data does not need to be preserved after the project duration since the reuse value is low after that period. What is essential to preserve is the code that conducts the functional tests and ensures the reproducibility of the results.

Dataset	Preservation needs
• **DS. 2.2, DS. 2.3, D. 2.4 - Socio-technical analysis with dispatcher and traffic controller** • **DS. 4.4 - Data of network operator interactions and feedback produced during Human-AI evaluation scenarios**	There is no need for preservation. The output of data processing (such as knowledge mapping, functional resonance model, etc.) may need to be preserved.
• **DS. 2.5 - Knowledge-assisted AI performance** • **DS. 2.6 - Hierarchical RL in relational domains**	Data can, in principle, be re-generated as it is from a synthetic environment. Preservation is desirable but not critical.
DS. 1.3 - Interactive AI with order-agnostic architectures	Possible human interaction is the most crucial part to be preserved. It cannot be re-generated and offers the possibility of being reused (e.g., to train and evaluate other algorithms).
DS. 5.2 - Dataset provided for AI competition.	The raw data should be preserved for 12 months. The consolidated results should be preserved for at least two years after the project. The code in GitHub will ensure the reproducibility of the test scenarios.

TABLE 10 – PRESERVATION NEEDS

3. ETHICS AND LEGAL COMPLIANCE

AI4REALNET will adopt data protection by design and by default approach in compliance with the General Data Protection Regulation regarding the processing of personal and sensitive data and the respect for data subjects' rights. In addition, it embraces an ethics-by-design approach, taking special account of the possible implications of AI-based solutions to be developed or validated in the scope of this project.

Table 11 outlines expected ethics or legal issues regarding AI4REALNET datasets.

Dataset	Ethical or legal issues
DS. 1.1 - Feedback from participants in the stakeholders' workshops	Possible legal issues are foreseen if the data is publicly disclosed since it is personal and not anonymized; this is why it is not expected to be disclosed.
DS. 2.1 - Experimental and operational data collection for operational personalized biomarkers	Only anonymous and aggregated results may be disseminated/published in scientific publications (which may involve research teams from different institutions) with the consent of the participants (information available on the informed consent document).
DS. 1.3 - Interactive AI with order-agnostic architectures	If a dataset is extended to include human interaction data, standards for anonymization and privacy will be followed. Only decisions made will be logged; this explicitly excludes personally identifiable information.
• **DS. 4.4 - Data of network operator interactions and feedback produced during Human-AI evaluation scenarios** • **DS. 4.5 - Feedback collected regarding the adoption of AI**	Ethical or legal issues shall be investigated since data is generated following the intervention of humans or influence in a physical environment; in any case, this dataset must be anonymized.
DS. 5.2 - Dataset provided for AI competition.	Possible ethical or legal issues will be managed before the publication in the context of AI competitions.

TABLE 11 – EXPECTED ETHICAL OR LEGAL ISSUES

3.1 DATA MINIMIZATION AND STORAGE LIMITATION PRINCIPLES

AI4REALNET must only collect essential personal information data, which shall be retained no longer than necessary. This means that personal data should only be kept for the time necessary to carry out the purposes for which it was collected or the time required to comply with applicable law.

Personal data must be reviewed periodically to decide whether unnecessary identifying information is retained. Data retention limits are set until the end of the period needed to conduct the research.

Personal information must be safely deleted or destroyed if it is no longer needed. AI4REALNET will consider using appropriate software for the deletion of data and encryption key deletion to prevent the recovery of data stored in an encrypted container or disks.

Specific data protection procedures for destroying data must be defined in compliance with GDPR and the applicable legal framework.

3.2 APPROPRIATE SAFEGUARDS: ANONYMIZATION AND PSEUDONYMISATION

Personal data processing for scientific research purposes or statistical purposes shall be subject to appropriate safeguards for the rights and freedoms of the data subjects.

These safeguards include technical and organizational measures such as pseudonymization and data anonymization procedures, which must be evaluated and carried out by project partners, while considering the realization of the project objectives. Where those purposes can be fulfilled by further processing which does not permit or no longer permits the identification of data subjects, those purposes shall be fulfilled in that manner. Therefore, for instance, when the identifying information is no longer needed, direct identifiers should be removed, where possible, by deleting them or replacing them with pseudonyms.

3.3 VOLUNTARY PARTICIPATION AND INFORMED CONSENT

AI4REALNET partners shall implement ethical procedures to ensure the voluntary participation of human subjects in project activities/ studies, as well as the respect for the fundamental ethical principles of autonomy, beneficence, non-maleficence and justice and applicable data protection legal requirements. The partners involved in activities or studies involving human subjects or personal data shall comply with their internal procedures for ethical assessment and clearance (namely, the approval of an institutional review board / ethics committee) before those activities take place. As a default rule, the leading partner in such an activity or study will be in charge of getting ethical approval.

AI4REALNET partners will always ask for informed consent in all activities involving human participants, provided the approval of an ethical committee, explicitly stating that participation is voluntary and that anyone has the right to refuse to participate and to withdraw their participation or data at any time without consequences. Annex 5, provides an illustrative provisional example concerning a specific study to be led by the Coordinator, INESC TEC.

Human participants will be informed about the aims, methods, and objectives of data collection, the benefits, and risks, as well as in which conditions the data may be shared, in accordance with the requirements stated in Article 13º or 14º of GDPR.

Notwithstanding the above, each partner shall assess the legal basis applicable to the respective activities/ studies whenever personal data processing is required, and perform in a timely manner the assessment as to the need to perform a DPIA. Furthermore, we recall, quoting the EC guidelines on identifying ethics issues in EU Funded research, that: *"Even if service providers and external collaborators are engaged in the research, the obligation to safeguard data subject's rights and freedoms rests with the principal researchers (e.g., the beneficiary and partners of a consortium). This obligation cannot be 'outsourced' or delegated (e.g., when surveys are conducted or data is processed or hosted by third parties or subcontractors)."*.

3.4 ROLES AND RESPONSIBILITIES

Partners involved in personal data processing activities shall assess their roles and responsibilities in relation to such activities according to the GDPR (as data controllers, joint controllers, or data processors) and, whenever necessary, shall celebrate the adequate contractual instruments in accordance with the law and the consortium agreement. Also, internally, project **partners shall take care of an adequate attribution of roles and responsibilities among teams involved in data processing and put in place suitable policies, procedures, and organizational measures, including continued** training, in order to ensure compliance.

Project partners collecting personal data must have a person responsible for monitoring GDPR compliance. The Data Protection Officer (DPO), when applicable.

3.5 TRUSTWORTHY AI

As explained above, AI4REALNET project will deal with and study the development, deployment, and use of AI and other new and emerging technologies in areas of high risk, such as the operation of essential critical infrastructure and services. Therefore, the project adopted the Assessment List for Trustworthy Artificial Intelligence (ALTAI) as a comprehensive tool for self-assessment across various dimensions.

Specifically, in Task 1.1 (*"Definition of AI4REALNET Assessment List for Trustworthy Artificial Intelligence ET conceptual framework"*), ALTAI is used to shape the conceptual framework, providing a robust foundation for the project's goals. Simultaneously, in Task 1.2 (*"Design of use cases and KPIs"*), ALTAI has been adeptly employed within the context of use cases, serving as a mechanism to identify and address gaps within both functional and non-functional requirements. For the use cases design, it will be used with the following process:

- All involved parties should review the responses in the ALTAI document and add insights, comments, and perspectives. Based on the responses to the questionnaire, a conscientious decision must be made:
    - Is the issue raised relevant and must be addressed in the Use Case?
    - This decision and its supporting arguments must be recorded in the ALTAI document for the Use Case.
    - If the issue is relevant, the partner should indicate which requirement(s) in the Use Case to address it.
- This procedure and the made decisions facilitate:
    - Is the issue raised relevant and must be addressed in the Use Case?
        - Ethical considerations and assessment satisfying the ALTAI.
    - This decision and its supporting arguments must be recorded in the ALTAI document for the Use Case.
        - Documentation of these considerations and assessment for the project.
    - If the issue is relevant, the partner should indicate which requirement(s) in the Use Case to address it.
        - Provide the means to validate that the identified issues have been adequately addressed.

This comprehensive evaluation extends to critical aspects such as privacy and data governance.

REFERENCES

European Commission (2016). H2020 Programme. Guidelines on FAIR Data Management in Horizon 2020. Version 3. July, 2016

Aguiar Castro, J. (2024). AI4REALNET: Data Management Plan structure and objectives. Zenodo. https://doi.org/10.5281/zenodo.10804943

Bessa, Ricardo (2024). AI4REALNET Project presentation: Ethics and Data Protection Committee (EDPC) meeting. Zenodo. https://doi.org/10.5281/zenodo.10804373

European Commission (2021). Identifying serious and complex ethics issues in EU-funded research. Version 5. July, 2021

ANNEX 1 – DMP INFORMATION GATHER TEMPLATE

DMP information – Partner

This document, based on the *The Data Curation Profiles Toolkit Interview Worksheet*[1] and the *Guidelines on FAIR Data Management in Horizon 2020*[2], aims to gather information to support the development of the DMP.

For any doubt, please contact: João Aguiar Castro (DMP manager), joao.a.castro@inesctec.pt

Data Summary

Please provide a description of the datasets you are expected to generate. Please add as many lines and dataset information as you seem fit. Identify subtask if needed.

Dataset description

Task	Dataset	General information
	Name of the dataset (as detailed as possible) The dataset name will be used throughout the form tables.	**Type of data:** **How the data will be collected:** **Software, instrument or tool required to process the data:** **Expected size and format:** **Update frequency:** **Personal or Sensitive data:**
		Type of data: **How the data will be collected:** **Software, instrument or tool required to process the data:** **Expected size and format:** **Update frequency:** **Personal or Sensitive data:**
		Type of data: **How the data will be collected:**

---

[1] Carlson, Jake (2010) The Data Curation Profile Toolkit: Interview Worksheet. Purdue University

[2] European Commission (2016) Guidelines on FAIR Data Management in Horizon 2020

		Software, instrument or tool required to process the data:
		Expected size and format:
		Update frequency:
		Personal or Sensitive data:

How the dataset relates to the project´s objectives? A brief description to contextualize the dataset in the overall AI4REALNET objectives.

Dataset	Aim and relation with the overall objective of the project

FAIR data

Can you specify <u>when</u> and <u>under which conditions</u> the datasets can be shared within the project? Taking the data lifecycle into account, it is expected that other project partners can access the data.

Dataset	Conditions for sharing the dataset within the project

Can you specify <u>when</u> and <u>under which conditions</u> the datasets can be made publicly available? For instance, whether the dataset can be made available in a data repository as soon as results are published (at the time of submission, or publication), or whether access conditions have to be defined (under embargo or request).

Dataset	Expected availability of the dataset

Will the dataset be created with interoperability in mind? In which way? If the datasets use formats and are collected with methodologies that promotes the integration with other data, applications or workflows. For instance, standardized disciplinary vocabularies, or other standardized practices, promotes interoperability.

Dataset	Interoperability

What kind of documentation will be produced to enable easy access to and the use of the datasets? Data documentation examples are readme.files, methodology information, codebooks, variables definitions, and others, any disciplinary standards to enable data sharing.

Dataset	Needed data documentation

Are there any ethical or legal issues that can have an impact on data sharing? If so, what?

Dataset	Ethical or legal issue

Are you aware of any data policies of the journals where it is expected to publish results? Can you please name some of these journals? This information can be used to identify possible editor's data policies to timely address them and include them in the DMP. For an example, see the Elsevier research data policy

Data storage and organization

Will the dataset be created or used in collaboration with different project partners? If yes, fill in the table with the partners corresponding to each dataset.

Dataset	Partners

What data storage and backup strategies will be adopted? If there is any backup periodicity, how many copies of the datasets and what storage solutions (e.g. partners institutional drives combined with offline storage devices).

Dataset	Storage and backup strategy

What are the most important parts of the data to preserve (manage and maintain over time) and under which conditions? Not all the data needs to be preserved or published in a data repository. Does a particular dataset have future reuse value? Is the processed data more critical than the raw data?

Dataset	Need for preservation and conditions

Who are the people responsible for data management? For instance, the people responsible for systematically ensure that the data is fit for use. A point of contact within the WP (or task) to data management related issues.

Dataset	Responsible for the dataset management

Other

Any additional notes you may think necessary to provide more information about your datasets.

ANNEX 2 – EDPC MINUTES OF MEETING

Date: 1 March 2024

Number of participants: 17 participants (14 partners represented)

Online comments and feedback

- ALTAI framework: Who is answering it? The work was divided by domain and partners, but it is mainly AI designers and end-users.

- Data management plan will be shared in Word format.
- How many human operators or experts are involved? We expect around 6 or 8 persons (~2 human experts per domain).
- Who is the responsible data controller?
  - We may need an additional contract concerning data protection between the partners, defining roles and responsibilities (e.g., joint controller and data controller roles).
  - We should analyze if we are joint controllers.
- Data collection: Will an informed consent be shared? It should be included in the document. Include a template in the appendix of the document
  - INESC TEC can provide a provisional version of the consent form to be used by INESC TEC in the activity that was mentioned in the meeting - the one involving the use of INESC sensors to collect data from a group of participants. So far, my understanding is that this activity is not yet fully confirmed, but it will be conducted exclusively by INESC TEC and therefore, at least for now, it would indicate only INESC TEC as data controller. But mentioning its provisional nature.
- Data storage: how long will it be? This could be specific to each data type and should be defined in future versions of the DMP.
- Anonymization: Are there any standards from the consortium, or is it up to the consortium to do this?
- Types of AI: No generative AI involved.
- Data access rights within the consortium. How can we share this with the partners? It must be made clearer in future versions.
- The need for a Data Protection Impact Assessment or separate assessments shall be evaluated by project partners as soon as possible, given that such kind of assessment should be made at the beginning and not at the end of the project. So far, a Fundamental Rights impact assessment has been made as a part of the ALTAI framework).
- The ALTAI framework is being made in parallel with the DMP. We are applying it in the design phase.
- How frequently the DMP can be updated? Ideally, it should be a working document. Updated every 6 months. It should depend on the lifecycle of each dataset. The DMP gather information template (Annex 1) will be used to support the DMP updates.

ANNEX 3 – DIGITAL ENVIRONMENTS DESCRIPTION

Grid2Op: power grid

RTE developed the open-source Grid2Op environment to model and study a large class of power system-related problems and facilitate the development and evaluation of controllers (or agents) that act on power grids. Any type of control algorithm in interaction with a virtual version of the electrical grid can be used to overcome gaps between research communities.

Through different L2RPN competitions, calibrated virtual environments have been instantiated for testing over robustness to adversarial attacks, adaptability for increasing renewable energy share, or agent alert trustworthiness. Such "autonomous" agent scenarios can already be visualized and analyzed through the Grid2viz module. Moreover, it is also possible for a human to play live scenarios, assisted by an AI agent with the Grid2Game module. A human can choose contextual triggers for alerts and get recommendations from the agent. It can also do its manual simulation if needed. Ultimately, this will be run through the OperatorFabric Hypervision Interface in real operations. There is also a repository hosting a set of reference baselines.

Chronix2Grid package allows the generation of synthetic but realistic consumption, renewable production, electricity loss (dissipation) and economic dispatched productions chronic for a given power grid.

Flatland: Railway

The Flatland environment is a comprehensive framework developed (by industry partners like SBB, DB, and AI community) for easy development and experimentation on the vehicle rescheduling problem for railway networks. Flatland represents railway networks as 2D grid environments with restricted transitions between neighboring cells. On the 2D grid, multiple train runs must be performed for a given set of goals and circumstances. Trains are represented as agents that make decisions on movement and navigation.

Flatland is a discrete-time simulation, i.e., it performs all actions with constant time steps. A single simulation step synchronously moves the time forward by a constant increment, thus enacting exactly one action per agent. The Flatland environment is tailored towards RL. It provides observations and rewards to any controlling agent, and it expects one discrete action per agent per step. Flatland, in its current state, provides a set of global and local observations. It provides generators for generating railway networks and demand for trains (scenarios), an evaluation system, and mechanisms to inject disturbances into rail operations. These disturbances are represented as malfunctions of trains, i.e., trains being unable to move on the track for several time steps. The occurrence of these is distributed according to configurable distribution at scenario definition.

After three competitions with Flatland, a comprehensive set of basic (mostly) AI solutions for Flatland exists and can be used as baseline/benchmark models.

BlueSky: Air traffic management

The BlueSky environment is an open-source ATM simulator that has been developed since 2013 with TU Delft as its main developer. It contains open source, open data aircraft performance models and a

global navigation database including airports; it is also compatible with Base of Aircraft Data (BADA) v3.xx files (containing performance and operating procedure coefficients for 295 different aircraft types). Although BlueSky started out as a simulator aimed at conventional aviation, in recent years it has been extended with several Drone/Urban Air Mobility models and functionality and has since been applied in several UAM/UTM-related projects.

Through its modular setup, an extension of each of the components of BlueSky (e.g., autopilot, FMS, performance model, conflict detection and resolution, environmental modeling, visualization, etc.) can be reimplemented or extended. In the same way, it is also possible to add completely new functionality to the simulator. By default, BlueSky has its own Qt/OpenGL-based interface that allows the user to control the simulation and get an overview of the simulated traffic. Through its client/server network implementation, BlueSky can also easily interface with separate ATM user interface applications and piloted blip driver stations.

ANNEX 4 – INESC TEC DRIVE DESCRIPTION

Technical and security measures met by INESC TEC drive ensure:

- Deny unauthorized persons access to data processing equipment used for processing personal data (equipment access control). Nextcloud includes edition/collaboration tools that enable users to process data on the server side. Only authorized users can access, create, store, or edit those files. Computers at INESC TEC may be used for local data processing; these computers are restricted to users with INESC TEC accounts; each user has a local area without administration privileges to prevent them from accessing areas from other users.

- Prevent the unauthorized reading, copying, modification or removal of data. The Nextcloud instance enforces access control at folder or file levels to authorized users only. The permission control may restrict users from deleting or modifying data.

- Prevent the unauthorized input of data and the unauthorized inspection, modification or deletion of stored data (storage control). Data storage cannot be accessed directly by users, only through the Nextcloud instance, which prevents unauthorized input of data and the unauthorized inspection, modification or deletion of stored data. Servers hosting Nextcloud are hosted on a datacenter at INESC TEC with biometric access control, restricted to IT staff.

- Prevent the use of automated data processing systems by unauthorized persons using data communication equipment (user control). The Nextcloud instance prevents automated data processing by unauthorized persons.

- Ensure that it is possible to verify and establish to which bodies data have been or may be transmitted or made available using data communication equipment (communication control). The INESC TEC Nextcloud instance ensures traceability via activity logs, such as downloads, modification, access, data sharing, including automatic e-mail notifications.

- Ensure that it is subsequently possible to verify and establish which data have been input into automated data processing systems and when and by whom the data were input (input control).

- Ensure that the functions of the system perform without fault, that the appearance of faults in the functions is immediately reported (reliability) and that stored data cannot be corrupted by means of a malfunctioning of the system (integrity). The Nextcloud instance is hosted on servers being actively monitored, with redundant power supplies connected to UPS, and with redundant network connectivity. The Nextcloud instance provides file versioning and recover from data file deletion; moreover, the data is stored on a filesystem with file checksum, on a storage server with redundant disk parity, redundant power supplies connected to UPS, with redundant network connectivity, and with daily backups.

Data on the Nextcloud instance can only be inserted, stored, accessed, inspected, modified, or deleted by authorized users only.

Data transfers between the Nextcloud instance and user client computers/applications is done over secure, encrypted/authenticated communications (HTTPS).

Servers hosting Nextcloud are hosted on a datacenter at INESC TEC with biometric access control restricted to IT staff. The Nextcloud instance provides file versioning and recovery from data file deletion; the data is stored on a filesystem with a file checksum and daily backups. Therefore, depending on the level of data recovery needed, data can be recovered from the mechanisms provided by Nextcloud or from external backups. Besides short-term data backups, data can be stored on a long-term encrypted tape archive, with a second copy off-site for disaster recovery.

ANNEX 5 – INFORMED CONSENT

INESCTEC

INFORMED CONSENT TO PARTICIPATE IN A RESEARCH PROJECT

Please read the following information. If you think something is incorrect or unclear, don't hesitate to ask for more information through the email: (Duarte Filipe Dias duarte.f.dias@inesctec.pt)

If you wish to participate in the study we request your consent, which you can give by signing the document.

Participation in the study is voluntary. You can quit it at any time without any consequence, needing only to contact the person responsible through the email specified above.

1.PROJECT DESCRIPTION

Title: AI4RealNet human-in-the-loop integration

Responsible Entity: INESC TEC

Principal Investigator: *Duarte Dias* – **duarte.f.dias@inesctec.pt**

General Description: The Center for Biomedical Engineering Research (C-BER) at INESC TEC has been developing for 10 years several wearable technologies focused on the monitoring of first responders in hazardous environments, which is integrated into the research line of quantified occupational health. Besides the focus on wearable devices, this research line has also made a strong effort to combine algorithms based on physiological signals with psychological information, to understand the impact of psychosocial risks, such as stress in user's health, using psychophysiological information. Our wearable devices are capable of monitoring physiological signal, namely electrocardiogram (ECG), respiration and temperature.

The current study is integrated in the **AI4RealNet European project** (Horizon Europe; 101119527) with the aim to research innovative methodologies for the integration of human conditions in the Decision Support system that aim to support the operator of critical infrastructures during the use of digital environments from three different domains: electric power grid, railway and air traffic management.

2.PROCESSING OF PERSONAL DATA

Objectives: The current study aims to understand the user psychophysiological changes (e.g., stress and fatigue levels) during the different trials that will be made.

- **Procedure:** Participant will be requested to collect data during 3 periods according to the methodology defined: 1) baseline data collection procedure – data collected outside the working environment; 2) standardized laboratory stress procedure – Trier Social Stress Test - TSST, along with a 2-Choice Reaction Time Task; 3) Operator monitorization in real-time during the use of digital

environments. The following systems, fully controlled by INESC TEC, will be used during these procedures:

1) **VitalSticker** that measures ECG, Heart Rate, Respiration, core temperature estimation, activity and posture based on a chest band with textile conductive electrodes (no need for gel electrodes).

2) **Smartphone**: will contain an App to easily connect to the devices and place it in the user's pocket. The application does not aim to be used during the test; it just needs to be initiated to collect all the information from the devices and then stopped and the end of the tests.

3) **Surveys:** a sociodemographic questionnaire will be used, along with a stress and fatigue scale. The sociodemographic questionnaire should be filled at the beginning of the tests and the scale at the beginning and end of tests.

4) **Computer:** only used on standardized laboratory stress procedure by the research team to perform the above-mentioned tasks.

Personal Data:

- ECG, Heart Rate, Respiration, Core temperature estimation, activity and posture will be collected with Vitalsticker, a chest band with textile conductive electrodes;
- Sociodemographic, stress, and fatigue levels information will be collected using surveys

Purposes of Personal Data Processing: The data collected will be processed in accordance with the applicable national and EU legislation and will only be used by the researchers for scientific research purposes in the domains of quantified occupational health and AI decision support systems.

Data Controller(s): INESC TEC – Instituto de Engenharia de Sistemas e Computadores, Tecnologia e Ciência, Campus da FEUP, Rua Dr. Roberto Frias, 4200 - 465 Porto.

Confidentiality, Security Measures and Data Retention Period: All data acquired will be analysed for scientific purposes within the framework of the study in question; all those involved in this study have a commitment to confidentiality and non-disclosure of data or personal information taken from it; the data collected will be pseudonymised and aggregated at a later stage so that your individual participation in the study will not be identified. Some steps will be taken to protect your privacy, namely: a) each participant will be assigned a randomized code; b) all data collected will be stored on safe local servers at INESC TEC with restricted, controlled access, limited only to members of the research group involved; c) the data collected is intended to be anonymised after 6 months, at which point all information that could directly or indirectly identify participants will be destroyed.

Personal Data Sharing: Only anonymous and aggregated results may be disseminated/published in scientific publications, which may involve research teams from different institutions.

Incidental Findings:

In addition, if any problems or physiological anomalies are detected in the data collected, would you like to be informed?

Please indicate: _____YES _____ NO

If you answer yes, please indicate your contact details (email or phone):_____.

Rights of the Data Subject: As the subject of the data, the law grants you the following rights: Information, Access, Rectification, Portability, Erasure, and Restriction of processing. In the event of withdrawal, there is no prejudice to the processing of data collected up to that point, but you may withdraw your consent and thus request the deletion of your personal data, provided that you exercise this right before the respective and irreversible anonymisation takes place, which will occur after 6 months. To exercise any of your rights, please use the following e-mail address: duarte.f.dias@inesctec.pt.

Data Protection Officer: For any questions regarding the processing of your personal data, please contact our Data Protection Officer at: dpo@inesctec.pt.

Under the terms of the Article 77 of the GDPR, the data subject also has the right to lodge a complaint with a supervisory authority in the European Union. In Portugal, the supervisory authority is the CNPD (www.cnpd.pt).

3. INFORMED CONSENT FORM

1. I read and understood the information about the project, including the identity of the controller, the type of data collected, the purpose of the collection and the respective processing. ☐

2. I read and understood the information about how the data is stored and for how long, including what will happen to my data in the case of ceasing my participation in the project. ☐

3. I have been given the opportunity to ask questions and to clarify any doubts about the project. ☐

4. I understand I can quit participating in the project at any point, without needing to provide any justification and without suffering any penalty or having my motives questioned. ☐

5. I understood how to communicate my decision to quit, as well as how to exercise my rights as the data subject. ☐

The Participant:

I declare I have read and understood this document, as well as the verbal information that have been given to me previously. As such, I accept to participate in this study and allow the use of the data I offer voluntarily.

Name: … … … … … … … …... … … … …... … … … … … … … … … … … …

Signature: … … … … … … … …... … … … … … …... … … … … … … … … … … … …   Date: …… /…… /………..