Funded by
the European Union

Project Acronym	**RAISE**
Project Name	**Research Analysis Identifier SystEm**
Project Number	**101058479**
Topic	**HORIZON-INFRA-2021-EOSC-01-04**
Type of Action	**HORIZON Research and Innovation Actions**
Granting Authority	**European Research Executive Agency**
Project starting date	**1 October 2022**
Project end date	**31 January 2026**
Project duration	**40 months**

D1.2 Data Management Plan

(Version 3.0, 31/03/2023)

Deliverable Number	D1.2	
Deliverable Name	Data Management Plan	
Work Package No	WP1	
Due Date (month)	31/03/2023	
Submission Date	31/03/2023	
Lead Beneficiary	OpenAIRE AMKE	
Version	3.0	
Status	Review	
Author name(s)	Elli Papadopoulou (OpenAIRE)	
Contributor(s)	Maria Kontopidi (ATHENA), Evdokimos Konstantinidis (AUTH), Christoniki Manga-Nteve (CERTH), Gorka Epelde (VICOM)	
Reviewer(s)	Panagiotis Kartsidis (AUTH)	Spyros Karakitsios (ENVE.X)
Approved by	Evdokimos Konstantinidis (Project Coordinator)	
Keywords	data management, data management plan, DMP	
Type	R – Document, report	
Dissemination level	PU - Public	

The **RAISE** Consortium consists of:

Participant #	Participant organisation name	Short Name	Country
1 COO	ARISTOTELIO PANEPISTIMIO THESSALONIKIS	AUTH	EL
2 BEN	FUNDACION CENTRO DE TECNOLOGIAS DE INTERACCION VISUAL Y COMUNICACIONES VICOMTECH	VICOM	ES
3 BEN	NETCOMPANY-INTRASOFT SA	INTRA	LU
4 BEN	PANEPISTIMIO DYTIKIS MAKEDONIAS	UOWM	EL
5 BEN	UNIVERSITY OF MACEDONIA	UOM	EL
6 BEN	TECREANDO B.V	TECREANDO	NL
7 BEN	WITA SRL	WITA SRL	IT
8 BEN	ETHNIKO KENTRO EREVNAS KAI TECHNOLOGIKIS ANAPTYXIS	CERTH	EL
9 BEN	INNOV-ACTS LIMITED	INNOVACTS	CY
10 BEN	VILABS (CY) LTD	VILABS	CY
11 BEN	KYKLIKI VELTISTOPOIISI IKE	Cyclopt PC	EL
12 BEN	KAROLINSKA INSTITUTET	KI	SE
13 BEN	ATHINA-EREVNITIKO KENTRO KAINOTOMIAS STIS TECHNOLOGIES TIS PLIROFORIAS, TON EPIKOINONION KAI TIS GNOSIS	ATHENA	EL
14 BEN	AINIGMA TECHNOLOGIES	AINIGMA	BE
15 BEN	UNIVERSITEIT HASSELT	UHASSELT	BE
16 BEN	OPENAIRE AMKE	OPENAIRE	EL
17 BEN	CENTRO EUROPEO DI FORMAZIONE E RICERCA IN INGEGNERIA SISMICA	EUCENTRE	IT
18 BEN	ENVE.X SINGLE MEMBER PC	ENVE.X	EL
19 AP	UNIVERSITY OF SOUTHAMPTON	SOTON	UK
20 AP	ROYAL INSTITUTION FOR THE ADVANCEMENT OF LEARNING MCGILL UNIVERSITY	McGill	CA

Revision history			
Version	**Date**	**Modified by**	**Comments**
0.2	20/03/2023	Elli Papadopoulou (OpenAIRE AMKE)	Table of Contents and Initial Draft
0.7	24/03/2023	Maria Kontopidi (ATHENA), Evdokimos Konstantinidis (AUTH), Christoniki Manga-Nteve (CERTH), Gorka Epelde (VICOM)	Contributions
0.9	27/03/2023	Elli Papadopoulou (OpenAIRE AMKE)	Formatting and preparation for internal review and Final Changes
1.0	28/03/2023	Elli Papadopoulou (OpenAIRE AMKE)	Document for Internal Review
2.0	31/3/2023	Elli Papadopoulou (OpenAIRE AMKE)	Integrate reviewers comments
3.0	31/3/2023	AUTH	Ready for submission

Table of Contents

Abbreviations

EC	European Commission
EU	European Union
GA	Grant Agreement
REA	European Research Executive Agency

WP	Work Package

Executive summary

This DMP reflects the data management activities of the RAISE Horizon Europe project. RAISE will develop a system for data analysis that adheres to the FAIR principles, i.e. is FAIR-enabling, following best practices and recommendations that connect data outputs and workflows offering to researchers a trustworthy resource for reproducible science.

RAISE deals with data handling and processions and therefore a specific task has been introduced to make sure that all data is properly accessed, curated, preserved and eventually made public for analysis after any possible ethical issue has been cleared. The RAISE DMP will capture those elements and continuously describe the research data, code and workflows resulting from use cases and pilots of RAISE components to the research community, together with underlying data of RAISE publications (e.g., the outcomes of the mobility pilots will be publications and studies on mobility patterns and driving behaviour). It will also provide insight on the different ways that the system addresses FAIR principles for data and workflows and will be openly published to be utilised as a guide for researchers to consult when describing the FAIRness of RAISE outputs in their own DMPs.

The RAISE DMP will be treated as a “living document” that will be continuously updated to record the changes in the decisions as well as data and software management practices followed by the consortium.

1 Introduction

1.1 The RAISE Project

The mission of RAISE is to provide the infrastructure for a distributed crowdsourced data processing system, moving from open data to open access data for processing. RAISE will provide the mechanism for sending the algorithm to the dataset instead of sending the data to the algorithm. The real value of open data for the research community is not to access them but to process them as conveniently as possible to reduce time-to-result and increase productivity. RAISE aims at promoting a transparent way of sharing and processing data, enabling the research community to publish their work with evidence-based authenticity of the data-analysis performed, ensuring at the same time the accreditation of their work.

RAISE will be grounded on the fundamental principle defined in the FAIR Guiding Principles for scientific data management and stewardship (Findability, Accessibility, Interoperability and Reusability). To do so, RAISE brings the processing algorithm (small size) to the dataset (large size) instead of downloading the dataset to the computer where the processing algorithm is. To increase the processing capacity of the dataset repositories, RAISE borrows the crowdsourcing concept where researchers can easily integrate in the existing workflows computers serving both their datasets and the processing capacity.

1.2 Description of WP1

Project management will cover all financial, administrative, scientific, as well as knowledge and innovation aspects. The different Management levels are divided among the WP leaders and include: (1) the project coordinator (AUTH), (2) the quality manager (AUTH), (3) the technical manager (VICOM), (4) the data protection and security manager (UOWM), (5) the piloting and mobility manager (CERTH), and (6) the dissemination and exploitation manager (VILABS). The day-to-day activities of RAISE project will be monitored by Project Coordinators, by Work Package Leaders and by the Executive Board. They are supported by the International Advisory Board / External Expert Advisory Board and the Ethics Committee and all accountable to the Plenary Board/General Assembly.

1.3 Structure of the deliverable

The structure of D1.2 “Data Management Plan” follows the European Commission’s Horizon Europe Data Management Plan Template topics answering to the questions with information that members have in the initial phase of the project. It should be noted that answers relevant to Section “3. Other Outputs” from the EC’s template are embedded in all sections where relevant answer could be provided for RAISE software and code. Follow up versions of the current DMP will be made available in future iterations showing further progress and decisions made around data management processes and practices.

A version of this DMP can also be found as a machine actionable and FAIR output produced by ARGOS service (argos.openaire.eu): https://zenodo.org/record/7789650#.ZCcRXnZBy5c.

2 Data Summary

2.1 Will you re-use any existing data and what will you re-use it for? State the reasons if re-use of any existing data has been considered but discarded.

RAISE will not generate new data, but instead re-use existing data and combine them with algorithms to produce new outputs.

2.2 What types and formats of data will the project generate or re-use?

The reused data will be exploited in the pilots that address the following themes: a. Health, b. Environment, c. Smart mobility, and d. Cross-disciplinary. Below, we provide the types and formats for those data:

A. Joint research in Health Domain

This pilot utilises the following datasets to harmonise them, run analysis with the same objectives and joint exploitation of results and explore shared IPR for algorithms development by companies with datasets that belong to other organisations.

- VITALISE dataset
    I. From Withings body+ scale we collect in .json format weight, fat percentage, hydration estimate, bone mass and BMI.
    II. From Fitbit Charge 4 we collect for 1 to 10 days in .json format the Heart Rate, SpO2 and sleep duration.
    III. From Smart Soles, we collect for 2 hours in .json format data regarding balance estimation and number of steps.
    IV. From WITA Mentorage we collect for 1 to 10 days in .json format features regarding body movement while in bed, measured by a depth camera.
    V. Demographics and Health record data in .csv format.
    VI. Questionnaires regarding Quality of Life, Functionality, Physical activity and cognitive status in .csv format.
- MS-Neuroplast dataset
    I. From the 128-channel Nihon-Kohden Electroencephalograph, we collect Event Related Potential (ERP) and Resting-State EEGs in .edf format.
    II. From Fitbit Charge 4 we collect for 90 days in .json format the step count, Heart Rate, SpO2 and sleep duration.
    III. From the BrainHQ cognitive training platform we collect 3-month cognitive training data (including training progression, the magnitude of training effect, rate of learning, progression against training levels, etc.) in .csv format.
    IV. From the Somatometric and Psychometric evaluations, we collect the results of said evaluations in .csv format.
    V. Demographics and Health record data in .csv format.

B. Joint research for Environment

Data, incl. personal data, have been acquired by individual GPS sensors measuring personal movement/activity and intensity of activity, environmental temperature and relative humidity. These datasets will be analysed and exploited to support modelling and simulation to derive accurate personal exposure assessment to ambient pollutants at individual level.

- Datasets per Sensors were the following:
    a. From the Garmin Vivosmart wearable sensors we collected data on intensity of activity, heart rate, stress level, steps, distance, sleep level, metabolic equivalent of an activity/task (MET).
    b. From the PPM IoTECH wearable sensors we collected data on Particulate Matter (PM1, PM2.5 and PM10) personal exposure levels, temperature and relative humidity.

c. From the silicone wearable wristband sensors we collected data on several organic compounds such as PAHs, dioxins, furans.
d. From the uHuoo static sensors (i.e., placed in the volunteer's households) we collected data on temperature, relative Humidity, PM2.5, Total Organic Compounds (TVOC), CO2, CO, NO2 and air pressure.
e. Additional datasets include daily diaries (time activity diaries (TAD)) as well as datasets on the housing conditions.

C. Joint research in Smart mobility solutions
The following datasets will be added to the RAISE system and will be exploited to support modelling and simulation, to train ML prediction engines, to extract mobility patterns, to provide routing services, to detect conflicts and abnormalities and to count pedestrians.

- Telematic data from vehicles (taxis, buses, e-scooters, bicycles).
- Video data from various sources (dashcam, external camera in cars, bikes and e-scooters, fixed surveillance cameras, drone cameras).
- Additional datasets include travel diaries; environmental sensors; data related to risk-related physiological indicators (driver related background factors, and driving environment and traffic complexity indicators); and digital transport network infrastructure (road network, public transport routes, schedules…).
- Data from external repositories

D. Cross-disciplinary research for the city of Thessaloniki
Data for Health, Mobility, Environment will be cross analysed in order to analyse the effect of climate and environment conditions, traffic and mobility in the health of the citizens of Thessaloniki.

E. Co-creation activities
There will be a number of co-creation activities to realise RAISE platform, workflows and features.

- Data will be collected by researchers during co-creation activities, which will include their opinions with regard to RAISE in text, .csv or .mp3 format.
- Video of participants in co-creation activities will be collected in .mp4 formats.
- Photographs of participants in co-creation activities will be collected in .jpg format.

2.3 What is the purpose of the data generation or re-use and its relation to the objectives of the project?

The main purpose of data reused in RAISE is to test the system, its different components and results (e.g., tools and services) in real life scenarios (O8) aiming to better support researchers pathways (O5) by reducing the analysis time (O7) and building trust in science and data sharing practices (O3).

2.4 What is the expected size of the data that you intend to generate or re-use?

The aggregated size from reused data in RAISE, so far, is collectively estimated to be 350GB.

2.5 What is the origin/provenance of the data, either generated or re-used?

RAISE will collect, make available and share secondary data, i.e. data that have already been collected through primary sources. The data will be made available from the RAISE platform to be used in the pilots.

More specifically, the data have been collected in the context of:

- H2020 projects, such as VITALISE, ICARUS, URBANOME, Big Data Europe
- Horizon Europe projects: iDREAMS, iCHANGE, Search&Rescue

- other initiatives and projects, such as INPERCEPT and Movilidad2030 from the Spanish CDTI.

Data also come from laboratory activities in AUTH, from Living Labs infrastructure, such as Big Data Warehouse for Mobility (BD W4M) in CERTH-HIT, or even from public bodies, such as MIA: Flemish Ministry of Transport.

2.6 To whom might your data be useful ('data utility'), outside your project?

Overall, RAISE data are useful for every researcher working in the fields of health, environment and smart mobility who would like to analyse their data without compromising privacy and time required for analysis. RAISE data are useful also to public bodies and authorities that perform similar activities to improve infrastructure and policies by utilising algorithms and technological means for evidence policymaking. Additionally, more Urban Living Labs can use RAISE data to tackle various levels of environmental exposures, age-dependent susceptibility windows, inter-individual variability, gender differentiation of exposure, and socio-economic disparities. Likewise, transport related institutes or even SMEs can use RAISE data to improve mobility footprint, create added value and services on the top of the collected data to return value to the ecosystem members and to create mobility services and apps for the citizens.

3 FAIR data

3.1 Making data findable, including provisions for metadata

3.1.1 Will data be identified by a persistent identifier?

All RAISE data will be assigned persistent identifiers from the Research Analysis Identifier (RAI) service deployed in the RAISE system. The RAI service supports a unique identifier of any result along with the dataset information and the processing script (RAI Registration Service) without disclosing any source code or raw data.

RAI identifier is a new, custom-made unique identifier for datasets, and the accompanying published analysis, that provides accreditation of the work. The results of each execution (RAI identifier) are linked with the encrypted processing script and the dataset of the application.

When the requested experiment computation is completed, the RAI Certified Node registers the results (together with source dataset and processing algorithm identifier and description) to a blockchain network through the RAI Registration Service, receiving a Research Analysis Identifier (RAI) (E) as the outcome of the registration process.

Moreover, RAISE will work on FAIR publishing of RAIs to ensure the DOI - RAI interoperability, so that the different elements involved in the RAI identified research experiment results (i.e. source dataset, algorithm and produced results) can be referenced and resolved using scientific community standard DOI identifiers. This is to ensure that RAISE datasets that already have a DOI can be interoperable with the RAI.

3.1.2 Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

RAISE will enable findability of RAIs through the RAI Finder Service, the Open Access Data for Processing Finder Service and the RAISE FAIR Catalogue and Portal for data discovery.

- The RAI Finder Service will classify and retrieve its processed results and source dataset identifier and metadata description according to a set of descriptive metadata following the OpenAIRE and Dublin Core/Datacite schemas.

- The Open Access Data for Processing Finder Service enhances the RAI Certified Nodes uploaded data with some metadata pertaining to size, format, origin and the RAI Certified node hosting the data.
- The RAISE FAIR Catalogue and Portal for data discovery will enable browsing and discovering datasets (available in different RAI nodes) indexed according to the RAISE model.

The Raise metadata model will capture information pertinent to the project use cases. The aim is to describe metadata regarding the resources, entities and relations that will be defined on the pilot's use cases. Initially, the metadata model will be based on the OpenAIRE data model (https://graph.openaire.eu/docs/data-model/entities/result/), where the provided dump files will be re-used i.e. result scheme to describe a dataset. In addition, the metadata model will be accompanied by a set of descriptive statistics which will enrich and annotate the existing data elements. Specifically, community specific data standards dedicated to each use case will be used to precisely represent users' data measurements deriving from different devices (i.e. heart rate and FHIR standard), which will be followed by descriptive coefficients.

Furthermore, the RAISE FAIR Results repository will enhance its metadata by interconnecting to the OpenAIRE Graph (https://graph.openaire.eu/).

3.1.3 Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use?

The Open Access Data for Processing Finder Service will make all the datasets of the network searchable, to facilitate research finding the proper dataset based on a number of search criteria.

3.1.4 Will metadata be offered in such a way that it can be harvested and indexed?

The indexing of the datasets available in the network of the RAI Certified Nodes in the Open Access Data for Processing Finder Service increases the findability of valuable datasets for the research community. Additional activities support information exchange and harvesting to reach out to wider communities, such as RAISE assets that will be published in Crossref through Zenodo (by integrating Zenodo API) and EOSC through the OpenAIRE Graph.

3.2 Making data accessible

3.2.1 Will the data be deposited in a trusted repository?

RAISE will develop a FAIR Results Repository to provide persistence and access (with required permission) to results produced by different RAI nodes, even when they’re down or not participating in the RAISE network. The final aim is to be able to provide uninterrupted access and response to RAI identifiers resolving requests. It will be developed on top of existing storage services (considering EOSC offered services adoption), implementing RAISE established data access policies. For repeatability, when the data owner repeats experiment with no script uploading, and replicability, when others test same algorithm against another dataset (if opt-in by owner), each processing script will be encrypted, stored and linked to the RAI identifier.

The aim of the repository is to gradually become a trusted service provider of certification and long-term preservation.

3.2.2 Have you explored appropriate arrangements with the identified repository where your data will be deposited?

The repository is hosted by RAISE, hence appropriate arrangements are being made throughout its development process. Examples are the access rules defined by each RAI node (offering protected datasets and algorithm processing).

3.2.3 Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?

RAI Registration and Finder services obtain, find and resolve (blockchain-based) immutable identifiers of performed research experiments (composed of the dataset, algorithm, and results combination). Scientific community standard DOI identifiers are part of the different elements involved in the RAI identified research experiment results (i.e. source dataset, algorithm and produced results) which can be referenced and resolved respectively.

3.2.4 Will all data be made openly available? If certain datasets cannot be shared (or need to be shared under restricted access conditions), explain why, clearly separating legal and contractual reasons from intentional restrictions. Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if opening their data goes against their legitimate interests or other constraints as per the Grant Agreement.

Each dataset holder will choose the levels of access assigned to their datasets. The data holders will all be asked for the consent of the data owners to be able to open/share them in pseudonymized form. Note that there are some pilots that will combine open datasets with non-open datasets for analysis.

Additionally, IPR and ethical issues will be addressed in pilots, especially for health, and access conditions will be employed to address any identified issues. In addition, the RAI Blockchain will be accessed either by the RAI Registration Service with write permissions or the RAI Finder Service with read permissions.

3.2.5 If an embargo is applied to give time to publish or seek protection of the intellectual property (e.g., patents), specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

At this moment, it is not determined whether any partner would like to seek protection of their IPR. If and when that arises during the RAISE development, the DMP will be updated accordingly.

3.2.6 Will the data be accessible through a free and standardized access protocol?

Upon completion of the RAI registration, an email is sent to the researcher with information and instructions on how to access the produced results. RAISE is also working on delivering basic interfaces and endpoints to facilitate information exchange. Specifically for the repository, it will provide its content via OAI-PMH endpoints.

3.2.7 If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?

Overall, AUTH IT Centre, RAI Central Hub will be installed, will maintain the system for at least 8 years, including after the end of the project active period. More information about access and exploitation during and after the project will be available in the deliverables of T6.2 as the project matures. This information will become available in the DMP in its future versions.

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

RAISE uses encryption and blockchain technology to ensure that the identity of persons accessing the data is controlled. All stakeholders gain access to processed results and source dataset identifier and metadata description, without getting information on the processing script or the methodology of the study. Only if the researcher has opt-in for open algorithm, sharing the processing script becomes available. Otherwise, the processing script remains encrypted with the encryption key of the user that requested the data processing. The data origin information contains all the history of the data processing steps, ensuring traceability mechanisms to increase transparency and proper acknowledgment of the research work behind the results.

3.2.9 Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)?

The origin of the RAISE use cases is relevant to the handling of sensitive and private (closed) data hence it demands a Data Access Committee to be established. The responsibilities and actions to grant data access will be subject to the Data Access Committee strategy which is underway. For example, in the RAISE discovery portal, first access should be granted by the Data access committee of each RAI node, and once granted, requests will be forwarded and managed by each RAI Certified Node or alternatively to Complex Script Execution scheduler, if the requested experiment requires combining algorithm results from multiple RAI nodes. The RAISE portal will also track the status of the different data processing tasks, and save process involved agreements and results (Ethics assessment, IPR agreement, privacy risk assessment, and all signed data access agreements).

3.2.10 Will metadata be made openly available and licenced under a public domain dedication CC0, as per the Grant Agreement? If not, please clarify why. Will metadata contain information to enable the user to access the data?

Source datasets' and results' identifier and metadata description remain accessible to anyone. All metadata descriptions of the RAISE results will be made available in CC0, as per the Grant Agreement.

3.2.11 How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available?

All data and metadata included in the RAISE platform will be maintained for at least 8 years. More information will follow in the sustainability plan of the project, but if the system does not operate after a certain period of time, then all data, metadata and links will be moved and preserved on OpenAIRE's Open Access repository, Zenodo.

3.2.12 Will documentation or reference about any software be needed to access or read the data be included? Will it be possible to include the relevant software (e.g. in open source code)?

Open collaborative research (involving source code sharing) will be supported at RAI Node by Data versioning and Reproducible ML developed subsystem. In addition, RAISE will provide programming language compilers through the RAI Certified Nodes that will support the execution of processing scripts and programming code blocks of the most acknowledged programming languages in the research community. The RAI Programming language compilers will have a simple and intuitive developer experience. The researchers will write, test and debug their scripts locally, using synthetic data, in an environment of their choice. Initially a limited number of programming and scripting languages will be supported, including the open-source Python, NodeJS, R, Java, and proprietary ones such as MatLab and IBM SPSS.

3.3 Making data interoperable

3.3.1 What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data interoperable to allow data exchange and re-use within and across disciplines? Will you follow community-endorsed interoperability best practices? Which ones?

So far, the plan has been to use as a basis the OpenAIRE's data model to enhance it with e.g. FHIR standard for the health pilot.

3.3.2 In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining or extending them?

N/A

3.3.3 Will your data include qualified references[1] to other data (e.g. other data from your project, or datasets from previous research)?

All results of executions, i.e. RAI identifiers, are linked with the encrypted processing script and the dataset of the application.

3.4 Increase data re-use

3.4.1 How will you provide documentation needed to validate data analysis and facilitate data re-use (e.g. readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.)?

To ensure that all functionalities of the RAI Certified Node are transparent to researchers when interfaced through the RAI Central Hub, RAISE will develop the RAI SDK (Software Development Kit) for supported languages and plugins for popular IDEs such as Visual Studio Code. The SDK will simplify tasks such as downloading synthetic data, uploading processing scripts, and uploading datasets by providing tools, documentation, code examples, and guides. This will allow researchers to focus on their research and algorithmic work, with technical details such as API usage kept as transparent as possible. Meanwhile, the plugins will enable the integration of basic functionalities into popular IDEs.

3.4.2 Will your data be made freely available in the public domain to permit the widest re-use possible? Will your data be licensed using standard reuse licenses, in line with the obligations set out in the Grant Agreement?

The public data coming from municipalities and public authorities might already be available in the public domain. But, further reaching consensus on appropriate licence agreements will be sought among the involved parties, to align RAISE exploitation with call's destination-requested open source licensing of project developed assets.

3.4.3 Will the data produced in the project be useable by third parties, in particular after the end of the project?

From the D1.1: "Each partner may transfer ownership of its own results in accordance with the GA's procedures. A third party may be identified in the CA to whom it intends to transfer ownership of its results. In accordance with GA, other partners waive the right to prior notice and their right to object to a transfer to a listed third party. The transferring partner, however, shall inform other partners of such transfer at the time of such transfer, and shall ensure that such transfer will not affect the rights of other partners (see Section 8 of the CA for more information)."

Additional information will become available in the exploitation plan under T6.2 as the project matures. The DMP will be updated accordingly when this information is published.

3.4.4 Will the provenance of the data be thoroughly documented using the appropriate standards?

The data origin information contains a log of all processing steps, ensuring traceability mechanisms to increase transparency and proper acknowledgment of the research work behind the results. For

[1] *A qualified reference is a cross-reference that explains its intent. For example, X is regulator of Y is a much more qualified reference than X is associated with Y, or X see also Y. The goal therefore is to create as many meaningful links as possible between (meta)data resources to enrich the contextual knowledge about the data. (Source: https://www.go-fair.org/fair-principles/i3-metadata-include-qualified-references-metadata/)*

example, in the RAI Registration Service, the outcome of the experiment execution contains the results (together with source dataset and processing algorithm identifier and metadata description), as well as account and timestamp information of processing request for the authorship and timeliness matter. Moreover, the FAIR Catalogue and Portal for data discovery develops the logic to request and track the computation of developed algorithms against a RAI node (or distributed in various nodes) hosted dataset, obtaining an immutable identifier of the dataset, algorithm and results combination. Requests will also be able to target algorithm requests for reusability to the researchers that originally produced them.

3.4.5 Describe all relevant data quality assurance processes.

RAISE has a dedicated Quality Manager (AUTH) that defines the quality plan / policy (project charter) consisting of the requirements related to quality assurance and practical management of the project. The plan can be found in the Deliverable 1 of Work Package 1 (D1.1) For instance, internal processes related to standards of documents, proofreading, validation workflow, etc. Additional processes cover different levels where quality is ensured:

- At the level of the system - Availability
    - The resources for the hosting of this service are provided by the IT Center of AUTH to ensure 24/7 availability even if a RAI Certified Node goes down.

    - By design, blockchain mechanisms do not allow the data of the blocks to be altered retroactively, without the alteration of all subsequent blocks. This allows the users to verify and audit transactions independently.

- At the level of the researcher workflows - Processing Script Quality Analyzer
    - Prior to uploading a processing script on the RAI Certified Node code, quality analysis and malware/vulnerability identification will be performed by the Script Quality Analyzer. Through static and dynamic analysis, the component will act as a Quality Gate prohibiting the uploading of scripts that either: a) are of inferior code and design quality compared to predetermined thresholds or b) contain harmful or vulnerable scripts that can harm the host computer or remote cloud infrastructure or any users relying on the corresponding script. In case of code/design inefficiencies, guidance will be provided to the script developer so as to improve key software characteristics. In case of malware or vulnerability detection, the corresponding threats will be reported and logged along with an assessment of the exploitation probability. In addition, this component will provide estimations on resources required (time to run, CPU and RAM usage, etc) which will be fed to the Processing script manager for better managing the resources.

4 Allocation of resources

4.1 What will the costs be for making data or other research outputs FAIR in your project (e.g., direct and indirect costs related to storage, archiving, re-use, security, etc.)?

RAISE will develop the resources for storage, archiving, re-use, etc by buying a server dedicated for the RAISE outcomes for an 8 year period. Today, the equipment, mainly for the hosting of RAI Central Hub and the distributed RAI Certified Nodes is estimated to be 100K euros, but this is subject to change as things progress. Costing involves:

- High Performance Computing (HPC) infrastructure
    - The internal interconnect and the cluster uplink.
- Cloud (IaaS) infrastructure
    - The hypervisor hosts of the cloud infrastructure.
- Storage services

- Backup services

For covering the computational and storage needs, all servers will include dual power supplies for redundancy and dual or quad uplink interfaces.

Specifically for the proprietary compiler and runtimes (i.e., MATLAB Production Server, RStudio Server, IBM SPSS Statistics Professional) the licensing and the cost will be analysed and considered for the exploitation of the RAISE system.

4.2 How will these be covered? Note that costs related to research data/output management are eligible as part of the Horizon Europe grant (if compliant with the Grant Agreement conditions)

All costs for providing a FAIR-enabled system for privacy preserving data analysis, as explained in the FAIR section of this document, have been incorporated in the overall budget of the project, following the Horizon Europe grant eligibility criteria. The costs related to maintenance and sustainability of RAISE will be detailed in the exploitation plan (D6.2) and reflected in a future version of the DMP.

4.3 Who will be responsible for data management in your project?

All Work Package leaders are responsible for data management, each one in their own area of expertise and RAISE occupation. The Management team, i.e. AUTH, in collaboration with OpenAIRE support coordination that is relevant to Research Data Management (RDM). It's worth noting that RAISE has set up Task Forces for "Make data accessible", "Find datasets (and get access for processing)", "Prepare processing on synthetic data", "Upload (and remotely run) processing script" to more effectively plan and execute work while remaining updated of each other's progress.

4.4 How will long term preservation be ensured? Discuss the necessary resources to accomplish this (costs and potential value, who decides and how, what data will be kept and for how long)?

The current plan foresees for the RAI Central Hub to be installed and operational in AUTH central IT server for at least the next 8 years. The WP5 leader in collaboration with the coordinator and the technical coordinator will investigate the opportunities for extending the duration that the RAI Central Hub will be operational. The goal of the coordination team is to include all data stored in the RAI system and expand the network by inviting collaborators with new RAI nodes in the in an effort to ensure the longevity of the data sharing infrastructure via diversification. The progress of the efforts as well as cost estimations will be included in future updates to the DMP during the course of the project's lifecycle.

5 Data security

5.1 What provisions are or will be in place for data security (including data recovery as well as secure storage/archiving and transfer of sensitive data)?

RAISE uses a combination of techniques to secure data, especially data of special categories.

The RAI Blockchain network will store the successfully processed and registered experiment results in an increasing list of cryptographically linked records (blocks). Each block will contain a cryptographic hash of the previous block, a timestamp, and the information created by the RAI Registration Service (i.e., transaction data): the processing results, the processing encrypted script, the metadata describing the dataset and the account that initiated the data processing. Therefore, the RAI Blockchain Server will be resistant to modification of the data, which makes it a trusted source

for research analysis results. Other means of security to be employed are the hashing algorithm, the security measures of the RAI nodes, security measures for data transfer, back-ups etc.

In addition, to ensure privacy it will deploy a RAI Synthetic Data Generator. This service will generate a downloadable highly realistic and representative, and scaled synthetic version of requested datasets, allowing researchers to locally develop and test new algorithms, without compromising the sensitive information that might be contained within those datasets. Starting from existing open implementations SDV, Datasynthesizer, SynthPop and Generative Adversarial Networks (GANs) will be investigated and most appropriate methods implemented for the privacy-preserving data sharing goal in RAISE (Rankin et. al, 2020, Mikel et. al, 2021). More information can be found on https://github.com/Vicomtech/STDG-evaluation-metrics (code)). State-of-the-art generative deep neural networks and AI mechanisms with in-built privacy mechanisms will compose the synthetic data engine. The engine will also implement an anonymization process in order to provide a higher level of privacy protection as well as compliance with the GDPR. Optionally, OpenAIRE Amnesia open-source tool for anonymizing datasets (https://amnesia.openaire.eu/), which has already been successfully piloted in other projects and demanding environments, such as the DECIDO project for public data, will be integrated for immediate employment of k & k-m anonymity methods of the synthetic data.

5.2 Will the data be safely stored in trusted repositories for long term preservation and curation?

All data will be stored in the RAISE repository as described above.

6 Ethics

6.1 Are there, or could there be, any ethics or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).

The RAISE management has set up an ethical committee (apart from GDPR and technical ethical issues), linking, and enhancing project charter, to share, debate and address ethical issues.

The Internal Ethical Board are the following:

- Mr. Panagiotis Kartsidis, AUTH
- Mrs. Maria Nikolaidou, AUTH
- Dr. Gorka Epelde, VICOM
- Prof. Panos Sarigiannidis, UOWM

The committee is supported by an Independent Ethical advisor, Maria Iakovidou, who has a background in Law and is a member of Hellenic Association of Procedural Law since 2016.

The management together with the ethical committee and advisor will develop the consortium approach around different ethical issues and identify all ethical aspects that could be affected by the project. Special attention will be paid on the research exemption articles of the GDPR in order to allow for a better understanding on how GDPR applies to research studies as well as the ethical issues that arise from the secondary use of data, especially in the health domain.

6.2 Will informed consent for data sharing and long-term preservation be included in questionnaires dealing with personal data?

Co-creation activities are more likely to perform anonymous collection of requirements. Should this not be the case, either swift anonymization or consent for data sharing in pseudonymized form will be sought.

Appendices

Pilots

Joint Research in Health

Pilot description and background information

Pilot in healthcare: transitions of care and home care.

The Biomedical Research and Informatics Living Laboratory for Innovative Advances of New Technologies (BRILLIANT) in community mobility rehabilitation aims to generate evidence-based research to improve rehabilitation for individuals with acquired brain injury (ABI).

BRILLIANT will (1) identify the factors limiting or enhancing mobility in real-world community environments (public spaces, including the mall, home, and outdoors) and understand their complex interplay in individuals of all ages with ABI and (2) customize community environment mobility training by identifying, on a continuous basis, the specific rehabilitation strategies and interventions that patient subgroups benefit from most.

In the framework of VITALISE H2020 project, McGill and the BRILLIANT initiative collaborate with AUTH. The objective of this combines study is primarily to evaluate the feasibility and benefit of collecting multichannel data across Living Labs on the topic of transitional care and to harmonize data processes and collection. During RAISE we will do a common analysis of the data collected in BRILLIANT and AUTH premises to explore initial patterns in the data that demonstrate the potential to predict transition outcomes, such as readmissions.

Joint Research in Environment

Pilot description and background information

The Smart Mobility Solutions pilot is supported by three organizations located in three different countries: Hasselt University (UHASSELT) from Belgium, Hellenic Institute of Transport (CERTH-HIT) from Greece and Vicomtech from Spain. Therefore, we have a University, a Research Center and a Technological Center, all of them collecting and analysing mobility and transport-related data.

In the case of UHASSELT the data is acquired by the Transportation Research Institute (IMOB), an institute of the university, from various national and European projects. UHASSELT-IMOB has large mobility data from various stakeholders, operators and policy makers. Some examples of such acquired data are distraction (in-vehicle or external), fatigue and drowsiness, health concerns (e.g. illness, frailty, cognitive state) and extreme emotions (e.g. anxiety, stress, anger) captured using various in-vehicle sensors. Additionally, data focusing on the mobility footprint on various environmental aspect is also analyzed and incorporated within classical/non-classical transport models.

In the case of CERTH-HIT the data is created by the members of the Thessaloniki Smart Mobility Living Lab, including transport operators, public authorities and infrastructure managers, individual citizens and the equipment of the academia partners. CERTH-HIT hosts and manages the living lab infrastructure, creating added value and services on the top of the collected data to return value to the ecosystem members and to create mobility services and apps for the citizens of Thessaloniki.

In the case of Vicomtech, the participation in multiple research projects has allowed to gain experience working with different types of mobility and transport related data from heterogeneous stakeholders, such as large-scale on-board perception video and sensor data, through public transport operator fleet vehicle data to high level mobility measurements from city and regional authorities.

From the above organizations two main datasets will be added to the RAI system: telematic data from public and private vehicles and video data from various sources. These datasets will be exploited to support modelling and simulation, to train ML prediction engines, to extract mobility patterns, to provide routing services, to detect conflicts and abnormalities and to count pedestrians. More details about the datasets and processing tools/algorithms are provided in the below sections and in the Annex for Thessaloniki.

Contribution and role of partners

Within RAISE CERTH-HIT is the pilot theme leader as well as WP5 leader so it will orchestrate the whole WP and task, while ensuring to the harmonization of the pilot theme tasks together with the other WP5 task leaders (AUTH and ENVE.X). As for the operational aspects all three partners (UHASSELT, CERTH-HIT and VICOMTECH) will have the same role, deploying RAISE nodes and executing the cross-border experiments on the other RAISE nodes.

UHASSELT

Provisioning of various mobility datasets prepared in current and previous projects to the consortium is vital for the success of the mobility pilot. These datasets are from various European cities with focus on citizens engagement, transport management and traffic safety.

Thessaloniki Smart Mobility Living Lab

Most of the mobility eco-system of Thessaloniki is engaged in the living lab initiative since 2012, providing mobility-related data and participating in research and innovation projects to deploy new mobility services and solutions. CERTH-HIT centralizes the data collection and processing and it hosts an open portal where many non-sensitive datasets are openly available.

VICOMTECH

The availability of many data sources used by Vicomtech in previous and current projects is limited to be used in RAISE due to dependencies with external data owners. However, Vicomtech has some own datasets and knowledge of existing open data repositories around the Euskadi region.

Available datasets

The main datasets include Telematics and video data as follows:

- Telematic data from vehicles (taxis, buses, e-scooters, bicycles)
- Video data from various sources (dashcam, external camera in cars, bikes and e-scooters, fixed surveillance cameras, drone cameras)

Additional datasets include travel diaries; environmental sensors; Sdata related to risk-related physiological indicators (driver related background factors and driving environment and traffic complexity indicators); and digital transport network infrastructure (road network, public transport routes, schedules…).

More concretely,

I. UHASSELT

UHASSELT-IMOB is working to setup a framework for the definition, development, testing and validation of a context-aware ‘Safety Tolerance Zone’ for driving, within a smart Driver, Vehicle & Environment Assessment and Monitoring System. Taking into account, on the one hand, driver background factors and real-time risk indicators, and on the other hand, driving task complexity indicators, a continuous real-time assessment is made to monitor and determine if a driver is within acceptable boundaries of safe operation. The core objective is the measurement of risk-related

physiological indicators (e.g., fatigue, distraction, stress, etc.), driver related background factors (age, driving experience, safety attitudes and perceptions, etc.), and driving environment and traffic complexity indicators (e.g., time of day, speed, traffic intensity, presence of vulnerable road users, adverse weather, etc.) to assess driver capacity and task demand in real-time.

Additionally, UHASSELT-IMOB collects environmental data such as air quality close to congested zones using specialized sensors developed in a project and, also, conflict observation data at potentially dangerous intersections using drones.

Thessaloniki Smart Mobility Living Lab

Thessaloniki collects location tracking and tracing data of various vehicle fleets, including taxis, e-scooters and bicycles but also private vehicles and individuals. Within RAISE the trajectories of taxis will be made available. These trajectories are composed of location data every 6-10 seconds including vehicle ID and speed as well as other flags (stop ID, passenger…).

Additionally, Thessaloniki collects video data from micromobility and active mobility vehicle trips as well as (this is an ongoing activity) video from surveillance cameras in pedestrian areas.

II. VICOMTECH

Vicomtech is working in several projects working in route planning and battery consumption estimation for electric vehicles. It already owns a simulated dataset of electric vehicle trips through different routes with timestamped geolocation with battery consumption information and has currently plans for creating a set of real-life trips with private vehicles telematic information.

During the project, Vicomtech plans to record periodic measurements extracted from Open Data of parking occupancy in near real-time and extract traffic monitoring camera images in San Sebastian (Spain).

Proposed tools and algorithms

The main applications include the following:

- Modelling and simulation
- Prediction of different variables (congestion, travel time)
- Mobility patterns extraction
- Routing services
- Conflict/abnormalities detection
- Pedestrian counting
- Safety indicators and safety tolerance zones
- Behavioural intervention

More concretely,

I. UHASSELT

The datasets collected and processed by UHASSELT-IMOB aim to propose a conceptual definition and operational implementation of a safety tolerance zone based on the identified factors and indicators (i.e. context-aware), and operational implementation of safety and driver comfort related interventions to keep the driver in the safety tolerance zone. Interventions will be both immediate (i.e. real-time in-vehicle), and ‘delayed’ i.e. aimed at enhancing the knowledge, attitudes, perceptions and behavioural reaction of drivers with respect to safety-related technologies, situations and behaviours.

Additionally, the datasets collected using environmental sensors allow the citizens to make better travel choices and closely participate in the decision-making process. This data is also evidently used in transport models for several important policy decisions. The drones data allow to safely access the various traffic flow patterns at intersections of interest and make use of AI and other operational algorithms to make informed choices/decisions on a short and long term. The micro-mobility data collected by Thessaloniki Smart Mobility Living Lab can be used for inter-modal conflict observations.

II. Thessaloniki Smart Mobility Living Lab

The datasets collected in Thessaloniki are used to estimate (and predict) traffic conditions in the city but also mobility patterns. Within the taxi case, congestion and travel time as well as mobility patterns of the taxi customers are extracted. The congestion and travel time information is provided as real-time and historical in open sites and apps as well as on the open data portal hosted by CERTH-HIT, while the mobility patterns are provided to the Taxi association in a dedicated service. The prediction module can be used on the top of the parking occupancy data collected by VICOMTECH to estimate parking demand.

Additionally, routing services for pedestrians/micromobility and active mobility, private vehicles and public transport are provided in dedicated sites and apps. The routing case is related to the provision of safe routing to VRUs like pedestrians or e-scooter / bicycle drivers. The air quality data collected by UHASSELT can be used for a cross-border collaboration.

Finally, demand and trip data are used to find the optimum configuration and locations for shared and on-demand mobility infrastructure (stations, stops, number of docks, number of bicycles/e-scooters…). The location module can be used to calculate the number and optimum location of charging stations using the trip traces data collected by VICOMTECH.

III. VICOMTECH

Trip traces with information including vehicle consumption will be used to build and improve battery consumption estimation models for route planning and alert modules for on-route assessment. If the telematic data obtained in Thessaloniki were able to give battery consumption measurements (for e-scooters), this estimation models could be applied or improved. If the UHASSELT-IMOB dataset available from driver behaviours is associated to trip traces, they type of aggressive or defensive driving style could be evaluated (used by the battery estimation model of Vicomtech).

Parking occupancy data will be used to estimate short-term demand prediction and traffic monitoring images could be used as input as well. The short-term demand prediction could be applied to taxi pick-up and delivery data collected in Thessaloniki. Depending on the type of videos collected from micromobility and active mobility vehicle trips and video from surveillance cameras in pedestrian areas, their applicability could be evaluated.

In addition, Vicomtech is developing algorithms to detect and improve punctuality errors in low and high-frequency bus routes and suggest regulation actions. When real time bus telematic data is available, these algorithms could be adapted and tested.

Relevant initiatives and projects

I. UHASSELT

iDREAMS: Horizon Europe - **D**river and **R**oad **E**nvironment **A**ssessment and **M**onitoring **S**ystem aims to set up a platform to define, develop, test and validate a ‘Safety Tolerance Zone’ to prevent drivers from getting too close to the boundaries of unsafe operation by mitigating risks in real-time and after the trip.

iCHANGE: Horizon Europe - is based on the idea that citizens and the civil society have a central role in the definition of environmental protection and climate action and that their direct involvement is essential to drive a true shift and promotion of behavioural changes towards more sustainable patterns.

Search&Rescue: Horizon Europe - Emerging technologies for the Early location of Entrapped victims under Collapsed Structures and Advanced Wearables for risk assessment and First Responders Safety in SAR operations

MIA: Flemish Ministry of Transport - Mobiliteit Innovatief Aanpakken (MIA) to improve on road safety issues in Flanders, Belgium.

II. Thessaloniki Smart Mobility Living Lab

Big Data Europe: Integrating Big Data, software & communities for addressing Europe's societal challenges

Big Data Warehouse for Mobility (BD W4M)

III. VICOMTECH

INPERCEPT - Programa Tecnológico de Automoción Sostenible (PTAS) by Spanish CDTI (Centro para el Desarrollo Tecnológico Industrial)

Movilidad2030 – Misiones by Spanish CDTI (Centro para el Desarrollo Tecnológico Industrial)

Joint Research for Smart Mobility

Pilot description and background information

The Joint research for Environment pilot is supported by two organizations located in two different countries: ENVE.X from Greece (Pilot leader) and EUCENTRE from Italy. Therefore, we have a SME (ENVE-X) and a Research Center (EUCENTRE), both providing and analysing environment-related data.

From the above organizations the datasets ENVE.X and EUCENTRE collected on the frame of the H2020 project ICARUS will be added to the RAI system. These datasets include GPS sensors, personal movement/activity and intensity of activity, environmental temperature and relative humidity. In addition, environmental data from low-cost portable sensor for measuring three fractions of PM (1, 2.5 and 10 μm), and from indoor air quality monitors for air pollutants such as PM2.5 NOX, VOCs, CO and CO2, as well as for toxic organic compounds such as PAHs, dioxins, furans were collected from over 600 participants from 7 European cities for 7 days, in both summer and winter period.

These datasets will be analysed and exploited to support modelling and simulation to derive accurate personal exposure assessment to ambient pollutants at individual level.

More details about the datasets and processing tools/algorithms are provided in the below sections.

Contribution and role of partners

Within RAISE ENVE.X is the pilot theme 2 leader so it will coordinate the whole WP and task, while ensuring to the harmonization of the pilot theme tasks together with the other WP5 task leaders (AUTH and CERTH-HIT). As for the operational aspects both the partners (ENVE.X and EUCENTRE) will have the same role, analysing the available datasets and modelling them.

Available datasets

In the frame of the H2020 project ICARUS a user friendly and easy to use multi-sensor setup - consisted of personal-portable and static sensors – has been established in the seven European cities, providing both at home and personal monitoring for 7 consecutive days, in both summer and winter periods. Data were collected in Athens, Basel, Brno, Ljubljana, Madrid, Milan, and Thessaloniki from the following sensors:

- Garmin Vivosmart for physical activity monitoring
- PPM IoTECH sensors for exposure monitoring
- uHuoo for indoor air quality monitoring
- Silicone wristband for passive exposure monitoring

More in details:

IV. From the Garmin Vivosmart wearable sensors we collected data on intensity of activity, heart rate, stress level, steps, distance, sleep level, metabolic equivalent of an activity/task (MET).

V. From the PPM IoTECH wearable sensors we collected data on Particulate Matter (PM1, PM2.5 and PM10) personal exposure levels, temperature and relative humidity.

VI. From the silicone wearable wristband sensors we collected data on several organic compounds such as PAHs, dioxins, furans.
VII. From the uHuoo static sensors (i.e., placed in the volunteer's households) we collected data on temperature, relative Humidity, PM2.5, Total Organic Compounds (TVOC), CO2, CO, NO2 and air pressure.
VIII. Additional datasets include daily diaries (time activity diaries (TAD)) as well as datasets on the housing conditions.

All the above dataset are available to both ENVE.X and EUCENTRE

Proposed tools and algorithms

The main applications include the following:

- AI models for calibrating and harmonizing sensor data against validated instruments for measuring air pollutants and make them comparable. This includes data gaps filling.
- AI models for predicting exposure levels at individual level based on datasets available collected.
- Machine learning (ABM) models for predicting exposure for population sub-groups having similar features (e.g., vulnerable people, elderly, minors, lower Socio-Economic Status).

ENVE.X and EUCENTRE will be involved and will synergistically collaborate in all the above actions.

Relevant initiatives and projects

ENVE.X is involved in the H2020 project URBANOME (Urban Observatory for Multi-participatory Enhancement of Health and Wellbeing) a 48 months project started in February 2021. The overall objective of URBANOME is to promote urban health, wellbeing and liveability, through systematically integrating health concerns in urban policies and the activities of urban citizens, on the basis of detailed and comprehensive evidence on environmental health determinants, the spatial distribution of these in the city, and the social distribution of their impact among different population groups, accounting for different lifestyles and behaviours. The URBANOME approach will be applied through pilots built by the Urban Living Labs in Aarhus, Athens, Aberdeen, Madrid, Milan, Ljubljana, Stuttgart, Montpellier and Thessaloniki tackling various levels of environmental exposures, age-dependent susceptibility windows, inter-individual variability, gender differentiation of exposure, and socio-economic disparities. In the frame of the project a sensor campaign will be executed in the project pilots for personal exposure estimation coupling wearable sensors and remote sensing techniques.