Grant Agreement No.: 101093069
Call: HORIZON-CL4-2022-DATA-01
Topic: HORIZON-CL4-2022-DATA-01-03
Type of action: HORIZON-RIA

INCODE

Programming Platform for Intelligent Collaborative
Deployments over Heterogeneous Edge-IoT Environments

D1.2 Data Management Plan

Revision: v.1.0

Work package	WP 1
Task	Task T1.2 – Technical, Scientific and Data management [M01-M36]
Due date	30/06/2023
Submission date	30/06/2023
Deliverable lead	UBI
Version	1.0
Authors	Dimitrios Manolopoulos (UBI)
Reviewers	John Avramidis (UNIS), Flavia Maragno, Jean-Baptiste Milon (MARTEL)

Abstract	Data Management Plan (DMP) describes plans for organizing, documenting, storing, and sharing data in line with FAIR principles (Findability, Accessibility, Interoperability, and Reusability). It considers issues such as data protection and confidentiality, data preservation and curation, methodologies and standards applied, and provides measures to ensure reproducibility of research outputs. The Data Management Plan is delivered in M6 and is constantly updated through the project lifespan.
Keywords	Data, DMP, FAIR, GDPR

Document revision history

Version	Date	Description of change	List of contributor(s)
V0.1	10/3/2023	1st version of the template for comments	Dimitris Manolopoulos (UBI)
V0.2	30/3/2023	Abstract, Executive summary & Introduction	Dimitris Manolopoulos (UBI)
V0.3	31/5/2023	Creation of Section s 2-7	Dimitris Manolopoulos (UBI)
V0.4	10/6/2023	1st snapshot of INCODE live DMP repository	All partners
V0.5	10/6/2023	Final draft	Dimitris Manolopoulos (UBI)
V0.6	10/6/2023	Feedback received	John Avramidis (UNIS), Flavia Maragno, Jean-Baptiste Milon (MARTEL)
V1.0	26/6/2023	Final revisions	Dimitris Manolopoulos (UBI)

Disclaimer

Co-funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the other granting authorities. Neither the European Union nor the granting authorities can be held responsible for them.

Copyright notice

Project co-funded by the European Commission in the Horizon Europe Programme		
Nature of the deliverable:	DMP	
Dissemination Level		
PU	*Public, fully open, e.g., web*	**x**
SEN	*Sensitive, limited under the conditions of the Grant Agreement*	
Classified R-UE/ EU-R	*EU RESTRICTED under the Commission Decision No2015/ 444*	
Classified C-UE/ EU-C	*EU CONFIDENTIAL under the Commission Decision No2015/ 444*	
Classified S-UE/ EU-S	*EU SECRET under the Commission Decision No2015/ 444*	

* *DMP: Data management plan.*

Executive summary

The Data Management Plan (DMP) INCODE serves as a comprehensive guide for the effective management and utilization of data throughout the project's lifecycle. This document outlines our strategy to ensure data integrity, accessibility, and security, while promoting data sharing and re-use to maximize the impact of our research outcomes. The primary objective of our DMP is to establish robust procedures for data collection, organization, storage, and sharing. By adhering to best practices and industry standards, we aim to create a reliable and well-structured data repository that facilitates collaboration among project partners and enables the broader research community to benefit from our findings. In this first version, we address key aspects of data management, including data acquisition, metadata documentation, storage infrastructure, data sharing protocols, and data preservation. This deliverable also includes ethical considerations and data protection, ensuring compliance with relevant regulations and safeguarding sensitive information. Last, INCODE consortium within this document foresees measures to maximize the value of our research data by encouraging open data practices and fostering collaborations with stakeholders beyond the project's scope. To reflect the project advancements, INCODE DMP will be regularly reviewed and updated to adapt to evolving data management practices and emerging technologies.

Co-funded by
the European Union

Table of contents

Co-funded by
the European Union

List of figures

Co-funded by
the European Union

List of tables

Co-funded by
the European Union

Abbreviations

API	Application Programming Interface
DMP	Data Management Plan
DOI	Digital Object Identifier
EC	European Commission
FAIR	Findability, Accessibility, Interoperability, Re-usability
GDPR	General Data Protection Regulation
GNU	GNU's Not Unix
HE	Horizon Europe
ICT	Information Communication Technologies
IoT	Internet Of Things
IPR	Intellectual Property Rights
URI	Uniform Resource Identifier
URL	Uniform Resource Locator

1 Introduction

For INCODE the Data Management Plan (DMP) serves as a crucial component. As we embark on this journey to develop innovative solutions in the field of IoT, edge, and cloud continuum, it is essential to establish a robust framework for managing the data generated throughout the project's lifecycle. Hence, the DMP outlines our approach to data collection, organization, storage, sharing, preservation, and exploitation, ensuring that our project adheres to best practices and meets the requirements set forth by Horizon Europe funding framework and relevant regulatory bodies from EC.

1.1 Scope

This initial version of a living DMP sets the stage for effective data management, recognizing the significance of data as a valuable asset in our research. It highlights the need to establish a clear roadmap for data handling, considering the unique challenges and opportunities presented by the INCODE scientific and technical landscape. By developing a comprehensive DMP, we aim to foster transparency, collaboration, and data-driven innovation within our project.

The primary objectives of this DMP are to ensure the integrity, accessibility, and usability of our research data, promote open and responsible data sharing, facilitate data re-use and validation, and address data security and ethical considerations. By implementing sound data management practices, we strive to maximize the impact and value of our research outcomes, enhance collaboration with stakeholders, and contribute to the wider scientific community.

This DMP will guide our project team and collaborators in effectively managing data throughout its lifecycle, from data generation and collection to data preservation and potential re-use beyond the project's duration. It will serve as a reference document that outlines the procedures, methodologies, tools, and resources required to achieve our data management goals. Furthermore, the DMP will be reviewed and updated in a semester basis to adapt to evolving project needs and emerging data management practices, ensuring its relevance and effectiveness throughout the project's lifecycle.

1.2 Structure

Since INCODE DMP follows the related template provided by the European Commission the deliverable addresses the following issues:

- Data Summary
- FAIR data (Findable, Accessible, Interoperable, Re-usable)
- Allocation of resources
- Data security
- Ethical aspects
- Other issues
- Conclusions

Each of the previously defined aspect has its own set of questions that must be addressed. The proposed template states that it is not required to provide detailed answers to all the questions of the DMP that needs to be submitted by month 6 of the project, subject -also- to potential future updates.

2 Data summary

As described in the Fair Data Management Guidelines for Horizon Europe, a Data Management Plan is a key element as it promotes effective data handling, facilitates data sharing and collaboration, ensures data security and ethics, enables long-term data retention, and demonstrates compliance with funding body requirements. It also helps researchers maximize the value of their data, enhance research integrity, and contribute to the wider scientific community.

For this reason, in this section we will first define the type of data and other artefacts that will be created and collected as part of the project during the lifetime of the INCODE project. The first collection of data and artifacts to be collected/created are listed in the Appendix of this deliverable. As the project progresses, this list will be updated on a 6-month basis (adding or removing items) in relation to the various stages and developments of the project.

2.1 Existing data reused

Data will be collected throughout the project. Regarding the use of retrospective data sets collected and used to support the technical developments and testing activities that will take place in the four Application Areas, specific protocols will be documented and approved by the ethics committees and/or competent authorities of INCODE and its partners in order to align their use with the terms and conditions that have been agreed upon for the possible data sets. It is worth mentioning that Application Areas partners already stated that existing datasets will be updated an enhanced. Likewise, it is sought to avoid sharing data with non-EU project partners (unless otherwise specified and not in the case of INCODE's third country partners participating in the project) and to make corresponding decisions regarding technical implementation accordingly.

2.2 Types and formats

Table 1 INCODE Data types

ARTEFACT TYPE	EXPLANATION	WP#	FORMAT
Research Items	**Deliverables**: The project will produce several deliverables uploaded on the project website. These deliverables are either public (PU) or confidential (CO). **Scientific publications**: INCODE partners will produce scientific publications that will be made publicly available for the wider audience. **Other dissemination and communication publications**: In the scope of WP6, other publications will be produced like website pages, promotional materials, press releases, website news, and blogs	WP1-WP6	MS Word (.doc/.docx), Adobe PDF (.pdf), MS Excel (.xls, .xlsx), Comma - separated values (.csv), Hypertext Mark - up Language(.html), JPEG (.jpeg, .jpg), GIF (.gif), PNG (.png), MPEG-4 (.mp4)

Software	Code, APIs, microservices, libraries, dashboard	WP2-WP5	More common formats for software-related data include code files (e.g., Java, Python, C++, etc.), configuration files (e.g., YAML, JSON, XML, etc.), database files (e.g., SQL, NoSQL, etc.), log files (e.g., text, CSV, JSON, etc.), and various types of binary files (e.g., executables, libraries, etc.). The specific format may also depend on the tool and its purpose, as well as any relevant standards or conventions that apply to the particular domain or industry.
Dataset	Datasets for the management and orchestration, datasets from the creation and collection of events, datasets from the deployment and evaluation of the AAs, models and meta models, Policies, Questionnaires. In the context of WP6 basic user datasets will be collected within the INCODE website and Newsletter. The INCODE website collects the following data: • User account data: i.e. account of users authorised to publish content on the website. These are in general users part of MARTEL and may include: • Name and relevant titles • Email address • Contact data: i.e. data provided by website visitors filling in contact forms or sending email to "info@incode.eu" to request information to the INCODE consortium. The INCODE Newsletter collects the following data : • E-mail address of visitors registering to the newsletter.	WP2-WP6	May vary depending on the type of data, but they could include structured data in databases or spreadsheets, unstructured data in log files or text documents, and multimedia data such as images or videos. Indicative formats expected include .xls, .csv, .txt, .docx, .pdf, json, Binary, ASCII, MQTT.

	• During the registration for the newsletter, we also store the IP address of the computer system assigned by the Internet Service Provider (ISP) and used by the data subject at the time of the registration, as well as the date and time of the registration		

2.3 Purpose of the data generation

The purpose of data generation in INCODE is to collect and generate data from various sources within the project's ecosystem. This data serves as the foundation for achieving the project's objectives. The relationship between data generation and project objectives lies in the following:

Data-driven Innovation: INCODE aims to leverage data generated from IoT devices, edge computing systems, and cloud infrastructure to drive innovation. By collecting and analyzing real-time data, the project can uncover insights, patterns, and trends that can lead to the development of the envisioned programming tools, services, or applications.

System Optimization: INCODE is also focused on optimizing the performance and efficiency of its platform across the continuum. Data generation plays a crucial role in understanding system behavior, identifying bottlenecks, and optimizing resource allocation. By analyzing data, the project can fine-tune system parameters, enhance scalability, and improve overall system performance.

Decision Making and Intelligence: INCODE aims to develop a platform that allow end-users make informed decisions based on data analysis. Data generation provides the necessary input for training machine learning algorithms or implementing data-driven decision-making processes. By analyzing real-time data, the project can enable automated decision-making, predictive analytics, or intelligent resource allocation.

Validation and Evaluation: INCODE objectives include validating and evaluating the effectiveness and feasibility of the proposed IoT-to-Edge-to-Cloud architectures, protocols, and tools. Data generation will allow our partners to gather experimental evidence and performance metrics to assess the project's proposed solutions. By collecting data and analyzing its impact, the project can validate its hypotheses and evaluate the success of its approaches.

Dissemination, Communication, and exploitation: Data generation is important for the dissemination, communication, and exploitation objectives of INCODE. It can provide tangible evidence of the project's outcomes and achievements as it will allow the partners to showcase the functionality, performance, and capabilities of the INCODE solution. By generating relevant data, the project can effectively communicate its findings, results, and advancements to stakeholders. At the same time, INCODE generated data will promote collaboration and exploitation opportunities for the project. By providing access to relevant data sets, the project can engage with external stakeholders, industry partners, and potential users to explore collaborative opportunities, further research, and commercialization possibilities. The availability of data will enhance the project's attractiveness and facilitate the transfer of knowledge and technology to other organizations or industries.

2.4 Expected size of the data

It is expected that INCODE activities will result in research data sets (i.e., results stemming from the development and testing of technical components, services produced and tested under the Application Areas, etc.), publications, and dissemination materials. Due to the project size, scope of work, and complexity, the expected size is currently estimated at around 1 TB.

2.5 Data origin and provenance

For INCODE is crucial to document and establish the origin/provenance of data since this documentation helps ensure transparency, reproducibility, and compliance with ethical and legal requirements. It also enables proper attribution, acknowledgment of data sources, and adherence to data sharing and reuse policies. Bellow it can be found the latest snapshot of data sources concerning their origin/provenance used in INCODE:

a. **Generated Data** created or collected specifically for the research project. It may include data obtained through experiments, surveys, simulations, or other data-generating activities conducted by the project team. The origin of generated data can be traced back to INCODE itself.

b. **Secondary Data Sources** which involve reusing existing data from secondary sources. These sources can include publicly available datasets, research repositories, government databases, or other relevant data sources. It is important to identify the original sources of the data and comply with any licensing or copyright requirements associated with their reuse.

Even if INCODE has not recorded so far other types of data origin, the following categories cannot be excluded in the future as the project evolves:

a. **Primary Data Sources** which involve collecting data from primary sources outside of the project. For example, data can be collected from individuals, organizations, or devices through partnerships, collaborations, or direct data collection efforts. The provenance of primary data sources should be documented to establish their reliability and ensure ethical considerations.

b. **Collaborative Data Sources** that come because of collaboration with other institutions, organizations, or researchers. In such cases, data may be shared or exchanged between project partners. The provenance of collaborative data sources should be clearly documented to acknowledge the contributions and establish data ownership and usage rights.

c. **Open Data Sources** by leveraging open data sources, such as openly available datasets, public APIs, or data repositories. Open data sources typically have defined provenance, licensing, and usage terms that need to be followed when incorporating the data into the research project.

2.6 Target audience

Mapping audiences that have an interest in INCODE results will help the project identify the specific groups or individuals who have an interest or influence over the project's data management activities. By understanding their needs, expectations, and concerns can help the project to further tailor its data management plan to address their requirements. This ensures that the project's data is managed and shared in a manner that aligns with their expectations. It also helps in identifying potential barriers or challenges in data sharing and utilization, allowing the project to proactively address them. Ultimately, mapping target audiences within the data management plan helps foster transparency, trust, and collaboration, enhancing this way the project's impact, sustainability, and potential for successful adoption and exploitation of its innovations. So far, we have identified the following groups:

Industry Partners: Companies operating in the network/communication/IoT sector can leverage the project's data to gain insights into emerging trends, technologies, and best practices. The data can inform their product development, innovation strategies, and decision-making processes. So far INCODE identified 3 relevant sub-groups: (i) Operators (i.e., infrastructure owner and service providers), (ii) Vertical end user (i.e., Industry sector), (iii) system integrators.

Policy Makers and Regulators: Government entities and regulatory bodies can utilize the research project's data to inform policy development, standards-setting, and regulatory frameworks related to network, communication, and IoT technologies. The data can help shape policies that promote efficient and secure communication networks, data privacy, and interoperability.

Academic and Research Community: Researchers and scholars working in the field of network, communication, and IoT can benefit from the project's data for further analysis, validation of findings, and building upon the existing knowledge base. The data can contribute to scientific advancements, foster collaboration, and inspire new research directions.

Non-Profit Organizations and Advocacy Groups: Non-profit organizations and advocacy groups focused on areas such as digital rights, privacy, and internet access can utilize the project's data to support their initiatives. The data can help raise awareness, inform public discourse, and advocate for policies and practices that promote a safe and inclusive digital environment.

General Public: The findings and insights derived from the project's data can be communicated to the public through various channels such as reports, publications, or educational materials. This can enhance public understanding of network, communication, and IoT technologies, their impact on daily life, and potential benefits or risks associated with their usage.

Co-funded by
the European Union

3 FAIR data

3.1 Making Data Findable

INCODE will not only provide FAIR data where applicable (data should be "as open as possible and as close as necessary") but will mostly support the community in the uptake of research data sharing and practices, in alignment with FAIR principles.

The following provisions for making INCODE data and metadata findable and promote their reuse will be followed:

- Collected data will be discoverable with a unique and persistent URI and be available at the dedicated portfolio/catalogue system URL. Other aspects of the persistent identifiers will be implemented by periodical snapshots of the database that will contains the data.
- Created metadata for general research data will follow the Dublin Core and DataCite metadata schema.
- Keywords will be used in each dataset published so as to increase the possibilities of reuse.
- A clear versioning will be provided with unique and persistent URI.

3.2 Making data Accessible

By default, INCODE will openly provide data produced following the principle "as open as possible, as closed as necessary", to comply with ethical or security requirements and avoid related conflicting issues. The following actions will be implemented on three levels:

Repository

- INCODE will ensure data accessibility through its deposition in a trusted repository. The choice of repository will depend on factors such as the type of data, the intended user community, and any relevant disciplinary standards or practices. As a requirement the chosen repository should provide a persistent and citable identifier (such as a DOI) and should adhere to relevant data management best practices, such as providing long-term preservation and metadata standards. As the most commonly used by research projects repositories adhering to these requirements, the following are most likely to be used to:
    - **GitHub**: GitHub is a web-based hosting service for version control using git. It provides a platform for open-source software development and collaborative software projects.
    - **GitLab**: like GitHub, GitLab is also a web-based platform used for version control and collaborative software development. It provides a hosting service for Git repositories and supports continuous integration/continuous deployment (CI/CD) pipelines.
    - **Zenodo**: a general-purpose open-access repository that allows researchers to deposit data, software, and other research outputs. It is operated by CERN and offers a variety of features including persistent identifiers (DOI), versioning, and integration with other services such as GitHub.
- The project may also consider other means of data accessibility, such as making data available through a project website, data portal, or through integration with existing data infrastructures. The project should ensure that any data accessibility options selected are sustainable and meet the needs of both the data producers and users.

Data

- The specific details of data accessibility are dependent on the individual technological blocks of INCODE as well as on the nature of the data produced across the four Application Areas. However, as a general principle, the project aims to make data as widely accessible as possible while, also protecting any sensitive or confidential information.
- If there are restrictions on use of the data, access will be provided according to the specified terms and conditions set by the data owner or the repository where the data is deposited.
- INCODE will work with the repository and data owner to ensure that access is provided in accordance with any restrictions. Any such restrictions will be clearly communicated in the associated metadata and documentation to ensure that users are aware of them before accessing the data.
- All three platforms mentioned above (GitHub, GitLab, Zenodo) support Git repositories, which are commonly used for version control and collaboration in software development projects. Users can access the data through Git commands or through web-based interfaces provided by each platform. Zenodo is also specifically designed for research outputs and provides features such as persistent identifiers and versioning to ensure the data is discoverable and citable.
- Access will be given using user authentication and authorization, handling users' verification and access levels. This may require from users to agree to a license or providing access only to authorized individuals or organizations.
- A data access committee consisting of UNIS, UBI, and MARTEL (Project Coordinator, Technical Coordinator, and Ethics Manager) is assembled to take care of data access issues and ensure the compliance with GDPR issues associated.

Metadata

- Metadata that will be created in the project includes information about the datasets used and produced, such as:
  - Dataset title and description
  - Creator and contributor information
  - Date of creation and modification
  - Spatial and temporal coverage
  - Data format and size
  - Data access and use conditions
  - Persistent identifiers (e.g., Digital Object Identifiers) if applicable
  - Provenance information (e.g., data sources, processing steps, quality control)
- All of these metadata will be openly available and licensed under a public domain dedication CC0, as per the Grant Agreement
- Data is intended to be available for as long as it is allowed by project resources and infrastructure. This will be reviewed as the project progresses.
- The consortium partners will take the necessary steps to include documentation about the software needed to access the data. The documentation will include information on any dependencies, installation and usage instructions, as well as troubleshooting tips.

Research material provisions

With regards to the dissemination of the scientific results, and in alignment with the EC Guidelines on Open Access to Scientific Publications and Research Data in Horizon Europe, INCODE will follow a combination of Gold and Green Open Access strategy to its scientific publications, with a potential embargo period for peer-reviewed publications that will be agreed during the first months of project execution. Gold Access will be encouraged for high-impact journal publications while the self-archiving, Green Access will be granted for the rest of the publications.

INCODE partners defined a simple, deterministic process that decides if a result must be published or not. The term result is used for all kind of artefacts generated during INCODE like white papers, scientific publications, and anonymous usage data. By following this process, each result is either classified public or non-public. Public means that the result must be published under the open access policy. Non-public means that it must not be published. For each result generated or collected during INCODE runtime, the following procedure will be followed to classify it.

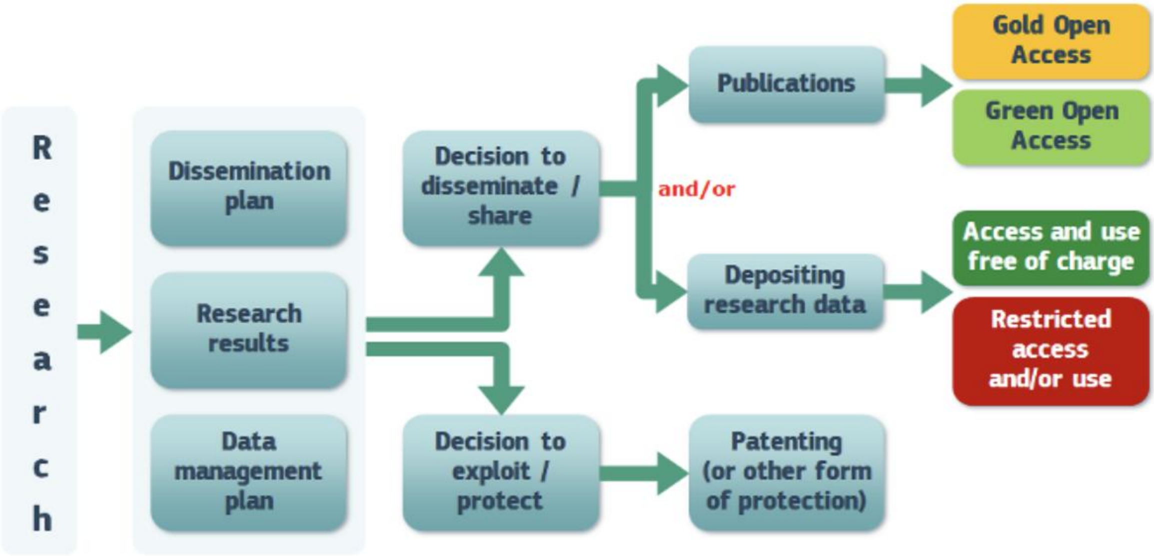

Research
Dissemination plan
Research results
Data management plan
Decision to disseminate / share
Decision to exploit / protect
and/or
Publications
Depositing research data
Patenting (or other form of protection)
Gold Open Access
Green Open Access
Access and use free of charge
Restricted access and/or use

Figure 1 Open access to scientific publication and research data in the wider context of dissemination and exploitation

3.3 Making data interoperable

INCODE recognizes the value of interoperability of its research data, as it is known to accelerate scientific discovery, promote collaboration and enable re-use and validation of research results. Interoperability fosters a more open and connected research ecosystem, where data can flow seamlessly across disciplines, research organizations, and geographic boundaries, driving innovation and unlocking exciting possibilities for scientific advancement. To address this requirement, INCODE will:

- Adhere to relevant standards for data formats to ensure interoperability between different datasets and facilitate data exchange and reuse between researchers, institutions, organizations, etc. We expect this to evolve dynamically during the project lifetime to ensure an ontology alignment within the Open Science Cloud (EOSC).
- Use Open-Source software applications wherever possible, to ensure compatibility and facilitate recombination with different datasets from different origins. In addition, efforts will be made to ensure that the data produced by the project can be integrated with existing datasets and infrastructure, particularly those related to the telecommunications industry.
- Make the data produced in the project interoperable by following relevant vocabularies, standards, and methodologies such as:
    - **OpenAPI**: The OpenAPI specification is a widely used standard for describing RESTful APIs. Using OpenAPI, we can provide a machine-readable description of the API, which makes it easier for other researchers and organizations to integrate with our data.

  - **Resource Description Framework (RDF)**: RDF is a standard model for data interchange on the web. It provides a framework for representing data and metadata in a machine-readable format, enabling data interoperability between different systems.
  - **Dublin Core**: The Dublin Core Metadata Initiative provides a set of metadata standards for describing resources on the web. Using Dublin Core, we can provide a standardized set of metadata for our data, making it easier for other researchers to find and use our data.
  - **Schema.org**: Schema.org provides a set of structured data vocabularies for use in web pages. By using Schema.org, we can provide structured data for our web pages, making it easier for search engines to understand and interpret our data.
  - **Linked Data**: This concerns a set of best practices for publishing and connecting structured data on the web. By following Linked Data principles, we can make our data more discoverable and easily integrable with other datasets.
- Provide mappings to more commonly used ontologies if uncommon or project specific ontologies or vocabularies are used.
- Add appropriate reference if existing data sets from previous research are used in the project.

1.1 Increase data re-use (through clarifying licenses)

INCODE is committed to increasing data re-use as it will enable the value and impact of collected data to be maximized. When data is re-used, it can be leveraged for multiple purposes, such as validating research findings, conducting further analysis, or developing new applications that will drive IoT-to-edge-to-Cloud continuity. Our actions in this area will promote efficiency, enhance collaboration, and encourage innovation within the community, avoiding unnecessary duplication of data collection efforts and enabling researchers to build on existing knowledge and resources. The following steps are foreseen to achieve this goal:

- Partners are committed to validate the produced data and facilitate data re-use, by implementing a series of steps. Firstly, each data producer will provide a thorough documentation of research methodology used, including readme files containing data collection methods, sampling techniques, and any preprocessing or cleaning steps applied to ensure transparency and replicability of the data analysis process. Secondly, INCODE's developers' will provide extensive documentation of any code or scripts developed for data analysis to explain its functionality and include information on dependencies and libraries used to validate the analysis and potentially reuse or modify the code. The documentation of the data processing steps, outlining the transformations, aggregations, calculations, or statistical methods applied is also an important step to achieve data re-use to enable third users to understand how the analyzed data was derived. Last but not least, INCODE partners will be asked to record the parameters and configurations used in the analysis, as well as clearly present the results and interpretations of the data analysis so as make easier the result validation and the reusability of the findings.
- Data will be available by using one of Creative Commons license options. The license for each dataset will be the one providing the widest re-use possible. For software, possibly Apache or GNU. More information will be specified in a following DMP version.
- Data produced or used in the project will be useable by third parties after the end of the project after a security and ethical parameters were assessed. For example, we expect data stemming out of critical infrastructure to have such restrictions. This will also be made clear as soon as there is a full and detailed view of the datasets achieved.

- Assure the provenance of data generated by INCODE by capturing and recording information about the origin, ownership, processing, and transformations applied to the data throughout its lifecycle to promote openness and facilitate future research and collaboration. This will be done by following the steps bellow: (i) select a provenance standard that aligns with project's and partners' needs and requirements like PROV-O ontology, W3C PROV standard, or Dublin Core Metadata Initiative (DCMI) metadata terms, (ii) identify the key elements to be captured in the provenance documentation such as data source, data collection method, data transformations, data processing steps, software tools used, data contributors, and timestamps for data events based on the provenance standard, (iii) capture and record the relevant provenance metadata by documenting the information in a structured format or by utilizing software tools that automatically capture provenance information during data processing, (iv) associate the captured provenance metadata with the corresponding data artifacts by linking the provenance information to the dataset files and creating metadata files alongside the data. Again, this will be correlated to the chosen provenance standard by each partner and the data file format being used, (v) ensure that the documented provenance information is preserved and maintained throughout the research project and beyond in the designated secure and accessible repository, alongside the associated and regularly updated data, (vi) share the relevant provenance information to enable others to understand the data's history and processing steps.
- Effort will be put in providing quality data ensuring the accuracy, consistency, completeness, and reliability of the collected data. Here are some key processes that INCODE partners will implement: (i) data collection planning by specifying data requirements, designing data collection instruments, and establishing protocols for data gathering, (ii) data validation and verification by checking the accuracy and integrity of collected data, (iii) data cleaning and preprocessing by identifying and correcting errors, inconsistencies, outliers, and missing values in the collected data, (iv) creation of metadata that describes the data collection process, variables, and their definitions, (v) quality control checks by reviewing the data for anomalies, conducting data audits, and performing cross-checks with other sources or experts, (vi) put effort on ensuring data security and privacy by anonymizing or de-identifying personal information by implementing access controls, complying with data protection regulations and overall safeguarding against unauthorized use or disclosure, and finally (vii) reporting activities by documenting any modifications made to the data, data cleaning procedures applied, and explanations of data anomalies or inconsistencies.

4 Allocation of resources

INCODE also includes provisions for the appropriate allocation of resources related to data management, as it will enable their effective and efficient use. With proper data management, including storage, organization and access, valuable information can be easily retrieved and used when needed by both internal and external stakeholders. This leads to improved decision making, optimized workflows and increased productivity. In addition, proper data management will enable INCODE to ensure data security, privacy and regulatory compliance, protecting sensitive information and maintaining trust among INCODE partners. Therefore, INCODE foresees the following actions for the proper allocation of resources:

- The cost of creating FAIR data can vary depending on a number of factors, including the size and complexity of the data, the level of documentation and metadata required, the type of storage used to deposit and maintain data, and any costs incurred related to open access publishing. INCODE provides the following costs to consider when implementing FAIR data in a project include:
  - Data management and curation costs associated with data storage, backups, and security, as well as costs for developing and maintaining data documentation and metadata
  - Repository costs associated to data deposition, preservation, retrieval, or access
  - Software and tools to support data analysis, visualization and dissemination can be used depending on the complexity of the data and the analysis methods used. In principle, INCODE will prioritize the use of open source over proprietary
  - Staff time is expected to be used to develop and implement data management plans, create documentation and metadata, and prepare data for sharing and publication, as these are time-consuming tasks and may require additional staff time or resources
- Data storage and maintenance costs are not going to require additional funding once the project is finished. Regarding the value of the data, it is important to consider that the topics covered by the project respond to a current need of the needs of the continuous IoT-to-Edge-to-Cloud. Therefore, the data that will emerge from this project will have an immediate impact in the coming years but may not be significant as the challenges are addressed or superseded by other priorities.
- UBI as the T1.2 leader and scientific coordinator assisted by the WP leaders will be responsible for updating this document and will develop a strategy to encourage: (i) the identification of the most appropriate methods of data sharing and preservation, (ii) efficient use of data by ensuring clear rules regarding its accessibility (iii) the quality of data stored and (iv) storage in a secure and user-friendly interface.
- Concerning long term preservation INCODE will pursue cost-effective preservation strategies that can balance the long-term storage and accessibility with available resources in hand. This may involve assessing different storage options, such as cloud-based solutions or institutional repositories, and evaluating their costs and scalability. Collaborating with existing partners' data infrastructures or initiatives like Zenodo, Figshare, Open Science Framework will also help reduce costs and ensure sustainable data preservation. Of course, the final decision will be made upon the review of each repository specific terms of service, storage limits, and any potential charges for additional services. Additionally, INCODE is committed to abide with the requirements and recommendations from your EC to ensure compliance with their data sharing policies.

5 Data Security

INCODE data already applies technological and organizational measures to secure processing of personal data against publishing to unauthorized persons, processing in violation of the law and change, loss, damage or destruction by adopting the following measures:

- Data collected and stored within INCODE cloud workspace (SharePoint) are encrypted using AES 256-bit encryption in geographically diverse areas. HTTPS protocol which runs on top of encrypted sockets (SSL/TLS) on the transport layer of the network stack (TCP/IP) is used for secure communication between endpoints as a standard.
- The password for each account exists on the platform only in a coded form.
- The platform offers the possibility to make data available in a read-only or downloadable format, hindering the access to information by unauthorized users.
- Complete back-ups are done every week. In addition, every time a modification is done an older version is saved.
- In the occasion of a catastrophic event that implies the partial or complete deletion of the data sets, the data from the most recent back up will be automatically restored. The last back-up won't be older than 60 minutes.
- In the occasion of accidental deletion or modification only the most recent document will be restored, so as data can be easily recovered.
- Only INCODE SharePoint administrators have the rights to delete or modify the information included in the datasets.
- Data will be moved to public storage such as Zenodo for long-term retention. As already mentioned, INCODE will base its repository selection on security features as well. For example, the Zenodo datasets are stored in CERN's EOS service on an 18 petabyte disk array. Each file copy has two copies located on different disk servers. For each file they store two independent MD5 checksums. A checksum is stored by Invenio, which is used to detect changes to files made outside of Invenio. The other checksum stored by EOS is used to automatically detect and recover from file corruption on disks.

Co-funded by
the European Union

6 Ethics

INCODE Consortium is aware of ethical, privacy, copyright and data protection issues that may arise from the activities to be carried out under the project. As a result, provisions related to ethical and legal compliance issues, such as consent to retain and share data, protecting individuals' identities and how personal data is handled to ensure its safe storage and transfer; as well as copyright and intellectual property rights (IPR) issues also must be made. These concerns are discussed in detail in the context of Task 1.3 where we define how research will be executed in the project regarding the ethics issues during the implementation of the project including, but not limited to confidentiality, integrity, validity, objectivity, accuracy, transparency, trustworthiness, authenticity, respect for autonomy, reciprocity, and equity. The task will also examine legal and regulatory issues related to project implementation and potential barriers that arise from them. In collaboration with the project partners and the ethics committee INCODE is committed to closely monitor and consult the consortium with regards to any activity involving ethics issues.

At this point in time, the consortium deems appropriate to refer on the legal framework that governs the activities of the project and directly affect INCODE's data management approach. These include, but are not limited to, the following:

(i) **General Data Protection Regulation** which is a comprehensive data protection regulation that sets out rules for the processing of personal data within the European Union. It applies to all research projects handling personal data and emphasizes principles such as data minimization, purpose limitation, data security, and individual rights.

(ii) **Data Protection Directive for the Police and Criminal Justice Sector directive** that addresses the processing of personal data for law enforcement purposes while it also establishes rules for the lawful and ethical handling of personal data in the context of research projects involving law enforcement agencies.

(iii) **Directive on Open Data and the Reuse of Public Sector Information** that promotes the openness and reusability of public sector information, including research data funded by public bodies while it also encourages the sharing and accessibility of research data, enabling its reuse and maximizing its societal and economic benefits.

(iv) **Directive on Copyright in the Digital Single Market** which addresses copyright rules and intellectual property rights within the European Unio also including provisions related to data mining and text and data analytics, allowing for the lawful use and extraction of data for research purposes.

(v) **European Code of Conduct for Research Integrity** which, while not a legislation or directive, provides guidelines and principles for researchers and institutions to ensure ethical research practices. It covers various aspects of research integrity, including data management, publication ethics, and conduction of responsibe research.

(vi) **Horizon Europe Rules for Participation and the Horizon Europe Model Grant Agreement** which outline the general principles and rules that govern the funding and implementation of research and innovation projects. While the specific ethics aspects are not explicitly listed in these documents, the include references concerning ethical considerations and highlight the importance of addressing ethical aspects in research projects like ethical compliance, review, informed consent, data protection, and dual use research.

In order to ensure that all ethical aspects are considered and that the INCODE project is compliant with all legal requirements and ethical issues, a general approach has been discussed. This involves an ad hoc monitoring process of the project development the "Socio-legal Approach" which is a risk approach to privacy and data protection issues that takes into account not only legal considerations but also social factors and societal norms. This approach recognizes that the protection of personal data goes beyond mere compliance with legal requirements and requires a broader understanding of the social and ethical implications of data processing. In the context of the General Data Protection Regulation (GDPR), the Socio-legal Approach emphasizes the need to assess the risks associated with the processing of personal data from both a legal and social perspective. It involves considering not only the legal obligations imposed by the GDPR but also the potential impact on individuals' privacy, autonomy, and societal values. In a brief this approach includes the following:

(i) **Legal Compliance** which starts with the GDPR's legal requirements, such as obtaining consent, providing transparency, implementing data protection safeguards, and respecting individuals' rights.

(ii) **Risk Assessment** by conducting a comprehensive risk assessment to identify potential risks to privacy and data protection. This includes evaluating the likelihood and potential harm of data breaches, unauthorized access, data misuse, or any other privacy-related risks.

(iii) **Social Implications** that consider the social implications of data processing activities. This involves considering the broader societal impact, ethical implications, public perception, and potential discrimination or stigmatization resulting from data processing.

(iv) **Stakeholder Engagement** which promotes the involvement of stakeholders, such as data subjects, privacy experts, civil society organizations, and regulatory authorities, in the decision-making process to ensure that different perspectives are considered and facilitate a more holistic approach to privacy and data protection.

(v) **Continuous Evaluation and Improvement** to emphasize on the need for ongoing monitoring, evaluation, and improvement of data processing practices to address emerging risks and changing societal expectations.

The ethical requirements set by the EC, during the ethics review process before the signing of the GA, have also been taken into account by the consortium, while the general guidelines for the project and the consortium have been provided for all aspects, covering data protection, privacy issues, safety of research participants. For example, the following provisions have been made:

- **Confidentiality**: INCODE partners must retain any data, documents, or other material as confidential during the implementation of the project. Further details on confidentiality can be found in respective article of the Grant Agreement along with the obligation to protect results.
- **Personal Data:** If personal data will be collected within this project, will only be stored, analyzed and used anonymously. The individuals will be informed about the use of the information collected by them and will have to agree to the data collection while providing their approval in the form of written consent. The identity of any individual interviewed or in any other way engaged in the project (e.g., by email, correspondence, newsletter) will be protected by this anonymization of the data.
- **Intellectual Property Rights**: INCODE ensures that data access and sharing activities will be rigorously implemented in compliance with the privacy and data collection rules and regulations, as they are applied nationally and, in the EU, as well as with the Horizon Europe rules. Concerning the results of the project, these will become publicly available based on the Access Rights as described in the Consortium Agreement.

7 Other issues

INCODE includes partners from third countries such as the UK (UWS and UMAN) and Switzerland (MARTEL) that they may have specific national, funder, sectorial, or departmental procedures for data management. These procedures may vary depending on the country and funding agency involved. In this chapter we briefly describe the main bodies that rule the data management procedures which will be more elaborated in the updates of this deliverable:

UK Research and Innovation (UKRI): UKRI is a funding agency in the UK that provides guidance on data management for research projects. It promotes the adoption of data management plans, data sharing, and open access to research outputs. UKRI expects in a similar fashion to Horizon Europe from researchers to adhere to the FAIR (Findable, Accessible, Interoperable, and Reusable) principles when managing and sharing data.

Swiss National Science Foundation (SNSF): SNSF is a major funding agency in Switzerland that supports research projects. They have specific guidelines on data management, including requirements for data sharing and data management plans. SNSF emphasizes the importance of making research data openly accessible and encourages data preservation.

Departmental or Sectorial Guidelines: Within third or associated countries, there may be specific departments or sectors that have their own procedures and guidelines for data management in research projects. For example, in the UK, specific research councils such as the Engineering and Physical Sciences Research Council (EPSRC) that may have their own data management requirements.

Hence, in the course of INCODE it is essential for researchers from these countries to adhere to guidelines and procedures provided by their respective national funding agencies, departments, or sectors. These guidelines may outline specific expectations for data management, including data sharing, preservation, and documentation. By following these procedures, these partners will ensure compliance with national regulations and contribute to the broader goals of promoting open science and research data sharing.

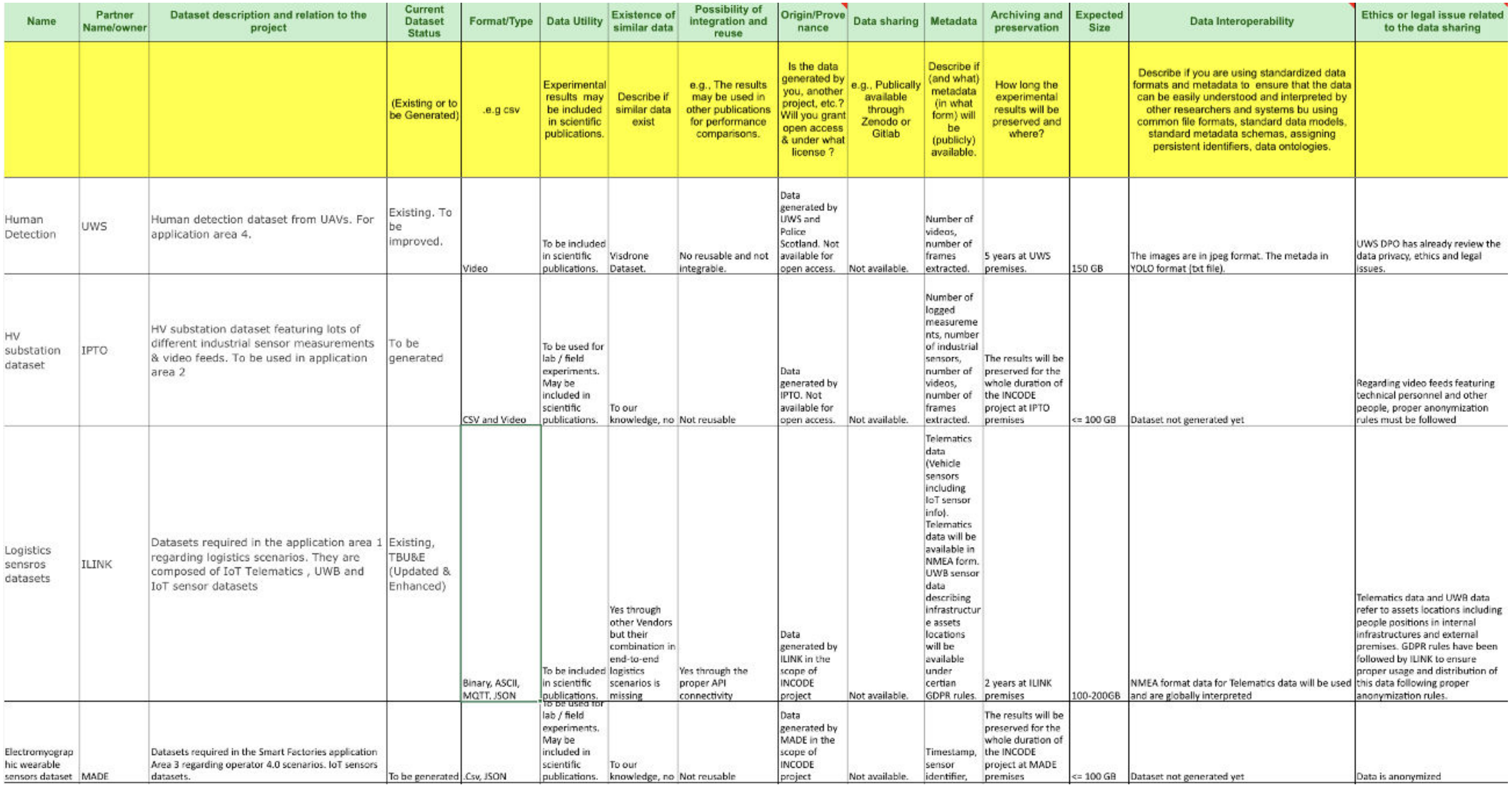
Co-funded by the European Union

8 Conclusions

This deliverable is the first version of the INCODE Data Management Plan and contains an initial description of datasets collected/generated in this first stage of the project. The described datasets may be of value for the project and will be exploited by the different tasks through the course of the project. The document will be updated every 6-months as the list of datasets is enriched with new information or datasets. Datasets use, sharing, preservation and dissemination aspects will be specified in all cases. All this updated information will be included in the future versions and revisions of the current document.

Co-funded by
the European Union

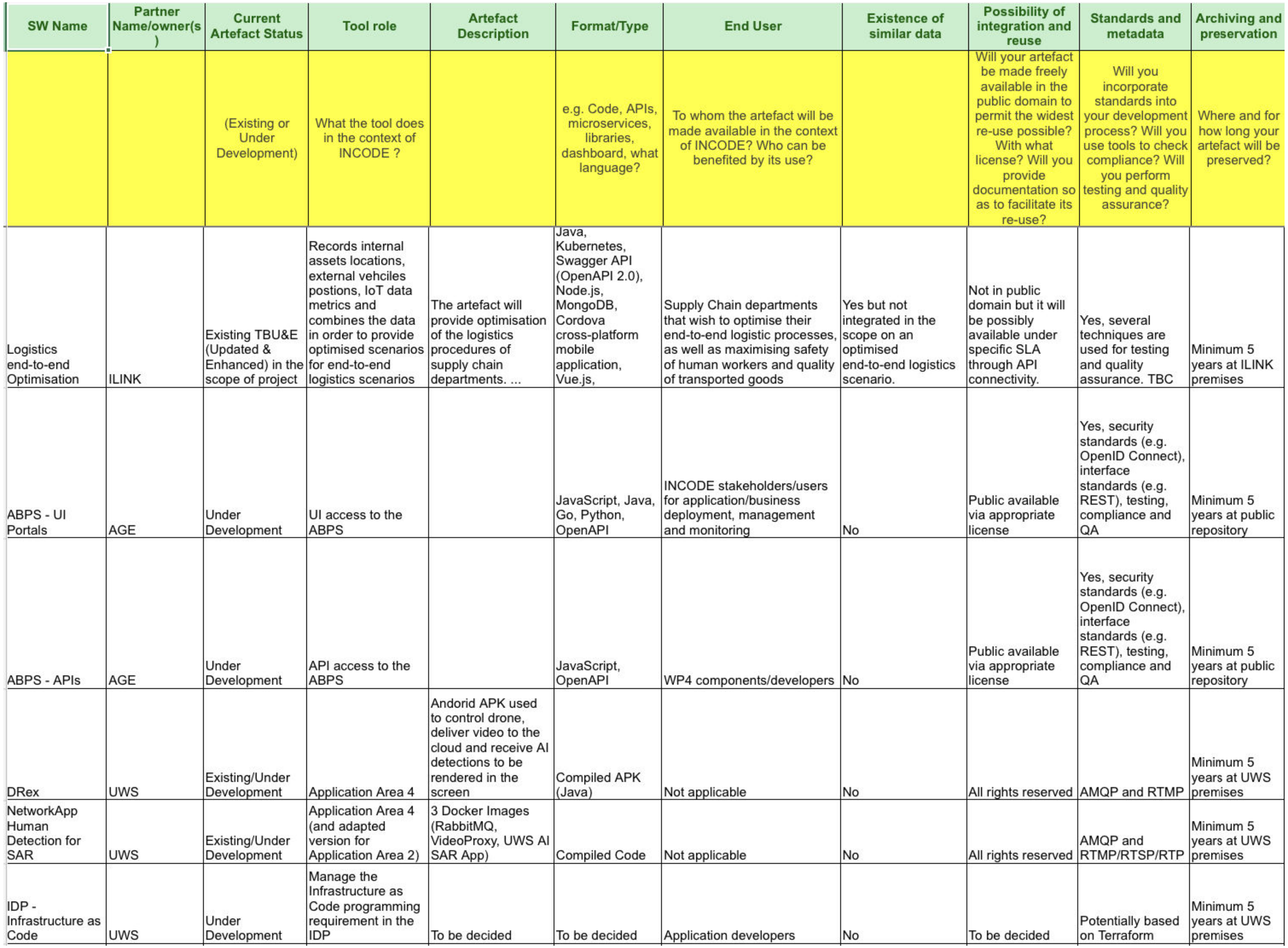
9 Appendix – Data Management Plan live repository



Name	Partner Name/owner	Dataset description and relation to the project	Current Dataset Status	Format/Type	Data Utility	Existence of similar data	Possibility of integration and reuse	Origin/Provenance	Data sharing	Metadata	Archiving and preservation	Expected Size	Data Interoperability	Ethics or legal issue related to the data sharing
			(Existing or to be Generated)	.e.g csv	Experimental results may be included in scientific publications.	Describe if similar data exist	e.g., The results may be used in other publications for performance comparisons.	Is the data generated by you, another project, etc.? Will you grant open access & under what license ?	e.g., Publically available through Zenodo or Gitlab	Describe if (and what) metadata (in what form) will be (publicly) available.	How long the experimental results will be preserved and where?		Describe if you are using standardized data formats and metadata to ensure that the data can be easily understood and interpreted by other researchers and systems bu using common file formats, standard data models, standard metadata schemas, assigning persistent identifiers, data ontologies.	
Human Detection	UWS	Human detection dataset from UAVs. For application area 4.	Existing. To be improved.	Video	To be included in scientific publications.	Visdrone Dataset.	No reusable and not integrable.	Data generated by UWS and Police Scotland. Not available for open access.	Not available.	Number of videos, number of frames extracted.	5 years at UWS premises.	150 GB	The images are in jpeg format. The metada in YOLO format (txt file).	UWS DPO has already review the data privacy, ethics and legal issues.
HV substation dataset	IPTO	HV substation dataset featuring lots of different industrial sensor measurements & video feeds. To be used in application area 2	To be generated	CSV and Video	To be used for lab / field experiments. May be included in scientific publications.	To our knowledge, no	Not reusable	Data generated by IPTO. Not available for open access.	Not available.	Number of logged measurements, number of industrial sensors, number of videos, number of frames extracted.	The results will be preserved for the whole duration of the INCODE project at IPTO premises	<= 100 GB	Dataset not generated yet	Regarding video feeds featuring technical personnel and other people, proper anonymization rules must be followed
Logistics sensros datasets	ILINK	Datasets required in the application area 1 regarding logistics scenarios. They are composed of IoT Telematics , UWB and IoT sensor datasets	Existing, TBU&E (Updated & Enhanced)	Binary, ASCII, MQTT, JSON	To be included in scientific publications.	Yes through other Vendors but their combination in end-to-end logistics scenarios is missing	Yes through the proper API connectivity	Data generated by ILINK in the scope of INCODE project	Not available.	Telematics data (Vehicle sensors including IoT sensor info). Telematics data will be available in NMEA form. UWB sensor data describing infrastructure assets locations will be available under certian GDPR rules.	2 years at ILINK premises	100-200GB	NMEA format data for Telematics data will be used and are globally interpreted	Telematics data and UWB data refer to assets locations including people positions in internal infrastructures and external premises. GDPR rules have been followed by ILINK to ensure proper usage and distribution of this data following proper anonymization rules.
Electromyographic wearable sensors dataset	MADE	Datasets required in the Smart Factories application Area 3 regarding operator 4.0 scenarios. IoT sensors datasets.	To be generated	.Csv, JSON	To be used for lab / field experiments. May be included in scientific publications.	To our knowledge, no	Not reusable	Data generated by MADE in the scope of INCODE project	Not available.	Timestamp, sensor identifier,	The results will be preserved for the whole duration of the INCODE project at MADE premises	<= 100 GB	Dataset not generated yet	Data is anonymized

Figure 2 INCODE Data sets at M6

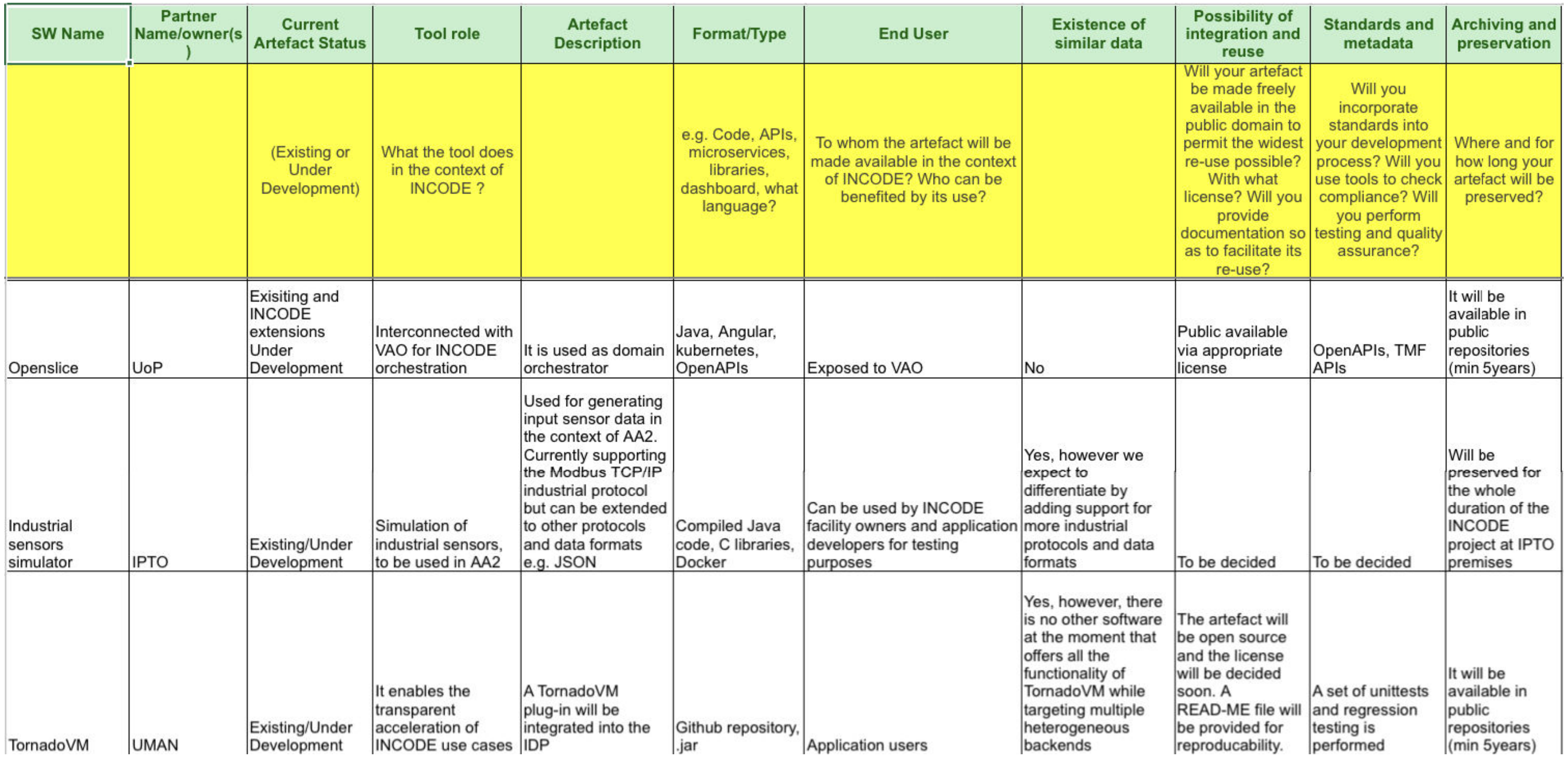
Co-funded by the European Union



SW Name	Partner Name/owner(s)	Current Artefact Status	Tool role	Artefact Description	Format/Type	End User	Existence of similar data	Possibility of integration and reuse	Standards and metadata	Archiving and preservation
		(Existing or Under Development)	What the tool does in the context of INCODE ?		e.g. Code, APIs, microservices, libraries, dashboard, what language?	To whom the artefact will be made available in the context of INCODE? Who can be benefited by its use?		Will your artefact be made freely available in the public domain to permit the widest re-use possible? With what license? Will you provide documentation so as to facilitate its re-use?	Will you incorporate standards into your development process? Will you use tools to check compliance? Will you perform testing and quality assurance?	Where and for how long your artefact will be preserved?
Logistics end-to-end Optimisation	ILINK	Existing TBU&E (Updated & Enhanced) in the scope of project	Records internal assets locations, external vehciles postions, IoT data metrics and combines the data in order to provide optimised scenarios for end-to-end logistics scenarios	The artefact will provide optimisation of the logistics procedures of supply chain departments. ...	Java, Kubernetes, Swagger API (OpenAPI 2.0), Node.js, MongoDB, Cordova cross-platform mobile application, Vue.js,	Supply Chain departments that wish to optimise their end-to-end logistic processes, as well as maximising safety of human workers and quality of transported goods	Yes but not integrated in the scope on an optimised end-to-end logistics scenario.	Not in public domain but it will be possibly available under specific SLA through API connectivity.	Yes, several techniques are used for testing and quality assurance. TBC	Minimum 5 years at ILINK premises
ABPS - UI Portals	AGE	Under Development	UI access to the ABPS		JavaScript, Java, Go, Python, OpenAPI	INCODE stakeholders/users for application/business deployment, management and monitoring	No	Public available via appropriate license	Yes, security standards (e.g. OpenID Connect), interface standards (e.g. REST), testing, compliance and QA	Minimum 5 years at public repository
ABPS - APIs	AGE	Under Development	API access to the ABPS		JavaScript, OpenAPI	WP4 components/developers	No	Public available via appropriate license	Yes, security standards (e.g. OpenID Connect), interface standards (e.g. REST), testing, compliance and QA	Minimum 5 years at public repository
DRex	UWS	Existing/Under Development	Application Area 4	Andorid APK used to control drone, deliver video to the cloud and receive AI detections to be rendered in the screen	Compiled APK (Java)	Not applicable	No	All rights reserved	AMQP and RTMP	Minimum 5 years at UWS premises
NetworkApp Human Detection for SAR	UWS	Existing/Under Development	Application Area 4 (and adapted version for Application Area 2)	3 Docker Images (RabbitMQ, VideoProxy, UWS AI SAR App)	Compiled Code	Not applicable	No	All rights reserved	AMQP and RTMP/RTSP/RTP	Minimum 5 years at UWS premises
IDP - Infrastructure as Code	UWS	Under Development	Manage the Infrastructure as Code programming requirement in the IDP	To be decided	To be decided	Application developers	No	To be decided	Potentially based on Terraform	Minimum 5 years at UWS premises

Figure 3 INCODE Software Artefacts at M6 (i)

Co-funded by the European Union



SW Name	Partner Name/owner(s)	Current Artefact Status	Tool role	Artefact Description	Format/Type	End User	Existence of similar data	Possibility of integration and reuse	Standards and metadata	Archiving and preservation
		(Existing or Under Development)	What the tool does in the context of INCODE ?		e.g. Code, APIs, microservices, libraries, dashboard, what language?	To whom the artefact will be made available in the context of INCODE? Who can be benefited by its use?		Will your artefact be made freely available in the public domain to permit the widest re-use possible? With what license? Will you provide documentation so as to facilitate its re-use?	Will you incorporate standards into your development process? Will you use tools to check compliance? Will you perform testing and quality assurance?	Where and for how long your artefact will be preserved?
Openslice	UoP	Exisiting and INCODE extensions Under Development	Interconnected with VAO for INCODE orchestration	It is used as domain orchestrator	Java, Angular, kubernetes, OpenAPIs	Exposed to VAO	No	Public available via appropriate license	OpenAPIs, TMF APIs	It will be available in public repositories (min 5years)
Industrial sensors simulator	IPTO	Existing/Under Development	Simulation of industrial sensors, to be used in AA2	Used for generating input sensor data in the context of AA2. Currently supporting the Modbus TCP/IP industrial protocol but can be extended to other protocols and data formats e.g. JSON	Compiled Java code, C libraries, Docker	Can be used by INCODE facility owners and application developers for testing purposes	Yes, however we expect to differentiate by adding support for more industrial protocols and data formats	To be decided	To be decided	Will be preserved for the whole duration of the INCODE project at IPTO premises
TornadoVM	UMAN	Existing/Under Development	It enables the transparent acceleration of INCODE use cases	A TornadoVM plug-in will be integrated into the IDP	Github repository, .jar	Application users	Yes, however, there is no other software at the moment that offers all the functionality of TornadoVM while targeting multiple heterogeneous backends	The artefact will be open source and the license will be decided soon. A READ-ME file will be provided for reproducability.	A set of unittests and regression testing is performed	It will be available in public repositories (min 5years)

Figure 4 INCODE Software artefacts at M6 (ii)

Co-funded by the European Union

Co-funded by
the European Union