InnovAite SLOVAKIA

Data Management Plan

Work Package 1: Project management

D1.3 Data management plan and its implementation report

Version: 0.1
Type of dissemination: N
Date of submission: M6
Authors:
Confidentiality: Public

www.innovaite.sk

Funded by the European Union NextGenerationEU

RECOVERY AND RESILIENCE PLAN

DEPUTY PRIME MINISTER'S OFFICE FOR THE RECOVERY PLAN AND THE KNOWLEDGE ECONOMY

VAIA RESEARCH AND INNOVATION AUTHORITY

Contents

1 Introduction

The Data Management Plan (DMP), developed under Task 1.3 of InnovAIte project, provides a structured framework for handling all data generated within the project. The first version was prepared at Month 6 (M6) and serves as a living document that will be regularly updated as the project evolves. The second version of DMP is expected to be created at Month 30 (M30). It ensures that data is managed in line with best practices and project requirements, supporting transparency, accessibility, and long-term usability.

The DMP covers both general aspects of data management and the specific requirements of the project's use cases. It addresses the full data lifecycle, including collection, storage, documentation, sharing, protection, and preservation. By defining clear procedures and responsibilities, it ensures consistency across partners and facilitates compliance with open science principles, European ethics recommendations, and relevant legal frameworks (Section 5). DMP was created in close cooperation with WP3 Cross domain AI and with application cases, WP4 AI-Driven buildings, WP5 AI-Driven Video, WP6 AI-Driven Software, and WP7 AI-Driven Insurance.

The consortium is currently in the process of negotiating specific conditions regarding the use of particular data with various stakeholders. At the same time, it is assessing the scope of implementation of the High-Level Expert Group's Ethics Guidelines for Trustworthy AI in relation to the chapter on ethics, as well as to another key output of the project, namely the Ethical AI Guidelines. For this reason, the relevant sections will be further elaborated and incorporated throughout the different phases of the project.

1.1. FAIR Principles and Open Science

Effective data management is necessary to ensure that research outputs are not only preserved but also maximized in their long-term value. Therefore, the project consortium unanimously agreed that the Data Management Plan will be guided by the FAIR Principles, which emphasize that data should be Findable, Accessible, Interoperable, and Reusable. Adopting these principles ensures that datasets are well-documented, stored in sustainable repositories, and shared in ways that make them easy to find, access, and combine with other resources so they can be confidently cited, reused, and built upon. Open access publications will be promoted to ensure that research findings remain accessible to the scientific community and the public. The approach seeks to enhance the impact of the research, facilitate collaboration, and contribute to an open and responsible scientific ecosystem.

- Research data will continue to be made available through appropriate data repositories, with proper documentation to facilitate data reuse.
- Research data will be deposited in appropriate repositories to ensure its accessibility for future researchers.
- Open access publications will be promoted to ensure that research findings remain accessible to the scientific community and the public.

1.2. Project Infrastructure and Storage

- HW and infrastructure of project partners.
- Repositories, “secure database”, open data.
- Possibility to use computation resources.
- For public data, Zenodo or other long-term repositories with curation policies will be used.
- Partners’ infrastructure: certification, service management and information security management,
- Trusted long-term repositories

  Internal and sensitive outputs will be archived in secure partner storage facilities with access restricted to authorised persons.

  Public data will share common agreement on security level.

  Partner´s data will have its own security agreement which is currently still under review of legal authorities on one hand in WP07 consortium partner processes and on the other hand on the side of data provider.

2 Data Summary

- Summary of data and metadata per WP.
- Input from the DMP form.
- Input data/output data (only a general description), intellectual property rights.

2.1. WP4 AI-Driven Buildings

This use case is focused on energy management sector which is undergoing a significant transformation due to growing pressure to reduce carbon emissions, lower energy consumption, and respond to volatile market dynamics. Traditional rule-based building management systems are no longer sufficient, as they often overlook system interdependencies. Artificial intelligence (AI) and machine learning (ML) offer a solution by enabling more advanced energy management through three key functions: observation (e.g., fault detection, activity recognition), prediction (e.g., energy demand, human behavior), and adjustment (e.g., system control, demand response). Reinforcement learning, in particular, shows strong potential for autonomous building management by leveraging large datasets and learning optimal policies without labelled data. The integration of AI/ML can significantly enhance efficiency, comfort, and sustainability, with potential cost savings up to $1.5 trillion by 2030. In a broader context, AI-based solutions also support dynamic, society-wide energy strategies that must adapt to the fluctuating nature of renewable sources and emerging trends such as electromobility.

Weather data such as temperature, wind, and solar radiation, energy consumption data, reference performance data (production, traffic, time of use, etc.), technical and operational metadata (consumption structure, appliance overview) will be used as input data for the digital twin of buildings, predictive and data analysis tools. Two types of data (i) data collected from the sensors and meters installed in the buildings, and (ii) simulation data will be generated with the purpose to create models and digital twins for energy efficiency management of buildings and industrial microgrids, perform analyses and simulations, and look for the optimal configuration of operating parameters. The expected size of the generated data is 10-100 TB. Re-used data (such as weather data) will be publicly available. Generated data will be obtained from sensors and meters installed in buildings, and from numerical simulations. The data will be useful for Researchers and developers of digital twins of buildings.

Findability

All data collected from sensors, as well as simulation data, will be equipped with a persistent identifier. Metadata will be created following established disciplinary and general metadata standards, where available, and will be structured to support the indexing by data repositories and catalogues. The metadata will include both descriptive and technical metadata. The plan is to adopt metadata standards relevant to building performance and environmental data, like:

- INSPIRE Metadata Implementing Rules (for geospatial datasets),
- ISO 19115 for geographic information metadata,
- IFC (Industry Foundation Classes) where BIM interoperability is involved,
- Dublin Core for general descriptive metadata,
- SensorML / OGC standards for sensor data streams,
- EnergyPlus or Modelica metadata schema for building simulation inputs and outputs,
- OpenBIM or COBie standards for facility management data.

If datasets are shared via repositories like ZENODO or national open-data repositories, their required metadata formats (often based on Dublin Core or DataCite) will also be adhered to. Search keywords will be provided in the metadata in a way to enable automatic harvesting and indexing.

Accessibility

At the end of the project, all data which could be made publicly accessible will be deposited in a trusted repository. Not all data generated during the project will be made openly available. This applies particularly to certain datasets collected through sensors and smart meters installed inside the buildings involved in the project, as well as simulation data generated by digital twins that are tailored to these specific buildings and usage patterns. Some of the data is subject to legal obligations that restrict open sharing. To balance openness with responsibility, the project will adopt the following measures:

- Metadata records of all datasets will be openly published, clearly indicating which data are restricted and why,
- Where possible, aggregated, synthetic, or anonymised versions of the data will be made available through project repositories or under controlled access,
- Requests for restricted datasets can be submitted via the formal application and reviewed by the data owner on a case-by-case basis.

Data will be accessible through a free and standardized access protocols. In the case of data restriction, the scope of the restriction will be specified (e.g. data can be used for research purposes but cannot be published publicly). Personal access will be limited to informed persons only, the data will be marked, and the extent of the restriction will be precisely described. Informed persons confirm their consent to the restricted use. In specific cases where it will be possible to expect the need for a specific SW tool for the access to data, the description will be in available at the metadata space.

Interoperability

To support interoperability, the following open and standardized data formats will be used:

- CSV, JSON, XML – for tabular sensor and monitoring data,
- IFC (Industry Foundation Classes) – for BIM models and architectural data,
- gbXML or IDF (EnergyPlus Input Data File) – for energy modelling data,
- FMU (Functional Mock-up Units) – for co-simulation of thermal systems,
- GeoJSON, GML – for geospatial data associated with buildings,
- NetCDF or HDF5 – for large-scale time-series or simulation outputs.

All formats chosen are machine-readable, well-documented, and have long-term community support. Metadata will follow the standards mentioned in Findability subsection above.

Semantic vocabularies and controlled terminologies will include:

- QUDT (Quantities, Units, Dimensions and Types) – for describing measured quantities,
- Brick Schema – for building metadata and ontology for sensors, zones, equipment,
- BOT (Building Topology Ontology) – for spatial relationships in buildings,
- SAREF (Smart Appliances REFerence ontology) – where energy appliances are involved.

Re-Usability

Each dataset will include a README.md or README.txt file that describes:

- Purpose of the dataset,
- Methodology used for data collection or simulation,
- Temporal and spatial resolution,
- Data source(s) and acquisition tools,
- File structure and naming conventions,
- Instructions for usage, including any software dependencies.

Structured codebooks will define:

- Variable names and full descriptions,
- Units of measurement,
- Value ranges or categories,
- Missing values treatment,
- Links to original standards (e.g. QUDT or Brick Schema when used).

A separate document (or README section) will explain:

- Data cleaning steps,
- Pre-processing logic and any filtering/interpolation applied,
- Assumptions and limitations,
- Model calibration or validation processes (if simulation data is included).

Where possible, data analysis will be conducted via scripts or notebooks (e.g., Jupyter, RMarkdown) that are version-controlled using Git. These will include inline comments and execution instructions to allow full reproduction of the analysis pipeline. Documentation will include the software versions, libraries, and dependencies used. For complex workflows, Docker images or environment.yaml files will be provided to ensure computational reproducibility.

Data Quality Assurance Process

For energy and weather data in the project, rigorous data quality assurance processes will be implemented. These processes will include the following steps:

- **Data Collection Standards**: Ensuring that data collection methods are standardized and consistent across all datasets to maintain high accuracy and reliability,
- **Validation and Verification**: Implementing thorough validation and verification protocols to check for errors, anomalies, and inconsistencies in the data. This will involve cross-referencing with historical data and other reliable sources,
- **Data Cleaning**: Performing regular data cleaning activities to remove any erroneous or irrelevant data points. This will ensure that the datasets remain accurate and relevant for analysis,
- **Metadata Documentation**: Documenting metadata comprehensively to provide context and provenance information. This will include details on data sources, collection methods, and any transformations applied to the data,
- **Automated Monitoring**: Implementing automated monitoring tools to detect and alert for any potential quality issues in real-time, allowing for prompt corrective actions.

These processes will ensure that the energy and weather data produced in the project are of the highest quality, making them reliable

2.2. WP5 AI-Driven Video

The aim of this work package is to enhance AI-based video intelligence through local data processing to enable faster reactions, increased energy efficiency and improved data security. The focus will be on two main areas: smart buildings and traffic safety. To maximise safety and security, algorithms and tools for object detection, human activity recognition and anomaly detection will be created and tested. A key feature will be the use of synthetic image data generated from virtual 3D scenes. This approach will enable our algorithms to be effectively

trained on complex scenarios that are difficult or even impossible to recreate. The computational capacity of the supercomputing centre will be used to generate and process this synthetic data.

Regarding traffic safety, the plan is to partially re-use existing 3D models of traffic objects. These models will be used to populate traffic scenes for generating training datasets. Reuse is appropriate to expedite development and ensure visual realism. The existing camera view scenes on the traffic infrastructure will be used as a basis for generating synthetic data. The use case will generate and use 3D model data in Blender-compatible formats (e.g., .blend, .obj, .fbx). Additionally, rendered 2D images in formats like .png or .jpg will be created from the 3D scenes. The 3D and 2D data will be used to simulate various traffic incident scenes. Image data will be utilized to train machine learning detection algorithms for traffic analysis and scene understanding. The total data volume is expected to range from 100 GB to 1 TB depending on the number of scenes and resolution of rendered images. Re-used 3D assets will be acquired from licensed third-party repositories. Generated data will be created in-house using Blender and a custom Blender addon. The data will be useful for researchers and developers working in computer vision, traffic analysis, simulation, and autonomous driving may benefit from the dataset.

Regarding the buildings, the plan is to partially re-use existing 3D models of objects related to buildings. These models will be used to populate buildings scenes for generating training datasets. The existing camera view scenes will be used in buildings or near to buildings as a basis for generating synthetic data. The use case will generate metadata over the inputs from cameras. This metadata will be linked to specific images and image sequences. Different levels and details of metadata will be generated. For example, a unique tagging of a person for re-recognition. The use case will use and generate 3D model data in Blender-compatible formats (e.g..blend, .obj, .fbx). Additionally, rendered 2D images in formats like .png or .jpg will be created from the 3D scenes. The 3D and 2D data will be used to simulate various security and safety incident scenes. Via LLM, a description of situations and their severity from the perspective of the situations being monitored will be generated. The total data volume is expected to range from 400 GB to 1,5 TB depending on the count of buildings and length of data collection. Generated metadata will be created in-house using camera, computing unit and SW. The data will be useful for researchers and developers working in computer vision, building safety and security analysis, simulation, and autonomous building security and safety systems, and for WP04 of this project (energy management of buildings may benefit from the dataset).

Findability

Data sets made public will be assigned DOIs or similar persistent identifiers through repositories such as Zenodo, etc. In both cases (traffic safety, building'), metadata will follow Dublin Core standards and include scene composition, object types, image parameters, and Blender configuration. Metadata will be structured in JSON or XML format.

Metadata will include appropriate keywords to optimize the possibility for discovery and will be deposited in harvestable formats compatible with major data repositories and indexing services.

Accessibility

All data which could be made publicly accessible will be deposited in a trusted repository. Preliminary considerations include Zenodo as a possible repository.

Not all data generated or collected in this use case will be made openly available. A subset of the data used and produced in this project is considered proprietary and is subject to legal and contractual restrictions. These data sets are owned or governed by third parties. Although raw or primary proprietary data cannot be shared, the project commits to making available derivative outputs, such as:

- Images generated from 3D models, visualizations, and post-processed renderings that do not allow re-identification or reverse engineering of the proprietary inputs,
- Aggregated or anonymized data summaries that serve the purposes of transparency and reproducibility while respecting the constraints.

These outputs will be published under appropriate licenses. The access will be provided via standard protocols used by repositories like Zenodo. Access to restricted parts will be granted upon request under licensing terms. No identity verification will be required for public data. For restricted parts, user identification may be handled through repository access controls.

For openly available data, their metadata will be made openly available under CC license. Metadata will contain information to access the data. The data and metadata will remain available for at least 5 years after the project ends. Metadata will be archived and maintained beyond the data retention period. References or documentation about software will be provided to access and read the data. Suitable open-source code will also be provided.

Interoperability

Data formats such as .blend, .obj, and image files will follow community standards. Metadata will follow Dublin Core, JSON-LD schemas, and FAIR data principles. Custom ontologies will be mapped to existing ones (e.g. schema.org) and published openly for reuse and extension. Data will include references to reused datasets and generated image derivatives for traceability.

Re-Usability

Documentation will include ReadMe files, necessary instructions, and dataset specifications. Publicly available data will be released with Creative Commons licenses (e.g. CC BY or CC0) wherever possible to maximize reuse. Data and tools will be documented and licensed to ensure post-project usability. Provenance records will be maintained using structured metadata formats and version control.

Data Quality Assurance Process

Validation will include visual inspections, consistency checks, and reproducible rendering pipelines.

Metadata Documentation: Documenting metadata comprehensively to provide context and provenance information.

2.3. WP6 AI-Driven Software

The software development process consists of several stages, including analysis, design, implementation, testing, deployment and maintenance. The volume of research into the inclusion of AI in software engineering increased before the emergence of LLMs and has increased even more since then due to the emergent abilities of LLMs, e.g. enabling AI-assisted code completion. This work package aims to explore the possibilities and propose novel approaches to utilising AI in the various stages of software development. The focus is on the automated transformation of software artefacts relevant to one development stage into artefacts relevant to another stage, and the AI-assisted modification of software development artefacts.

Within this work package internal data from Asseco Central Europe and Gratex International will be used, which includes software analytical artefacts, architectural designs, code, test scripts, and other SDLC components. This data will be used:

- as input for the development of specialised AI agents in individual SDLC phases,
- for prompt engineering of LLM models,
- in some cases, also as training data for fine-tuning local LLMs in a secure environment,
- for Retrieval Augmented Generation (RAG) databases.

If necessary, publicly available external data sources will be used (e.g., open-source repositories such as GitHub) for pre-training or fine-tuning LLM models. However, this external data will not be used without thorough verification of its quality, relevance, and security.

Various types of data related to the individual phases of the software development life cycle will be generated and used. These will mainly include text analysis documents, architectural models, structured designs, source code, test scripts, outputs from AI tools, and prompt templates. This data will represent the inputs and outputs of AI agents and LLM, which will be used to automate processes within the SDLC. The data will also include generated representations and outputs from models created during training or inference.

The goal is to automate activities throughout the entire software development lifecycle using AI agents and LLM. The data will enable:

- creation of domain-specific agents for individual SDLC phases (analysis, design, testing, maintenance),

- training and tuning LLM models to understand the structures, terminology and processes typical of software engineering in the Asseco/Gratex environment.

The expected data size includes a combination of reused internal artefacts and newly created data during the development and training of AI solutions:

- **Reused data**: approximately 30–50 GB, mainly in the form of existing documentation, development artefacts, models and source code from current internal projects,
- **Generated data**: approximately 50–100 GB, including synthetic prompts, embedded representations, AI agent outputs, source code versions, and training files for fine-tuning local LLM models,
- **Total scope**: up to 150 GB, including version history and model checkpoints. This estimate also considers the need to store data for experiments and repeated validation.

Reused data comes from internal projects of Asseco and Gratex (e.g. documentation, code, models). Generated data will be created during the project as outputs of AI agents and LLM models. Some data may also come from public sources (e.g. GitHub), but it will be thoroughly verified before use. Most data will be for internal use only due to confidentiality. Only anonymised or synthetic outputs that may be useful to researchers, AI tool developers or the academic community in developing and testing AI solutions in the field of software engineering may be published.

Findability

DOI or other persistent identifier will be used for public data. Detailed metadata describing the content, authorship and use will be created. Metadata will consist of types as title, description, author, version, date, data type, licence, access conditions. Dublin Core and DCAT standards wil be used. For specific data types, custom internal structures will be added. Relevant keywords will be added to optimize the possibility for discovery in a machine-readable format (e.g. JSON-LD) compatible with harvesting tools.

Accessibility

Selected data will be stored in a trusted repository with the possibility to assign a permanent identifier (e.g. DOI). Most data will remain accessible only within the consortium for reasons of intellectual property and sensitive information. The reasons are contractual, security and commercial. Only selected, anonymised or synthetic data will be shared publicly. An embargo may be applied for a period of up to 12 months if the data relates to publications or intellectual property protection (e.g. patent applications). This period is in line with the common practice of scientific publishers and EU framework programmes (e.g. Horizon Europe), which allow temporary restrictions on access to enable researchers to prepare for publication or secure legal protection for their outputs before they are made available. The public data will be accessible through the standardized access protocols like HTTPS, OAI-PMH, etc. In the case of restricted use,

the data will be available on request, with approval from the responsible person and, if applicable, signature of a license agreement. Verification will be done via institutional account, project identifier or contractual access. An approval mechanism will be created for access to sensitive or internal data.

Metadata for public data will be available and open under a CC0 license. Metadata will contain information on how to access them and contact details. Public data will remain available and findable at least 5 years after the end of the project to meet the minimum recommended archiving period in accordance with the requirements of grant schemes. Metadata will remain available even after the data itself has been deleted. The necessary documentation and any open-source software will be included in the sharing.

Interoperability

Generally accepted formats and standards such as JSON, XML, YAML, together with DCAT and Dublin Core metadata standards will be used. Standards such as UML, ArchiMate and JSON Schema will be used for models and architectures. FAIR principles, OpenAIRE recommendations, and methodologies used in open software development (e.g., versioning, API specifications, modularity) will be followed. If project-specific dictionaries or structures are created (e.g. for AI components or prompt types), they will be designed so that they can be mapped to common standards. The data will be linked to other relevant project outputs (e.g. outputs from WP3 and WP8), thereby promoting their reusability and contextualization.

Re-Usability

Each shared data set will be accompanied by a README file describing the methodology, data structure, variable types, processing methods and related outputs. Only selected data that is not confidential or contractually bound will be published. Public data will be licensed under standard licenses such as CC BY or CC BY-NC. Published outputs will be usable after the end of the project, including documentation, templates and selected training data. All data flows will be versioned and recorded with a history of origin, processing and purpose of use, primarily using Git and internal tools.

Data Quality Assurance Process

Data will be validated continuously, both manually and automatically. Uniform structures, checksums, versioning and feedback from AI outputs will be used. Key data outputs will be subject to peer review or internal approval.

2.4. WP7 AI-Driven Insurance

AI is increasingly used in the insurance industry to optimize risk assessment, automate claims processing, detect fraud, and enhance customer service through chatbots and virtual assistants. It supports predictive analytics for trend forecasting and strategic decision-making, while automating administrative tasks to boost efficiency and accuracy. Despite its potential, effective AI deployment faces challenges such as a shortage of skilled professionals, poor data quality, lack of standardized business rules, and the need for a solid business and IT architecture. Addressing these issues is essential for fully unlocking AI's benefits, which include improved operational efficiency, regulatory compliance, and competitive advantage in a rapidly evolving market.

Training data for this work package will be newly collected from data owners. No reuse of any existing data is possible due to the strict insurance market regulations. Intellectual rights protection is applied. Structured data and NLR business rules written in SBVR[1] notation mapped to the selected insurance reference model will be used. The purpose of the generated data is that synthetic data generation requires a better LLM model training to cover unusual or specific data and process states not present within real data. The expected size of data is under evaluation at the moment of writing this deliverable. Some specific results from publicly available training data might be useful to market regulators and existing or future customers of any insurer carrier.

Findability

Generated data will be identified by a persistent identifier. Metadata will be provided to allow data discovery.

Metadata are organised into these categories:

Metadata Category	Purpose
Administrative Metadata	Technical details to manage and retrieve the document
Descriptive Metadata	Content-related attributes for search & discovery
Structural Metadata	Defines how the document's parts relate to each other
Process Metadata	Information about the document's role in the claim workflow
Compliance & Security Metadata	Retention rules, confidentiality levels, access rights

Metadata definition will be executed compliant with the SBVR v1.5 by the RuleSpeak Business Rule Notation to create data dictionary for all metadata fields

Data dictionary will consist of these mandatory attributes: Metadata Field Name, Business Definition, Example, Data Type, ACORD Class, ACORD Attribute, Sensitivity Level, Ethical Preprocessing Algorithm, Keywords, Note , Controlled Vocabulary Source

Controlled Vocabulary will be enforced for common code list (preferable ACORD's)

[1] Semantics of Business Vocabulary and Business Rules: https://www.omg.org/spec/SBVR/1.5/About-SBVR

Metadata will keep link to the Master Data via ClaimId, PolicyId and DocumentId

All sensitive data are secured by depersonalisation, anonymisation or encryption base on Data Owner agreed Ethical Preprocessing Algorithm

Metadata will respect Version Control for audit trails

Appropriate keywords will be provided in the metadata by a way to allow automatic harvesting and indexing.

For any dataset we decide to publish (e.g., synthetic/constructed), we will assign a persistent identifier (DOI).

Accessibility

Data will be deposited in a trusted repository. The appropriate repository is under the decision and evaluation at the moment of writing this deliverable. Insurance carrier data is under embargo, selected synthetic data will be openly available and accessible through a free and standardized access protocol. Access to insurance data will only be granted via an agreement with the data provider.

Metadata of insurance carrier data are under embargo, metadata of synthetic data will be available.

Metadata will be available for time period defined specifically by agreement with the data provider. Generally from the date of agreement sign till end of development process of AI model training/first outputs of PoC on data providers side. In case of increased satisfaction rate on provider´s side through improved claims management process we expect to get data allowance till end of the project or even further to deploy the solution on provider´s environment.

Metadata will not be findable as suggested and agreed from accesses other than WP07 with specified agreement & approval for data management from provider.

Each of participants based on Ethical standards and data management agreement will define own reference and software used per data provider participating.

Real insurance data remain non-public under embargo and are accessible only under provider agreements. Only pseudonymised non-public datasets (not raw customer data) are hosted in trusted EU-based storage with role-based access. If third-party AI services are used, we operate under EU data residency and zero data retention modes where available.

Interoperability

Adopting common insurance reference model like ACORD eliminates custom mapping for each partner and new integrations with insurers, MGA, broker or claim service provider are cheaper and faster.

ACORD ensure consistent shared meaning and reduces misunderstanding between any collaborating business or technical units.

Example: “Effective Date” in ACORD is always policy start date in the contract term, not an underwriting approval date.

Re-Usability

Training Data from selected partners and participating insurers must not be shared or exposed to the public domain. There are also restrictions on sharing data between project members.

Licenses will be specified, based on the granted access via agreement with the provider of data. Synthetic data could be usable by third parties, with the documentation of the data provenance by using the appropriate standards.

Data Quality Assurance Process

Data quality assurance process will be evaluated during the project. Data Quality Rules will be specified compliant with SBVR v1.5 by the RuleSpeak Business Rule Notation.

Prior to sharing or publication, datasets pass automated quality checks (completeness, format, consistency) and maintain provenance/lineage records. Pseudonymisation/anonymisation is performed locally by WP07 prior to any onward processing or transfer; operational logs and monitoring exclude raw PII.

3 Other Outputs

WP4 -will also produce outputs other than data. These will mainly be digital outputs, such as:

- algorithms
- software / software tools
- procedures and methodologies
- will use own repository for source code

WP5 will also produce outputs other than data. These will mainly be digital outputs, such as:

- Algorithms
- Software / software tools

WP6 will also produce outputs other than data. These will mainly be digital outputs, such as:

- software tools (e.g. AI assistants, testing modules, refactoring tools),
- prompt templates and prototypes,
- AI-generated architectural designs,
- documented workflows for integrating AI into SDLC.

WP7 will also produce outputs other than data.

- software tools (e.g. AI assistants, testing modules, refactoring tools, ethical preprocessing data tools, Vision models and tools for unstructured data extraction from documents),
- prompt templates and prototypes
- documented claim processing workflows compliant with ACORD

4 Ethics

To facilitate the analysis of the potential ethical issues, the organizational structure of the project designated the DFKI Ethics Team to monitor the ethical aspects. The DFKI Ethics Team will ensure that all project activities will be conducted in accordance with the Horizon Europe Ethics Appraisal Process and will provide, when necessary, templates of informed consent forms, ethics information sheets, ethics approvals and ethics compliance votums, and any other required ethics documentation.

The project has launched a workstream to evaluate the scope and methods of implementing the HLEG's Ethics Guidelines for Trustworthy AI as its ethical foundation, since HLEG combines clear normative authority, regulatory alignment, practical tools, and stronger auditability. The following seven HLEG's principles: human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination and fairness; societal and environmental well-being; and accountability; provide a compact and widely recognized ethical reference. HLEG is not merely aspirational: it has regulatory relevance in EU policy and offers ready-to-use instruments such as the ALTAI (Assessment List for Trustworthy AI) that the project will apply to translate principles into verifiable actions.

InnovAIte will also ensure compliance with the European Code of Conduct for Research Integrity, the Charter of Fundamental Rights of the European Union and the European Convention on Human Rights together with its supplementary protocols. Any data processing activities during the execution phase of the research and development and pilot and validation cycles will be in compliance with the Regulation (EU) 2016/679 (GDPR), the ePrivacy Directive 2002/58/EC (and its forthcoming revisions).

With the overall framework established, the Data Management Plan proceeds by addressing the main aspects of data handling in greater detail. The sections that follow highlight the priorities that will shape how data is organized, safeguarded, and made available, ensuring consistency and coherence throughout the project's lifecycle.

4.1. Fairness and Non-Discrimination

Data will be collected and used in a manner that promotes fairness and prevents systemic bias. Proactive measures will be implemented to ensure that AI systems do not perpetuate or amplify unfair biases or discrimination based on race, gender, age, religion, or any other protected characteristic.

4.2. Transparency and Accountability

Comprehensive documentation will be maintained regarding data sources and data cleaning procedures. All data-related decisions and their impact on model outcomes will be made transparent and understandable to relevant stakeholders.

4.3. Privacy and Data Protection

- The collection and processing of personal data will fully comply with all applicable legal regulations, particularly Regulation (EU) 2016/679 of the European Parliament and Council (the General Data Protection Regulation, "GDPR").
- Core principles of the GDPR will be upheld, including the establishment of a legal basis for processing, specification of the processing purpose, data minimization, and the protection of the rights and legitimate interests of data subjects.
- Personal data will only be processed on the basis of:
    - the fulfilment of contractual obligations, or
    - legitimate interest, or
    - explicit consent, or
    - compliance with legal obligations.
- Personal data will be used exclusively for training purposes and in accordance with the principle of data minimization — only data that is strictly necessary and relevant to the stated purpose will be collected and processed.
- Data subjects will be clearly informed of their rights under the GDPR, including the rights to access, rectification, restriction of processing, data portability, objection, withdrawal of consent, and the right to lodge a complaint with a supervisory authority.
- The principles of privacy by design and by default will be applied. This includes encryption of data both in transit and at rest, and pseudonymization wherever feasible.
- Personal data collection will be based on informed consent, with individuals retaining the right to define the scope of data collected and to revoke their consent at any time.
- State-of-the-art encryption and access control mechanisms will be employed to protect data from unauthorized access, breaches, and misuse.

4.4. Purpose Limitation

Data will be used strictly for the specific purposes disclosed at the time of collection. Any use beyond the original purpose will require additional consent or a clearly articulated ethical justification.

4.5. Data Quality and Integrity

All data will be verified for accuracy, completeness, and currency prior to and during its use in AI systems. Inputs and outputs will be audited regularly to detect and correct errors, inconsistencies, or unintended effects.

4.6. Intellection property rights

- In cases where data constitutes a copyrighted work or involves objects of industrial property, all applicable intellectual property rights will be respected.
- Such content will be used only under valid licensing agreements, either through rights granted by the rights holder or under the terms of a public license.
- All entities involved will ensure compliance with relevant obligations to prevent unauthorized use or infringement of intellectual property rights.

4.7. Ethics in WPs

4.7.1. WP4 AI-Driven Buildings

For WP04, the initial analysis has determined that the relevant data will comprise records from measuring instruments and outputs from control and monitoring systems with the required resolution and duration, data on energy consumption, and reference performance data (including, but not limited to, production figures, visitor attendance, and utilization times).The processing of such data entails not only strict compliance with applicable legal obligations — including the protection of trade secrets, database rights, audiovisual recordings, and adherence to GDPR requirements — but also the observance of fundamental ethical principles of data management. This includes obtaining the explicit consent of the authorized data holder (where the data is not publicly available), ensuring fairness, transparency, and the protection of the rights of all data subjects. Such consent shall be formalized through a separate legal instrument. If the data is classified as a protected database, the appropriate consent and permissions from the rights holder shall likewise be secured prior to any use. Given the sensitive nature of the data, an ethical assessment will be conducted in parallel with legal and technical evaluations, with particular emphasis on privacy, accountability, and data minimization principles, especially in relation to the collection and processing of camera recordings. The procedure shall adhere to the general rules set out in paragraph 5.3, with additional consideration given to the principle of proportionality (collecting only data necessary for the intended purpose) and the ethical duty to prevent potential misuse. Where the designated entity holds the data on a valid legal basis, a clearly defined anonymization process shall be implemented for any subsequent use, ensuring both compliance with legal requirements and the ethical imperative of protecting individual identities. In cases where camera recordings also constitute audiovisual works, the producer's consent will be obtained, with full respect for intellectual property rights.

4.7.2. WP5 AI-Driven Video

Within WP05, the collected and analysed data have been divided into two main categories – building surveillance footage and traffic data. Each of these categories has specific ethical and legal considerations that will be considered during their processing. Building surveillance footage will be processed in compliance with the requirements of the GDPR and relevant national legislation, with particular emphasis on the protection of personal data, the right to privacy, and the protection of personality rights, as well as the safeguarding of the rights of the audiovisual recording's producer. All recordings will be used exclusively for the declared purposes of the project and will be subject to technical and organisational measures preventing unauthorised disclosure or misuse, including anonymisation or blurring of identifiable elements when necessary.

Traffic data will be processed with the same emphasis on compliance with legislative requirements. It is anticipated that the quality of the collected recordings will not allow the identification of individuals or vehicles; therefore, they will not contain personal data. Nevertheless, measures will be implemented to ensure the protection of the producer's rights to the audiovisual material, and in cases where the dataset meets the legal definition of a database, the corresponding rights to the database will be granted and duly regulated.

4.7.3. WP6 AI-Driven Software

The work within the WP06 will, in terms of the nature of the data, encompass all relevant information associated with the software development lifecycle. This includes, in particular, source code, analytical documentation, user manuals, technical specifications, as well as database structures and related datasets that are created or employed in this context.

The consortium will ensure that, whenever any of these outputs qualify as copyright-protected works under the applicable legislation—whether as a computer program, database, or literary work—they will be managed in full compliance with the relevant licensing conditions. This will be achieved either by obtaining explicit consent from the rights holder for the use of the copyrighted work, or, in cases where works are made available under a public license, by adhering strictly to the specific terms and conditions of the respective license.

Special attention will be given to transparent and legally sound handling of copyrighted materials, with the objective of minimizing the risk of infringing third-party rights. The processes will include maintaining detailed records of sources on which individual components are based, as well as

systematically verifying that the licensing regimes permit their integration into the project's final outputs.

4.7.4. WP7 AI-Driven Insurance

The integration of Artificial Intelligence (AI) into insurance operations offers unprecedented opportunities in underwriting, claims management, fraud detection, and customer engagement. However, AI systems must be designed and deployed in ways that protect policyholders, maintain trust, and comply with ethical and legal standards.

Figure 1 - WP07 - Ethics in AI-Driven Insurance

Strategic Imperative

Ethical AI in insurance is not only a compliance necessity, but also a competitive advantage. Insurers who integrate transparency, fairness, and trust into AI models will gain stronger customer loyalty, regulatory goodwill, and market differentiation.

Key Ethical Principles

1. **Transparency & Explainability**
   AI decisions — from premium pricing to claims settlement — must be explainable in clear, human terms. Black-box algorithms risk eroding customer trust and may breach regulatory requirements.
2. **Fairness & Non-Discrimination**
   AI must be tested to prevent bias based on age, gender, ethnicity, geography, or socio-economic status. Data sources and training sets should be representative and balanced.
3. **Privacy & Data Protection**
   Personal and claims data should be processed under GDPR-by-design principles: data minimisation, pseudonymisation, and strict access controls. AI models should use the least personal information necessary to achieve their objectives.
4. **Accountability & Governance**
   Insurers must maintain clear accountability for AI-driven decisions, with defined oversight roles, audit trails, and the ability to override automated outcomes when necessary.
5. **Security & Robustness**
   AI systems must be resilient to cyberattacks, manipulation, or data poisoning. Regular security testing, model validation, and monitoring are essential.
6. **Human Oversight**
   AI should augment, not replace, human judgement. High-impact decisions, especially in claims disputes or policy cancellations, must involve qualified human review.
7. **Social Responsibility**
   AI deployment should align with the insurance industry's broader role in risk mitigation and societal resilience, avoiding harm to vulnerable customers.

5 Data Outputs Plan

5.1. WP4 AI-Driven Buildings

Based on input data (artificial data, reference projects, energy consumptions, metadata), data models will be created for individual reference cases, analytical and control procedures will be compiled, and the resulting conclusions will be validated. For the purposes of the project, conclusions are understood as sets of datasets and procedures (obtained from the transmitted data) aimed at optimizing energy consumption, which generate specific knowledge conclusions through algorithms that identify the methods, places and time periods of inefficient energy use. Based on these findings, it is possible to determine exactly where and when losses occur, and how they arise. The conclusions provide the basis for designing effective measures that lead to increased energy efficiency, reduced operating costs and improved energy consumption management in the monitored objects or processes.

Item	Description
Dataset ID	WP4 _OUT_unefficiency
Dataset Name	Energy unefficency vectors
Dataset Reference	To be specified later
Standards, Format, and Metadata	Time series of calculated energy inefficiency (can be structured)
Data Sharing (incl. License)	NO
Is Dataset Accessible?	Only for POC partners providing input data (optional public after confirmation)
Is Dataset Reusable?	NO
Archiving and Preservation (Storage and Backup)	Stored in safety cloud space

Item	Description
Dataset ID	WP4 _OUT_model
Dataset Name	Paramaters of energy consumption models
Dataset Reference	To be specified later
Standards, Format, and Metadata	Model parametres, time series of energy performance indicators
Data Sharing (incl. License)	NO

Is Dataset Accessible?	Only for POC partners providing input data (optional public after confirmation)
Is Dataset Reusable?	NO
Archiving and Preservation (Storage and Backup)	Stored in safety cloud space

Item	Description
Dataset ID	WP4 _OUT_prediction
Dataset Name	Time series of predicted energy consumption
Dataset Reference	To be specified later
Standards, Format, and Metadata	Model parametres, time series of energy performance indicators
Data Sharing (incl. License)	NO
Is Dataset Accessible?	Only for POC partners providing input data (optional public after confirmation)
Is Dataset Reusable?	NO
Archiving and Preservation (Storage and Backup)	Stored in safety cloud space

Item	Description
Dataset ID	
Dataset Name	
Dataset Reference	To be specified later (DMP v2.0)
Standards, Format, and Metadata	Model parametres, time series of energy performance indicators
Data Sharing (incl. License)	NO
Is Dataset Accessible?	Only for POC partners providing input data (optional public after confirmation)
Is Dataset Reusable?	NO
Archiving and Preservation (Storage and Backup)	Stored in safety cloud space

5.2. WP5 AI-Driven Video

Item	Description
Dataset ID	10.5281/zenodo.17754828
Dataset Name	Street 3D Assets for Blender
Dataset Reference	To be defined
Standards, Format, and Metadata	BLEND, md, txt
Data Sharing (incl. License)	Yes (CC-BY)
Is Dataset Accessible?	Yes
Is Dataset Reusable?	Yes
Archiving and Preservation (Storage and Backup)	Zenodo

Item	Description
Dataset ID	DSID_WP05_002
Dataset Name	image_data
Dataset Reference	To be defined
Standards, Format, and Metadata	JPEG, PNG, JSON
Data Sharing (incl. License)	Yes (detailed specification – later)
Is Dataset Accessible?	Yes
Is Dataset Reusable?	Yes
Archiving and Preservation (Storage and Backup)	To be defined (Zenodo, Hugging Face, GitLab)

5.3. WP6 AI-Driven Software

Item	Description
Dataset ID	DSID_WP06_001
Dataset Name	usecase_to_testcase
Dataset Reference	https://github.com/navidh86/usecase-to-testcase

Standards, Format, and Metadata	Apache Arrow, JSON
Data Sharing (incl. License)	Yes (per standard scientific acknowledgement)
Is Dataset Accessible?	Yes
Is Dataset Reusable?	Yes
Archiving and Preservation (Storage and Backup)	GitHub

Item	Description
Dataset ID	DSID_WP06_002
Dataset Name	BeIR
Dataset Reference	https://huggingface.co/BeIR
Standards, Format, and Metadata	JSONL, TSV
Data Sharing (incl. License)	Yes (per standard scientific acknowledgement)
Is Dataset Accessible?	Yes
Is Dataset Reusable?	Yes
Archiving and Preservation (Storage and Backup)	HuggingFace

Item	Description
Dataset ID	DSID_WP06_003
Dataset Name	GTI – Internal Data
Dataset Reference	N/A
Standards, Format, and Metadata	Python, SQL, NL, JSON, XML, DOCX, XSLX
Data Sharing (incl. License)	No, except data we decide to share publicly, e.g. constructed datasets
Is Dataset Accessible?	No
Is Dataset Reusable?	No
Archiving and Preservation (Storage and Backup)	GitHub, GitLab

Further research may raise need for more benchmark or training datasets – will be added in future.

5.4. WP7 AI-Driven Insurance

Item	Description
Dataset ID	DSIW_WP07_001
Dataset Name	InsuranceCoverage
Dataset Reference	To be specified later (DMP v2.0)
Standards, Format, and Metadata	JSON
Data Sharing (incl. License)	No, except data we decide to share publicly, e.g. constructed datasets
Is Dataset Accessible?	No
Is Dataset Reusable?	Yes
Short summary	**Core policy & coverage attributes for model training/evaluation; confidential operational data.**
Provenance	**Real-world (confidential)**

Item	Description
Dataset ID	DSIW_WP07_002
Dataset Name	ClaimStream
Dataset Reference	To be specified later (DMP v2.0)
Standards, Format, and Metadata	JSON
Data Sharing (incl. License)	No, except data we decide to share publicly, e.g. constructed datasets
Is Dataset Accessible?	No
Is Dataset Reusable?	Yes

Short summary	**Event-level claim timeline from first notice of loss to closure (statuses & actions); confidential operational data.**
Provenance	**Real-world (confidential)**

Item	Description
Dataset ID	DSIW_WP07_003
Dataset Name	KnowledgeOnthology
Dataset Reference	To be specified later (DMP v2.0)
Standards, Format, and Metadata	JSON
Data Sharing (incl. License)	No, except data we decide to share publicly, e.g. constructed datasets
Is Dataset Accessible?	No
Is Dataset Reusable?	Yes
Short summary	**Controlled vocabulary & relationships for insurance terms to support consistent tagging and prompts; synthetic.**
Provenance	**Synthetic / Constructed**

Item	Description
Dataset ID	DSIW_WP07_004
Dataset Name	ClaimPerformance
Dataset Reference	To be specified later (DMP v2.0)
Standards, Format, and Metadata	JSON
Data Sharing (incl. License)	No, except data we decide to share publicly, e.g. constructed datasets
Is Dataset Accessible?	No
Is Dataset Reusable?	Yes
Short summary	**Aggregated KPIs/metrics to evaluate models and process outcomes; no PII.**
Provenance	**Derived / Aggregated (from confidential)**

Item	Description
Dataset ID	DSIW_WP07_005
Dataset Name	ComplianceRulesSBVR
Dataset Reference	To be specified later (DMP v2.0)
Standards, Format, and Metadata	JSON
Data Sharing (incl. License)	No, except data we decide to share publicly, e.g. constructed datasets
Is Dataset Accessible?	No
Is Dataset Reusable?	Yes
Short summary	**Business & compliance rules in SBVR v1.5 guiding preprocessing and decision logic; synthetic.**
Provenance	**Synthetic / Constructed (rules & taxonomy)**

6 Conclusion

- IPR problematics in some WPs will be solved during the project
- Continuous update of the DMP during the project

The InnovAIte Data Management Plan provides a comprehensive framework for managing all data generated within the project, ensuring alignment with the FAIR principles and Open Science standards. It defines clear procedures for collection, storage, documentation, accessibility, and preservation, while addressing legal, ethical, and security requirements.

Each work package (WP4 – AI-Driven Buildings, WP5 – AI-Driven Video, WP6 – AI-Driven Software, WP7 – AI-Driven Insurance) has identified its specific datasets, metadata standards, and quality assurance processes. While some outputs will remain confidential due to intellectual property rights, security, or regulatory constraints, the project commits to maximizing openness through trusted repositories, anonymized or synthetic data sharing, and transparent metadata records.

Ethical considerations remain at the core of data handling, ensuring fairness, privacy, accountability, and compliance with European regulations such as GDPR. The plan also recognizes the importance of interoperability and reusability by adopting open formats, standardized vocabularies, and reproducible workflows.

As a living document, the DMP will be continuously updated throughout the project lifecycle. This guarantees adaptability to evolving requirements, stakeholder agreements, and technological advancements, while supporting the long-term usability and impact of the project's outputs. Ultimately, the InnovAIte Data Management Plan not only secures responsible and sustainable data governance but also contributes to building a trustworthy, ethical, and innovative AI ecosystem in Slovakia and beyond.

About the project

The InnovAIte Slovakia project was launched in response to the call "Transformation and Innovation Consortia" announced by the Government Office of the Slovak Republic under the Recovery and Resilience Plan of the Slovak Republic (Component 9 – Investment 2: Supporting cooperation between companies, academia, and research and development organizations). The call aims to drive systemic transformation and increase the added value of key sectors of the Slovak economy through intensive collaboration among research institutions, innovative enterprises, the public sector, and internationally renowned partners.

InnovAIte Slovakia seeks to build a dynamic and sustainable innovation ecosystem in the field of artificial intelligence (AI), effectively linking cutting-edge research with practical applications. The project emphasizes the development of AI solutions that are not only technologically advanced but also ethical, environmentally sustainable, and socially beneficial. The consortium brings together leading research centres, universities, and businesses from Slovakia, Germany, and the Czech Republic.

Key focus areas include the development of AI algorithms for improving building energy efficiency, enhancing traffic safety through video analytics, automating software development, driving digital transformation in the insurance sector, and validating functional prototypes. Special attention is also given to AI education, talent development, and the incubation of startups.

The project's outcomes will significantly strengthen Slovakia's position in research and innovation, reduce environmental burdens, increase the competitiveness of the national industry, and modernize public services. By bridging the public and private sectors with international research excellence, InnovAIte Slovakia stands as a key instrument in addressing the social and economic challenges of today.

InnovAite
SLOVAKIA

Proin at molestie

Lorem ipsum dolor sit amet, consectetuer adipiscing elit sed diam nonummy nibh euismod tincidunt

www.innovaite.sk

Funded by the European Union
NextGenerationEU

RECOVERY AND RESILIENCE PLAN

DEPUTY PRIME MINISTER'S OFFICE FOR THE RECOVERY PLAN AND THE KNOWLEDGE ECONOMY

VAIA RESEARCH AND INNOVATION AUTHORITY