EDIH CTU

EDIH Czech Technical University in Prague

EDIH CTU

European Digital Innovation Hub in the Czech Republic in the field of Artificial Intelligence (AI) and Machine Learning (ML)

GRANT AGREEMENT NUMBER: 101083359

Deliverable D2.2

Data Management Plan

(version 2 – updated 31 October 2025)

Co-funded by the European Union

Funded by the European Union NextGenerationEU

CZECH RECOVERY PLAN

Inspire and make the Czech AI-driven Industry

www.edihctu.eu | www.edihcvut.cz

Project title	EDIH Czech Technical University in Prague (EDIH CTU)
Grant Agreement number	101083359
Funding scheme	Digital Europe Programme (DIGITAL) Call: DIGITAL-2021-EDIH-01 Topic: DIGITAL-2021-EDIH-INITIAL-01
Type of action	DIGITAL Simple Grants
Project duration	1 January 2023 – 31 December 2025 (36 months)
Project coordinator name	CTU - CESKE VYSOKE UCENI TECHNICKE V PRAZE
Deliverable number and title	**D2.2 Data Management Plan**
WP contributing to the deliverable	WP2 EDIH governance and operations
Deliverable type	DMP – Data Management Plan
Dissemination level	Public
Due submission date	30 April 2023 (Month 4), update 31 October 2025
Actual submission date	28 April 2023, update 31 October 2025
Deliverable Lead	Jan Wedlich
Contributor(s)	Martin Schano
	Tomáš Kejzlar
	Ondrej Smisek
	Petr Achs
Internal reviewer(s)	
Final approval	Barbora Zochova, Mikulas Cizmar

Status

This deliverable is subject to final acceptance by the EDIH Chairman and Business Development and Technology Transfer Manager.
This deliverable was approved on 28/04/2023 (V1), on 31/10/2025 (V2).

History of changes		
When	**Who**	**Comments**
26.04.2023	Martin Schano	The first draft of the document
26.04.2023	Barbora Zochova	Feedback from the PM
28.04.2023	Martin Schano	Version 1.0 release
30.10.2025	Petr Achs, Martin Schano	Revision of the document
30. 11. 2025	David Pešek	Revision of the document

Confidentiality	
Does this report contain **confidential** information?	Yes ☐ No ☑
Is the report **restricted** to a specific group?	Yes ☐ No ☑ *If yes, please precise the list of authorised recipients*:

Table of Contents

LIST OF ABBREVIATIONS AND ACRONYMS

BCP/DR	Business Continuity Planning / Disaster Recovery
DCAT-AP	Data Catalogue Vocabulary – Application Profile
DMP	Data Management Plan
DoA	Description of the Action
DOI	Digital Object Identifier
DPIA	Data Protection Impact Assessment
DUA	Data Use Agreement
EDIH CTU	European Digital Innovation Hub at the Czech Technical University in Prague
EPCIS	Electronic Product Code Information Services
ETL	Extract-Transform-Load
FHIR	Fast Healthcare Interoperability Resources
GDPR	General Data Protection Regulation
ISO	International Organization for Standardization
KPI	Key Performance Indicator
MFA	Multi-Factor Authentication
NGSI	Next Generation Service Interfaces
NKOD	National Open Data Catalogue
OData	Open Data Protocol
OSS	Open-Source Software
PID	Persistent Identifier
RAMI 4.0	Reference Architecture Model for Industry 4.0
RDF	Resource Description Framework
RO-Crate	Research Object Crate
RPO	Recovery Point Objective
RTO	Recovery Time Objective
SemVer	Semantic Versioning
SPDX	Software Package Data Exchange
TbI	Test before Invest
TRL	Technology Readiness Level
UN/CEFACT	United Nations Centre for Trade Facilitation and Electronic Business
WP	Work package

Executive Summary

The Data Management Plan (DMP) describes the recording and handling of specific research data and associated systems within the EDIH CTU project consortium members. Considering the nature and schedule of the EDIH CTU project, which aims to support the transfer of trusted solutions and services and whose nature ensures the gradual expansion of the services and solutions addressed, it is necessary to perceive this document as a framework for working with a heterogeneous set of data and other project outputs.

The DMP is explicitly linked to the project structure: it is connected to WP2 (T2.3 – Technology Transfer & IPR, T2.2 – Innovation Management) and to WP3–WP6.

This document describes the method of recording specific research data and other project outputs, evaluation criteria from the point of view of GDPR, cyber security, intellectual property and other relevant areas. Based on these facts, recommendations for publication, publication of anonymized or otherwise modified data, or non-publication and clarification of follow-up procedures in the event of these decisions follows.

1. Introduction

A Data Management Plan (DMP) outlines how specific research data are collected, managed, and shared throughout the project lifecycle. In the context of a European Digital Innovation Hub (EDIH), a DMP is essential for ensuring that data is managed effectively and securely throughout the innovation process.

A DMP for the EDIH CTU includes general information on the types of data that will be collected, how the data will be managed, who will be responsible for managing the data, how data quality will be ensured, and how data will be shared and preserved over time. It also addresses legal and ethical considerations related to data, such as data protection and privacy.

The development of the DMP for the EDIH CTU involves collaboration between various stakeholders, including researchers, data managers, and IT professionals. The DMP will be reviewed and updated throughout the innovation process, to ensure that it remains up-to-date and relevant.

This document helps minimizing the risk of data loss or breaches, ensures compliance with relevant regulations and ethical standards, and facilitates the sharing and reuse of data. It also contributes to the overall success of the EDIH CTU as a service provider, by enabling effective data-driven decision-making.

The DMP is a living document. Formal updates occur when necessary (e.g., launch of key services, legislative or security changes). The EDIH Office publishes changes in the internal repository. Designated persons at each partner are responsible for implementing the DMP at partner level.

2. Scope of Data Subject to the Data Management Plan

This chapter helps to identify which types of data sets are subject to this Data Management Plan in detail, and which just in short.

2.1 Purpose of Data Classification

Trigger-based application.

The DMP provides a framework that is activated when the project generates, collects, or needs to retain datasets or other outputs that meet the DMP-in-scope criteria (e.g., personal/sensitive data, external sharing, publication, long-term preservation, or reporting datasets with external exposure). It is not intended to impose full DMP procedures on all operational information handled during service delivery (e.g., routine internal engineering artefacts, transient logs, or client-provided data that remain under NDA and are not intended for external reuse), unless these are reclassified as in-scope.

This DMP applies selectively to datasets generated, collected, or processed within the project. Its purpose is to ensure that datasets which pose legal, ethical, security, or privacy risks are managed in compliance with applicable regulations and best practices, while avoiding unnecessary administrative burden for datasets that do not require such controls.

To this end, project data are classified into:

- Datasets subject to full DMP procedures (DMP-in-scope), and
- Datasets excluded from DMP procedures (DMP-out-of-scope).

This classification is based on the nature of the data, its intended use, and its potential impact if disclosed or misused.

2.2 Datasets Subject to Full DMP Procedures (DMP-in-Scope)

The following categories of data are strictly subject to the DMP and all associated procedures for storage, access control, documentation, retention, and (where applicable) sharing:

a) Personal Data

Any data that can directly or indirectly identify a natural person, including but not limited to:

- Names, contact details, identifiers, or account information
- Pseudonymised or linked datasets where re-identification is possible
- User interaction logs, behavioural data, or profiles associated with individuals

b) Sensitive or Special Category Data (if applicable)

- Data revealing racial or ethnic origin, health status, biometric data, or other protected attributes as defined by applicable legislation (e.g. GDPR).

c) Operational or Project-Critical Datasets with External Exposure

- Data shared with project partners, third parties, or external evaluators
- Data intended for publication, open access, or long-term archiving
- Data used for reporting, benchmarking, or scientific dissemination

All in-scope datasets must comply with the DMP requirements regarding:

- Lawful basis for processing
- Data minimisation
- Secure storage and access control
- Retention and deletion policies
- Documentation and traceability

2.3 Datasets Excluded from DMP Procedures (DMP-Out-of-Scope)

Client-provided data (default handling).

In most service deliveries, EDIH CTU processes data provided by the client either (i) under a Non-Disclosure Agreement (NDA) / contractual confidentiality, or (ii) as data that are already fully public. Client-provided confidential data remain under the client's control and are used strictly for the agreed purpose ("purpose limitation"); secondary use is not permitted unless explicitly agreed in writing. Such datasets are treated as DMP-out-of-scope by default (no cataloguing, PID, or publication planning), while applying appropriate security and retention as defined by the service contract/DUA. If client-provided data contain personal/special-category data, or if they are intended to be shared beyond the delivery team, they are classified as DMP-in-scope and the full DMP procedures apply.

The following categories of data are explicitly excluded from the scope of this DMP:

a) Internal Technical and System-Generated Data

Data used exclusively for the internal functioning, testing, or optimisation of project deliverables, including:

- Intermediate computation results
- Model weights, parameters, and internal representations
- Temporary logs generated for debugging or performance tuning
- Synthetic or randomly generated test data
- Non-Personal, Non-Shareable Engineering Artefacts
- Configuration files
- Build artefacts
- Internal system metadata
- Derived data that cannot be reverse-engineered to reveal personal or sensitive information

b) Data with No External Visibility or Reuse Intent Datasets that:

- Are not shared outside the development team
- Are not published or made accessible to users or third parties
- Are overwritten, aggregated, or discarded during system operation

The above-mentioned datasets are excluded because they:

- Do not contain personal or sensitive information
- Pose no privacy, ethical, or legal risk
- Have no standalone value outside the internal system context

As such, they do not require compliance with DMP procedures related to documentation, sharing, or long-term preservation.

2.4 Review and Reclassification

Data classification is subject to periodic review. If the purpose, content, or exposure of any dataset changes (e.g. internal data becomes externally shared or linked to identifiable individuals), the dataset will be reclassified and brought into scope of the DMP without delay.

3. Data Summary

This chapter includes the process of how to handle data generated or processed within the project. Each data set record includes a general description of the type of data (e.g., sensory data from production facilities, health record data, traffic movement and position data, etc.), the expected size of the data, the format used, the method of data collection (e.g., automated sensory collection, manual data entry, data collection from existing databases, etc.), the expected or measured data quality and the potential challenges associated with data collection and storage, including security and data protection.

The collection and storage of data itself can bring various challenges and risks, especially when it comes to sensitive or confidential information. Security and data protection are critical elements in data collection and storage. Data can be exposed to various threats such as cyberattacks, device loss or theft, misuse, and more. Specific strategies on how to protect data from such threats and how to minimize risks are developed in relation to a specific type of data.

Privacy is very important, especially if the data is used outside the project. It is important to determine who has access to personal data and how this data is stored and processed. Access should be limited to those in need and the technology used should be secure. It is also necessary to determine for what purposes the personal data will be used and whether these purposes comply with the laws and regulations relating to the protection of personal data. Where relevant, it may be important to take measures to minimise the risks associated with the use of personal data, such as network security, data encryption and data access monitoring. When working with personal data, it is important to respect the rights of the persons whose personal data are processed. These are, in particular, the right to information about data processing, the right to rectification and deletion of personal data and the right to restriction of processing. Violation of the Personal Data Protection Act can have serious consequences for businesses, including fines and the obligation to compensate injured persons.

The purpose of data generation or reuse applications for existing data in the context of the EDIH CTU project is for example the following:

- Benchmarking: Reusing industry-specific datasets can help in benchmarking the digital maturity of SMEs and identifying best practices.
- Identifying barriers and enablers: the project can identify common barriers and enablers to digitalization and develop targeted support services.
- Developing predictive models: Reusing datasets can enable the development of predictive models to forecast the impact of digitalization on SMEs, guiding strategic planning and decision-making.
- Improving product (service) design: Analysing existing data on the product (service) usage and feedback can inform product (service) development, identifying areas for improvement or new features that address customer pain points.
- Identifying skill gaps: Existing data on workforce skills and competencies can be reused to identify skill gaps in SMEs and design targeted training programs.
- Policy recommendations: Reusing data on the success and challenges of digital transformation initiatives can inform policy recommendations to support SMEs' digitalization journey.

- Monitoring and evaluation: Reusing data from previous digital transformation projects can provide a baseline for monitoring and evaluating the success of the EDIH CTU project and guide future improvements.

Some data may be useful to third parties. This potential is considered in this chapter, including potential entities or sectors that might be interested in the data. The data collected may be useful for other research projects in different fields that may use similar types of data. For example, sensor data from production facilities could be useful for research into the automation of industrial production. Alternatively, the data could be used for commercial purposes, such as the development of new products or services. For example, traffic movement and location data could be used to develop new navigation applications or to better optimise traffic flows. The data can also be useful for government purposes, such as better planning of urban transportation systems or monitoring health trends in the population. Last but not least, the collected data can be useful for non-profit organizations, such as for monitoring environmental trends or for helping in humanitarian crises.

The following Data description template needs to be filled in for each set of data included in the EDIH CTU project. The template is filled out by the person or institution managing the data, who is responsible for the correctness of the entry. The project manager is responsible for ensuring that the form is completed and properly filed.

3.1 Data description template

Each record of the data processed will contain the following information.

Service / Solution Name:

[Specify here within which service the solved data are processed.]

Unique identifier:

[Each dataset must be marked with a unique identifier for permanent traceability.; recommended pattern: EDIHCTU-<WP>-<SERVICE>-<YYYY>-<seq>.]

Persistent Identifier (PID):

[If applicable, the DOI/Handle assigned to the dataset or to its metadata record.]

Name of the data described:

[Provide a concise dataset title.]

General description of the type of data and the purpose of their collection:

[Here briefly describe the type of data and the purpose of its collection for the uninitiated reader. For example, it may be sufficient to simply inform that the data are data on the movement and location of selected means of transport.]

Method of data collection and processing:

[When describing the method of data collection, you can indicate, for example, whether it is automated sensory collection, indicate the frequency of data collection, etc.]

Expected data size and format:

[For example, if the format is changed during processing, please also describe it in this section. In the case of multiple data, it is important to include information on the size and format of the data that is passed on to other parties.]

Expected or measured data quality:

[This section can include any information about data quality, from assumptions through experience from working with data to specific outputs from quality control or other objective outputs.]

Provenance / source and processing:

[Summarize origins and key processing steps; link to a RO-Crate or equivalent if available.]

Version
[Use Semantic Versioning (e.g., 1.0.0). Describe what changed since the previous version.]

Data owner and contact person:

[Name of the data owner and contact person.]

Sensitivity / classification

[Choose: Public / Restricted / Confidential / Secret. Explain briefly.]

Relation to GDPR:

[Please indicate whether and, if so, to what extent data protection laws and regulations apply to such data.]

Relation to the protection of intellectual property, protection of classified information, trade secrets or protection by special laws:

[Indicate whether and, if so, to what extent such data are covered by the protection of intellectual property, the protection of classified information, trade secrets or protection by special laws.]

Licence (proposed / final):

[E.g., CC BY 4.0 / CC BY-SA / CC BY-NC / DUA only. If metadata only: CC BY 4.0.]

Access and security controls:

[Planned repository and access mode (open/restricted/closed), MFA/encryption requirements, allow-list, logging.]

Publication plan (data/metadata):

[Repository, target URL/DOI, embargo (if any), and expected publication date. If data cannot be published, state that only metadata ("tombstone" record) will be published with a contact point.]

Retention and disposal (archiving/erasure):

[Minimum retention period and conditions; what is archived (e.g., aggregates/models), what is erased (e.g., raw data) and when.]

Potential third parties / re-use scenarios :

[Indicate what possible third parties could use this data in case of disclosure.]

Other potentially relevant information:

[If applicable, please provide additional information not provided in the previous points.]

Recommendations for publication in the form of open data:

[Based on the above, in particular the protection of personal data and the protection of intellectual property, please provide a recommendation to publish in the form of open data, publish anonymized or otherwise modified data or not publish. Please also include a short justification.]

3.2 Dataset onboarding workflow

Dataset onboarding workflow applies to every new DMP-in-scope dataset and to any dataset that is later brought into scope through reclassification. For DMP-out-of-scope datasets, the project applies proportionate controls (contract/NDA terms, secure workspace, and agreed retention), but does not require full DMP documentation, cataloguing, PID assignment, or publication planning. Recommended steps:

1. Registration → create a record using this template (incl. WP/Task, Service, KPI).
2. Legal & ethics check → confirm GDPR legal basis, DPIA trigger/decision, IPR/licence intent.
3. Technical check → validate formats, metadata completeness (DCAT-AP + domain), and quality controls.
4. PID assignment → mint DOI/Handle (as applicable) and set the initial version.
5. Publication → publish data/metadata per the Publication plan; if data are restricted/closed, publish at least a metadata "tombstone" with DUA contact.

4. FAIR data

4.1 Legislative framework for data processing and accessibility

Data accessibility in the Czech legal environment is shaped by several legislative instruments that look either at which data should be public by default or, conversely, which data should be protected by default.

Public access to data is governed by Act No. 106/1999 Coll., on Free Access to Information, as amended to transpose Directive (EU) 2019/1024 on open data and the re-use of public sector information. This transposition introduced new rules on data openness and reuse across public institutions and publicly controlled entities. In this context, EDIH CTU follows the national open-data policy and publishes at least metadata in the National Open Data Catalogue (NKOD) wherever feasible.

The Czech amendment implementing Directive 2019/1024 EU also introduced the concept of a “public enterprise” (veřejný podnik), which extends certain transparency and data-sharing obligations to organisations that are publicly owned or controlled and perform activities of public interest – such as public transport operators, infrastructure managers, or public-service providers.

For projects like EDIH CTU, operating in areas such as logistics, mobility, and digital infrastructure, this means that datasets produced under publicly funded activities may, in some cases, fall under open-data obligations, unless they qualify for statutory exceptions (e.g. personal data, trade secrets, or security-sensitive information).

At minimum, metadata describing such datasets should be published in the NKOD or the institutional catalogue, and if direct publication is not possible, access may still be granted under a DUA defining the permitted use, safeguards, and audit rights.

Readers and data producers within the project are therefore advised to consult Act No. 106/1999 Coll. and Directive (EU) 2019/1024 when determining whether a dataset qualifies as open, restricted, or non-public.

In borderline cases – particularly in the logistics and transport domains where data may include operational or security-critical information – appropriate risk assessment, including anonymisation or pseudonymisation, should be conducted prior to any release.

With respect to the priority areas of data processing defined above, each dataset must be assessed to determine whether it falls into the category of (i) information for public access, (ii) open data, or (iii) one of the statutory exceptions where publication is not required. Such exceptions include, for example, industrial property, classified information, trade secrets, or data protected by special laws.

This framework is particularly relevant for the logistics and transport priority area: the information law newly introduces the concept of a “public enterprise,” which includes, inter alia, public transport service providers. In such cases, decisions about openness must balance transparency obligations with protection of sensitive operational and security information. For health-related and logistics datasets that may contain personal or safety-critical information, EDIH CTU applies anonymisation or pseudonymisation prior to any publication; only non-identifying data (or metadata alone) are released when risks cannot be adequately mitigated.

The protection of data and the information they contain is defined by several legal standards.

The basic and most general is the Civil Code (Act No. 89/2012 Coll.), which establishes the general protection of personality rights. Protection of personal data is governed in detail by Regulation (EU) 2016/679 (GDPR), including principles such as lawfulness, purpose limitation, data minimisation, storage limitation, integrity and confidentiality, and accountability. Where personal data are processed, the dataset record must state the legal basis (e.g., contract performance, legitimate interest, legal obligation, consent) and the DPIA status (required/not required/completed).

Personal data protection follows GDPR (EU 2016/679) and Act No. 110/2019 Coll. on the Processing of Personal Data; security measures reflect the Act on Cybersecurity (No. 181/2014 Coll.) and related NIS2 requirements. These references complement existing GDPR and civil code citations and align with DPIA/DUA processes.

The key legal regulation stipulating access to data arising from development, experimental research and innovation is Act No. 130/2002 Coll., on the Support of Research and Development from Public Funds. Section §12a addresses access to research data. In EDIH CTU, this translates into a default commitment to make research outputs as open as possible and as closed as necessary: open datasets with clear licences when legally and contractually feasible; otherwise, controlled access via DUA with appropriate safeguards.

Residual references and sector-specific regulations applicable to particular datasets (e.g., road transport, medical devices, cybersecurity) remain in force and must be cited in the dataset's "Legislative/Regulatory context" field within the template.

4.2 Making data findable, including provisions for metadata

Each dataset must be marked with a unique identifier for permanent traceability. It is advisable to use a naming convention for the identifier, thanks to which it will be possible to identify the dataset in general based on the identifier itself.

There are different standards for metadata, that is information that describes and provides context to other data in different contexts and areas, such as information science, digital librarianship, web services, archiving, and more. For the project, it will be useful to use the Resource Description Framework (RDF). RDF is a standard for describing metadata on the web. It uses a semantic model to describe the relationships between resources on the Web and enables interoperability between different systems and applications that use RDF.

The metadata will include keywords for a higher probability of finding the right result and repeatability.

The catalogue containing datasets will provide an application interface for interoperability with other systems, which will be able to harvest the catalogue automatically. At the same time, metadata will respect the established standards mentioned above.

It is proposed to use the following specific standards for the most common areas EDIH will deal with:

Industrial production by the principles of Industry 4.0

Industry 4.0 metadata standards focus on standardizing metadata related to digital technologies, automation and digitization of industrial processes. In particular, the following will be appropriate:

FIWARE NGSI (Next Generation Service Interfaces): It is a standard developed by the

European consortium FIWARE for data management in the context of the Internet of Things (IoT) and Industry 4.0. FIWARE NGSI defines a model for describing and exchanging metadata about entities such as sensors, devices, locations, and more, and provides an interface for their management and manipulation.

RAMI 4.0 (Reference Architecture Model for Industry 4.0): It is a model developed by the German company Plattform Industrie 4.0, which defines the reference architecture for Industry 4.0. RAMI 4.0 includes the definition of the layout and structure of metadata in industrial systems, including a description of physical and virtual objects, their interrelationships and functions.

ISO 8000: It is a standard developed by the International Organization for Standardization (ISO) for data management and data quality in various sectors, including industry. ISO 8000 provides a framework for metadata management, including the definition of metadata, the basic rules for its use, and the classification of metadata.

Public health

FHIR (Fast Healthcare Interoperability Resources): This is a modern standard developed by HL7 for the exchange of health information in electronic form. FHIR is based on web technologies and provides a standardized way to define, publish, exchange, and manage health data, including metadata related to the structure, content, security, and other aspects of health information.

Energetics

ISO 15926: This is a standard developed by the International Organization for Standardization (ISO) for describing and integrating data in industrial processes, including the energy industry. ISO 15926 provides a semantic model for describing physical and conceptual objects in various industries, including energy.

Another suitable standard for energy may also be the Open Data Protocol (OData): It is a standard developed by the OASIS for publishing and exchanging data using Web services. OData provides a protocol and model for describing and accessing data in energy systems, such as energy consumption data, production facilities, and more.

Logistics and transport

GS1 EPCIS (Electronic Product Code Information Services): This is a standard developed by GS1 for the definition and exchange of metadata related to the movement of goods and information about goods in the logistics chain. GS1 EPCIS allows the recording of events related to goods, such as arrival, dispatch, transfers, storage, etc., and provides structured metadata about these events, which can be used for tracking, tracing and analysing logistics operations.

UN/CEFACT (United Nations Centre for Trade Facilitation and Electronic Business) XML: This is a set of standards developed by the United Nations (UN) for electronic data interchange in the field of international trade and logistics. UN/CEFACT XML defines formats and structures for exchanging data related to logistics processes, such as information on inventory, transport, mail, customs declarations, etc., and provides standardized metadata for these processes.

ISO 28000 (Specification for security management systems for the supply chain): This is an international standard developed by the International Organization for Standardization (ISO) for security management in the logistics chain. ISO 28000 provides a framework for the

definition, implementation and evaluation of security measures in logistics processes, and also defines the requirements for metadata management related to the safety and protection of inventory, transport, storage, etc.

EDIH CTU ensures that datasets and other research outputs are findable by adhering to a consistent metadata, identification, and naming strategy:

1. Metadata standards and cataloguing
    - EDIH CTU maintains a lightweight internal register for DMP-in-scope datasets and selected outputs that are intended for external sharing, reporting with external exposure, or long-term preservation. Where openness is feasible, a public-facing metadata record (DCAT-AP) may be created and, where applicable, harvested to the National Open Data Catalogue (NKOD). DMP-out-of-scope operational artefacts are not catalogued. Each catalogue record follows DCAT-AP (Data Catalogue Vocabulary – Application Profile) to support interoperability and automated harvesting to the National Open Data Catalogue (NKOD).
    - Where applicable, experimental bundles (data + code + documentation) are packaged using RO-Crate, and the catalogue record points to the crate landing page.
    - Domain-specific metadata (e.g., NGSI/FHIR/EPCIS/OData) may be included in addition to DCAT-AP; mandatory elements are mapped into the DCAT-AP profile.
2. Persistent identifiers (PIDs)
    - Each public dataset (or, if needed, its metadata-only record) receives a PID, preferably a DOI minted via the institutional repository or Zenodo.
    - The PID and the dataset version (Semantic Versioning, e.g., 1.2.0) are recorded in the catalogue to enable precise citation and traceability.
3. Naming convention
    - To improve human and machine findability, dataset IDs follow the convention:
      EDIHCTU-<WP>-<SERVICE>-<YYYY>-<seq>
      Examples: EDIHCTU-WP3-TbI-2025-001, EDIHCTU-WP5-Training-2025-014.
    - The ID is included in filenames, repository records, and citations.
4. Keywords and discoverability
    - Catalogue entries include controlled keywords aligned with the project taxonomy (services, domains) and free-text tags to support search.
    - Each entry links to related work packages/tasks, services, and KPIs so users can navigate from objectives to the underlying data.
5. Landing pages and cross-links
    - Every PID resolves to a landing page with rich metadata (title, abstract, creators, version, licence, access conditions, related outputs).
    - If data cannot be shared openly, a tombstone record is still published with contact details and conditions for access (e.g., DUA).

6. Operational practice
    - New or updated records are validated against the DCAT-AP profile during onboarding. Where NKOD harvesting is applicable, metadata publication and any automated transfer will be scheduled case-by-case (not for all datasets), based on the publication plan and legal/contractual constraints.
    - The catalogue maintains change history so users can find previous versions and read a changelog.
7. Citation
    - Recommended citation format includes authors, year, title, version, PID (DOI), and licence to facilitate reuse and attribution.

4.3 Making data accessible

Repository

A trusted data store must meet several important requirements to ensure data security, privacy, availability, and integrity. Some of these requirements are:

1. Data security: A trusted data store must have adequate data security, including technical, organizational and physical measures to protect the data from unauthorized access, loss, theft or damage. This includes, for example, data encryption, user authentication and authorization, data access monitoring, and data backup.
2. Privacy and data protection: A trusted data store must protect privacy and protect personal data by applicable laws and regulations, such as the General Data Protection Regulation (GDPR) in the European Union. This includes, for example, anonymizing or pseudonymizing data, restricting access to personal data to authorized users only, and ensuring compliance with applicable data protection laws and regulations. This issue is described in more detail in the chapter Legislative Framework.
3. Availability and continuity of services: A trusted data store must ensure the availability and continuity of services so that data is available to the extent and at the appropriate time. This includes, for example, redundant architecture, data backup, monitoring service availability and performance, and scheduling backup solutions for recovery in the event of an outage or disaster.
4. Data integrity: A trusted data store must ensure data integrity, which means that data remains unchanged and is not tampered with or corrupted. This includes, for example, data integrity checks, data backup and data integrity verification at storage and processing.

Access model and repository decision

EDIH CTU applies a simple decision model to determine how each dataset is made accessible:

Access class	When to use	Repository / delivery	Licence & conditions

Open	No legal/contractual restrictions; public value maximised	Zenodo or CTU institutional repository with DOI	C BY 4.0 (preferred) or CC BY-SA; metadata under CC BY 4.0 **C**
Restricted	Legal/contractual/safety constraints prevent full openness, but controlled sharing is possible	Controlled repository/secure workspace with access management	Access via DUA (purpose limitation, security minima), full audit log of access
Closed (metadata-only)	Data cannot be shared (e.g., trade secrets, sensitive personal/safety-critical data)	No data release; metadata "tombstone" only	N/A for data; metadata under CC BY 4.0

Healthcare and transport/logistics data. Where special regulation applies, anonymisation or pseudonymisation is performed before any publication. If such steps would invalidate the content or risk cannot be mitigated, the dataset is Restricted or Closed (metadata-only).

Publication and harvesting

- Catalogue → NKOD. All public catalogue records follow DCAT-AP and are harvested to the National Open Data Catalogue (NKOD) by an automated job (cron) once per week.
- PID/DOI. Open datasets (and, where appropriate, metadata-only records) receive a PID (preferably DOI) to ensure citability and long-term resolvability.

Access channels

- Open data. Discoverable via the project catalogue and NKOD, downloadable over HTTPS and/or API as provided by the repository.
- Other public datasets. Served through the catalogue's interfaces (REST API, HTTPS).
- Restricted data. Accessed upon request under a DUA, with identity verification and logged access.

Catalogue analytics and feedback

The catalogue provides basic anonymised usage analytics (e.g., views/downloads per dataset) and enables voluntary user feedback for open datasets to improve quality and re-use.

Metadata

The metadata will be publicly available under a free Creative Commons (CC BY 4.0) license. By default, the lifetime and availability of metadata should exceed the lifetime and availability of the datasets themselves.

Data will be provided in accordance with Act No. 130/2002 Coll., on the Support of Research and Development from Public Funds and on the Amendment of Some Related Acts (the Research and Development Support Act).

The catalogue will not provide access to software that allows reading and processing datasets.

4.4 Making data interoperable

4.4.1 Formats and frameworks

At a minimum, the following formats, frameworks and methodologies will be used to ensure the highest possible interoperability for exchange and reuse within and across areas:

Formats:

JSON (JavaScript Object Notation), XML (eXtensible Markup Language), CSV (Comma-Separated Values) and generally other open formats that have publicly and freely available documentation.

The EDIH CTU catalogue records may also follow DCAT-AP. When datasets originate from domain standards (e.g. NGSI, FHIR, EPCIS 2.0, OData).

Frameworks:

FAIR (Findable, Accessible, Interoperable, Reusable) provides principles and guidelines for creating interoperable data across different regions.

4.4.2 Methodology

Extract-Transform-Load (ETL) and Linked Data provide techniques for linking data from heterogeneous sources. The ETL process involves three steps:

- Extract: Extract data from a variety of sources, such as databases, files, web services, or other data sources. The extracted data is usually converted into a format suitable for further processing.
- Transform: Transform the extracted data into the desired format or structure. This may include data cleaning, normalization, merging data from different sources, calculations, and other data modifications.
- Load: Load the transformed data into the target system or storage, which can be databases, data warehouses, or other data storage and processing systems.

The ETL process is often used to automate the transfer, transformation, and retrieval of data from different sources into target systems, allowing data from different sources to be integrated and analysed, thereby facilitating decision-making and retrieval of information from data sources. ETL is widely used in areas such as business intelligence, data warehousing, data analysis and other areas where there is a need to integrate and analyse large amounts of data from various sources.

Linked Data is a way of organizing and publishing data on the Web that focuses on standardized linking and joining of data using identifiers (URIs) and links (links). Linked Data is based on a set of principles and techniques that have been designed to achieve interoperability and connectivity of data on a global scale. Linked Data uses open standards such as the Resource Description Framework (RDF) to represent data in a machine-readable format and the Uniform Resource Identifier (URI) to uniquely identify data sources on the Web. Data published as Linked Data is enriched with links to other related data and sources that allow them to be interconnected with other data sources and thus create semantic networks of interconnected data.

The standards used are described in the metadata chapter.

In the case where specific and not quite common ontologies and dictionaries are used, a mapping to more common ontologies will be created. The created ontologies and dictionaries will be publicly available for further use, improvement and expansion.

For the Czech implementation, DCAT-AP-CZ, the data model on the "Open data" website serves as the primary reference.

Mandatory elements for a dataset (dcat:Dataset)

- **Title** (dct:title): A human-readable name for the dataset.
- **Description** (dct:description): A textual description of the dataset's content.
- **Publisher** (dct:publisher): An identifier for the organization publishing the dataset (in the Czech context, typically a URI from the Register of Rights and Obligations).
- **Contact Point** (dcat:contactPoint): Contact details for a person or organization (e.g., email, name).
- **Category/Theme** (dcat:theme): An identifier for the topic (e.g., from the MDR Data Themes classification).
- **Distribution** (dcat:distribution): A link to at least one distribution.

Mandatory elements for a distribution (dcat:Distribution)

- **Access/Download URL** (dcat:accessURL or dcat:downloadURL): The URL where the distribution can be accessed or downloaded.
- **Media Type** (dcat:mediaType): The standard identifier for the file format (e.g., text/csv).

Mandatory elements for a catalog (dcat:Catalog)

- **Title** (dct:title): The name of the catalogue.
- **Description** (dct:description): A textual description of the catalogue.
- **Publisher** (dct:publisher): The publisher of the catalogue.
- **Dataset** (dcat:dataset): A link to the datasets contained within the catalogue.

Transfer to a mapping table

A mapping table for DCAT-AP should include columns that link the original schema (e.g., from an internal system) to the target DCAT-AP elements. For each mandatory element, the table should contain:

1. **Source Element:** The field name in the original system.
2. **Target Element:** The name of the DCAT-AP element (e.g., dct:title).
3. **Mandatory/Required:** An indicator of whether the element is mandatory (e.g., M = Mandatory).
4. **Mapping Rule:** How the value from the source field is transformed for DCAT-AP.

When creating the mapping table, it's also advisable to include recommended elements, which improve the quality of the metadata. Many DCAT-AP standards distinguish between mandatory and recommended elements, which should be provided if available.

4.5 Increase data re-use

For validation of data analyses and to facilitate the reuse of data, documentation will be available for the datasets, which will contain at least:

- Description of the analysis methodology
- Description of assets
- Description of data transformation and cleaning
- Description of data pre-processing
- Analysis outputs
- Code and overview of used tools
- Validation methods

The data will be available in the public domain to allow for the widest possible use. The data will be provided under a license allowing their reuse, even by third parties and after the end of the project.

The metadata will also include information about the origin of the data.

The guarantee of data quality will be ensured by standard tools of the catalogue following the example of NKOD. The basic tool will be appropriately structured metadata informing, for example, about the responsible person or entity, the frequency of data updates, the last change, etc. The structure of descriptive metadata is described in more detail in the introductory chapter.

4.5.1 Licensing guide

Default for metadata: CC BY 4.0.

Recommended for data (choose the most open option permitted by law/contract):

- CC BY 4.0 (preferred; attribution only), or
- CC BY-SA 4.0 (share-alike), or
- CC BY-NC 4.0 (non-commercial) when strictly required.

When openness is not possible: provide metadata-only and enable controlled access via DUA defining purpose limitation, security minima, retention, audit, and breach handling.

Record the chosen licence (and, where relevant, DUA URL) in the catalogue record (dct:license, dct:rights), and include machine-readable licence identifiers (e.g., SPDX/Creative Commons URLs).

4.5.2 Citation and versioning

Every public dataset (or metadata-only record) must have a PID (DOI preferred). Use Semantic Versioning (SemVer) for datasets and derived artefacts (models, code): MAJOR.MINOR.PATCH. Provide a CHANGELOG (summary of changes) and link it from the landing page.

Recommended citation format (also placed on the landing page and in dct:bibliographicCitation):

Author(s)/Organisation (Year). Dataset title (Version X.Y.Z). DOI: 10.xxxx/xxxxx. Licence: CC BY 4.0.

Example: *EDIH CTU* (2025). TbI vibration signals for robot cell QA (v1.2.0). DOI: 10.5281/zenodo.9999999. Licence: CC BY 4.0.

4.5.3 Re-use enablement

Provide clear readme/"How to use" notes, sample code or notebooks (where feasible), and machine-readable schemas. Ensure interoperable formats (open standards) and complete DCAT-AP records with keywords, temporal/spatial coverage, and dct:conformsTo for domain standards. For restricted data, publish a tombstone record with a contact point and DUA request process.

4.5.4 Metrics and KPI linkage

To evidence impact and guide improvements, track and report the following re-use metrics (catalogue or repository analytics; aggregated, anonymised where applicable):

- Discoverability: catalogue views, NKOD referrals.
- Access/Downloads: unique downloads/API calls by distribution.
- Citations: DOI resolutions and scholarly/industry citations (via DOI and manual curation).
- External re-use signals: forks/stars of companion code, references in case studies/standards, industry integrations.
- Quality feedback: user ratings/comments, issue reports resolved.

5. Other research outputs

Beyond datasets, EDIH CTU may generate software, machine-learning models, data/ML pipelines, and digital twins. As a default, technology outputs developed by EDIH CTU (software, models, pipelines, documentation) remain the property of CTU or the respective partner that created them, and the client receives a non-exclusive, royalty-free licence for use for the agreed purpose. Where a technology output is developed or substantially customised specifically for a client under a contract that transfers rights, the client receives the agreed rights up to full ownership as defined in the service agreement. The applicable IPR and licence terms are recorded in Template B (and linked from any related dataset record in Template A where relevant).

For models, provide a brief Model Card (intended use, limitations, risks), a Data Card for the training/validation data (with dataset DOI), and an evaluation summary with key metrics and thresholds. Pipelines should describe orchestration (e.g., DAG/graph), idempotency, and lineage; container images and exact dependency locks must be referenced from the release.

Access follows the project's three classes: Open (public repo + archived DOI, metadata under CC BY 4.0), Restricted (private distribution under DUA with audit logging and MFA), and Closed (metadata-only) where disclosure is legally or contractually impossible. All catalogue records include dct:license, dct:conformsTo (standards/API), owl:versionInfo, and the PID/DOI; releases expose a CITATION.cff with the recommended citation (Author/Org, Title, vX.Y.Z, DOI, Licence).

Quality assurance is proportionate: software includes basic tests and code/dependency scanning; models report evaluation against stated datasets/versions; pipelines document validation and recovery behaviour. Outputs are screened for export-control/dual-use concerns and, where applicable, are classified as Restricted/Closed with access via DUA. Privacy risks for model memorisation are noted with any mitigations applied.

6. Data security

EDIH CTU protects data throughout their lifecycle with controls proportionate to sensitivity. All transfers use TLS, all repositories maintain access logs, and personal or otherwise sensitive data are minimised, pseudonymised or anonymised where appropriate. We classify datasets into Public, Restricted, and Confidential, and apply escalating safeguards:

- Public data are served over TLS with audit-ready access logging.
- Restricted data add multi-factor authentication (MFA), encryption at rest, role-based access, and are shared only under a DUA defining purpose, retention and security minima.
- Confidential data inherit all Restricted controls and additionally enforce IP allow-lists and periodic access reviews (at least monthly), with least-privilege by default.

Keys and credentials are rotated regularly; repositories and services are patched on a defined cadence; backup and logging systems are segregated from production. For third-party or client-provided data (e.g., TbI), isolated workspaces are used, outbound transfer is restricted, and raw inputs are erased according to the dataset's retention plan once deliverables are accepted.

Incident management follows a simple playbook: detect → contain → eradicate → recover → post-mortem. Suspected incidents are reported within 72 hours, triaged by the security lead and the EDIH Data Manager, tracked in the ticketing system, and closed with documented corrective actions and customer notification where required.

Business continuity and disaster recovery (BCP/DR) targets are RPO ≤ 24 hours and RTO ≤ 48 hours. We maintain 3–2–1 backups (three copies, two media, one off-site) with quarterly restore tests; critical services have run-books, designated alternates, and contact trees. Security and access controls are reviewed at least quarterly (and after any incident), and results feed back into the DMP change log.

7. Ethics

There may be several ethical and legal issues that may impact data sharing. Common problems that can occur include, for example:

- Privacy concerns: Sharing personal data or sensitive information may raise privacy concerns and there may be legal or ethical restrictions on how such data is shared or used. When sharing personal data, it is important to ensure proper consent and anonymization measures.
- Intellectual property rights: Sharing data may also include sharing intellectual property rights, such as copyrights, patents or trademarks. It is important to ensure that any intellectual property rights related to the data are properly identified and managed when data is shared.
- Confidentiality and security: Confidentiality or security concerns may arise when sharing data, especially if the data relates to national security, trade secrets or other sensitive information. It is important to ensure that appropriate security measures are in place to protect data from unauthorised access or disclosure.

EDIH CTU applies “as open as possible, as closed as necessary” with clear accountability for people and data. Ethical compliance is embedded in service delivery (especially **Test before Invest**, WP3) and in Open Calls: before any collection or processing, we confirm lawful basis, necessity and proportionality, risks to individuals, and safeguards (privacy-by-design, security, transparency).

Roles and accountability

For each service or activity, the dataset record explicitly states who is **Controller** and who is **Processor**. As a rule of thumb:

- in TbI engagements, the client is Controller and the relevant EDIH partner acts as Processor;
- in training/events and EDIH-run studies, EDIH (or the named partner) is Controller;
- in Open Calls, the applicant is Controller for their submitted data; the EDIH partner delivering the service is Processor. Any joint controllership or sub-processing is documented, including contacts and contract references.

DPIA triggers and decisions

DPIA is performed when triggers apply—e.g., processing of special-category data; large-scale or systematic monitoring; profiling with significant effects; combining data from multiple sources; AI models with potential impact on individuals (recognition, scoring, diagnostics). Where a DPIA is not required, the record includes a brief rationale; when required, the DPIA is completed before processing and revisited upon material change.

Informed consent and transparency

Where consent is the lawful basis (e.g., marketing communications; publication of identifiable photos/videos from events), we use concise short-form notices at the point of collection, backed by a long-form notice accessible via link/QR. Notices explain purpose, scope, retention, recipients, international transfers, contact points, and withdrawal options. Consent is specific, granular, freely given, and logged; alternative bases (contract, legitimate interest,

legal obligation) are used where more appropriate, with the balancing test recorded for legitimate interest.

Data minimisation and safeguards

Only data necessary for the stated purpose are collected. Personal data are pseudonymised or anonymised whenever feasible—especially for health-related or safety-critical logistics/transport contexts—before any sharing or publication. Access is role-based; Restricted/Confidential classes follow the security measures defined in Chapter 5. De-identification limits and re-identification risks are considered in DPIA and documented.

Documentation and evidence

Partners must upload the following artefacts to the dataset's repository (or linked record) and keep them in sync with the catalogue entry: (i) DPIA (if applicable) or DPIA-not-required note, (ii) consent materials (short/long form), (iii) lawful-basis statement, (iv) data-sharing terms (e.g., DUA), and (v) retention/erasure plan. The catalogue record links these artefacts and states the DPIA status and legal basis.

Special cases

- **TbI projects:** isolate client data, restrict secondary use, erase raw inputs after deliverable acceptance per retention plan, and share only aggregates/models permitted by contract. Publication, if any, is metadata-only ("tombstone") with a DUA contact.
- **Open Calls:** applications include data-ethics and GDPR sections (lawful basis, DPIA trigger outcome, intended sharing/licence). Processing starts only after these checks are complete.

Oversight and review

The EDIH Data Manager coordinates ethical compliance with partner Stewards, ensures DPIA quality, and tracks corrective actions. Ethics controls are reviewed at least quarterly or after any incident or material change; outcomes feed into the DMP change log.

8. Conclusion

In conclusion, the DMP describes the recording and handling of a heterogeneous set of data and associated outputs within the EDIH CTU project. It sets out the areas to be considered, how they affect processed data, and who is responsible for recording, safeguarding, and deciding on publication.

Given that EDIH CTU aims to create a library of trusted solutions and services in a changing environment, the DMP is designed to operate across diverse data types and situations. Principles are formulated to be specific enough for operational use yet adaptable as services mature and information becomes available.

Updates and reviews

The DMP is a living document: it is reviewed and updated at key milestones aligned with WP1/WP2 and, post-project, undergoes an annual re-audit. Each update is recorded in a public change log, with roles and responsibilities (Data Manager/Stewards) accountable for implementing agreed improvements.

This approach ensures continuity beyond the project, keeps the catalogue reliable and reusable, and aligns openness with legal, ethical, and security obligations.

Annexes – New/updated templates and procedures

Template A – Data description (extended)

Service / Solution / WP / Task; KPI linkage; Unique ID (e.g. EDIHCTU-<WP>-<SERVICE>-<YYYY>-<SEQ>); DMP applicability status (Out-of-scope / In-scope / Pending – treat as in-scope until confirmed); Classification decision date; Classified by (name/role); Classification rationale (reference to contract/DUA/GDPR assessment); Title & version; Purpose; Provenance (RO-Crate link); ETL and frequency; Size & formats; Quality (QC metrics); Owner/Steward (Data Steward); GDPR (legal basis, categories, DPIA status, pseudonymisation/anonymisation); IPR & restrictions; Security class & controls; Metadata profile (DCAT-AP + domain); Keywords; PID; Licence; Publication (repository, date, embargo); Disposal/archiving; Re-use scenarios; Notes.

Template B – Other outputs (software/models)

Type; Repo & release (Git*, Zenodo DOI); Licence & third-party components (SPDX); Model/Data Card; Evaluation metrics; Risks/bias & mitigation; Export/dual-use restrictions; Links to related dataset(s).

Template C – DUA (skeleton)

Parties; Data description; Purpose; Term; Permitted and prohibited uses; Minimum security measures; Audit & logging; Incidents & notifications; Sanctions; Termination & deletion; Governing law & jurisdiction.

Template D – DPIA trigger list

Personal/special-category data; Profiling; Large-scale systematic monitoring; Combining data across sources; AI training/validation with impacts on individuals. Outcome: DPIA required / not required + justification.

Checklist – Dataset onboarding

[ ] Template A completed

[ ] GDPR legal basis and DPIA status

[ ] Classification and security controls

[ ] DCAT-AP + domain metadata

[ ] PID assigned

[ ] Licence/DUA in place

[ ] Publication executed/planned

[ ] KPI linkage recorded

[ ] Backups & disposal plan

Project-specific data flows

The following data flows illustrate typical situations relevant for DMP purposes; the full DMP procedures apply only when the processed data meet the DMP-in-scope criteria defined in Section 2.

1. **Contact/enquiry:** identification and contact details, description of need → **purpose:** presales and service qualification → **legal basis:** negotiation of a contract / legitimate interest → **retention:** 12 months from last interaction, then delete/anonymise.
2. **Request forms (TbI, ADS, IEN, SFI):** company identification, description of data/technology, TRL, de minimis, and possibly personal data of contact persons → **purpose:** service design and delivery → **legal basis:** contract performance → **retention:** minimum period per donor rules + 5 years of archiving.
3. **TbI experiments:** technical raw data streams (IoT, PLC, images), derived feature sets, models, reports → **legal basis:** contract performance; if personal data are involved, DPIA and pseudonymisation/anonymisation are required → **retention:** per contract/DUA; by default, delete raw data within 90 days after acceptance of the deliverable, keep only aggregates/models if the contract permits.
4. **Training and events:** registrations, attendance, evaluation, photos/video → **legal basis:** contract performance (participation), legitimate interest for documentation; marketing only on the basis of consent → **retention:** 24 months (operational); selected outputs archived per the communications plan.
5. **Publicity/Case studies:** public summaries of benefits and recommendations → **legal basis:** legal obligation/legitimate interest per the communications plan and grant rules; legal and IPR review required before publication.
6. **Reporting to DTA and donors:** progress and impact indicators → **legal basis:** legal obligation; data minimisation and aggregation; publish aggregated open statistics (without personal data).