# 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

Co-funded by
the European Union

*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)<br>Call: DIGITAL-2021-EDIH-01<br>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) |
| **Actual submission date** | 28 April 2023 |
| **Deliverable Lead** | Petr Achs |
| **Contributor(s)** | Martin Schano |
| | Tomáš Kejzlar |
| | Ondrej Smisek |
| **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.

| 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 |
| | | |
| | | |
| | | |
| | | |
| | | |

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

# Table of Contents

LIST OF ABBREVIATIONS AND ACRONYMS ..... 5
Executive Summary ..... 6
1. Introduction ..... 7
2. Data Summary ..... 8
2.1 Data description template ..... 9
3. FAIR data ..... 11
3.1 Legislative framework for data processing and accessibility ..... 11
3.2 Making data findable, including provisions for metadata ..... 11
3.3 Making data accessible ..... 13
3.4 Making data interoperable ..... 14
3.5 Increase data re-use ..... 15
4. Other research outputs ..... 16
4.1 Other project outputs template ..... 16
5. Data security ..... 18
6. Ethics ..... 18
7. Conclusion ..... 20

# LIST OF ABBREVIATIONS AND ACRONYMS

| | |
|---|---|
| DMP | Data Management Plan |
| EDIH CTU | European Digital Innovation Hub at the Czech Technical University in Prague |
| EPCIS | Electronic Product Code Information Services |
| FHIR | Fast Healthcare Interoperability Resources |
| GDPR | General Data Protection Regulation |
| ISO | International Organization for Standardization |
| NGSI | Next Generation Service Interfaces |
| NKOD | National Open Data Catalogue |
| RAMI 4.0 | Reference Architecture Model for Industry 4.0 |
| RDF | Resource Description Framework |
| UN/CEFACT | United Nations Centre for Trade Facilitation and Electronic Business |

# Executive Summary

The Data Management Plan (DMP) describes the recording and handling of 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 document describes the method of recording 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 data will be 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 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 regularly throughout the innovation process, to ensure that it remains up-to-date and relevant.

This document will help to minimize the risk of data loss or breaches, ensure compliance with relevant regulations and ethical standards, and facilitate the sharing and reuse of data. It will also contribute to the overall success of the EDIH CTU as a service provider, by enabling effective data-driven decision-making.

# 2. 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.

## 2.1 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.]*

**Name of the data described:**

*[Name the dates briefly and concisely here.]*

**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.]*

**Data owner and contact person:**

*[Name of the data owner and contact person.]*

**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.]*

**Potential third parties for whom the data may be relevant:**

*[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. FAIR data

## 3.1 Legislative framework for data processing and accessibility

Data accessibility in the Czech legal environment defines several basic legislative forms that view this accessibility either from the perspective of which data should be public in principle or, conversely, which should be protected in principle.

Public access to data is stipulated by Act No. 106/1999 Coll., on Free Access to Information, the current wording of which is based on the transposition of European Directive 2019/1024.

Concerning the four priority areas of data processing defined above, it will always be necessary to assess whether the processed data fall into the category of data intended for public access, open data or one of the exceptions defined by law for situations where data do not have to be published. These exceptions are, for example, the protection of industrial property, the protection of classified information, trade secrets or data protected by special laws.

This law will be particularly important for the fourth priority area - logistics and transport, as the information law newly introduces the term "public enterprise", which includes, inter alia, providers of public transport services.

The protection of data and the information contained therein is defined by several legislative standards. The basic and most general is Act No. 82/2012 Coll., the Civil Code, which generally sets out the principle of protection of personality. The protection of data containing personal data is determined in more detail by Regulation No. 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and the free movement of such data.

The key legal regulation stipulating access to data from development, experimental research and innovation is 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). Access to research data is described in §12a.

## 3.2 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.

## 3.3 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.

**Data**

Primarily, most data should be available publicly, ideally in the form of open data, as defined by the relevant legislation. The data published in this way will be traceable not only in the project catalogue itself but also in the National Open Data Catalogue (NKOD). For this purpose, the catalogue must have an interface created for harvesting by NKOD.

Some data may be subject to special regulation, in particular data from the healthcare sector. In this case, it will be necessary to anonymise or pseudonymize the data before they can be published. If such steps would invalidate the content of the data, it will be necessary to restrict access to this data.

Data from the open data category will in principle be accessible in a standard form via NKOD. Other publicly published data through the described interfaces that the data catalogue will provide. In particular, it is an application interface (API) and a standard HTTPS protocol.

For publicly accessible data, these will be freely available even after the end of the project implementation.

The catalogue will be equipped with a tool for basic anonymized analytics regarding the utilization rate of individual datasets. At the same time, it will allow feedback from users, which, however, will be primarily voluntary for freely available datasets.

**Metadata**

The metadata will be publicly available under a free Creative Commons 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.

## 3.4 Making data interoperable

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.

**Frameworks:**

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

**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.

## 3.5 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. Other research outputs

The project itself can generate not only data but also other outputs that can be used in various areas. These outputs can be digital or physical in nature. Digital outputs include, for example, software, workflow, protocol, model, etc. Physical outputs may include new materials, antibodies, samples, etc.

In addition to data management, providers of individual services and solutions must consider whether these other project outputs can also be provided to third parties in accordance with the FAIR (Findable, Accessible, Interoperable, Reusable) principles and further evaluate to what extent these other outputs need to be protected.

When digital outputs are opened, it is important to ensure that they are available and usable for other researchers and practitioners in the field. This can be done, for example, by releasing open-source code, sharing online libraries, or providing access to software.

Physical outputs, such as new materials or samples, should also be correctly labelled and stored so that they can be easily found and reused. If these outputs are sensitive, privacy and/or intellectual property protection measures should be considered when storing and sharing them. Alternatively, these outputs should also appear in the inventory of other project outputs and be made available to third parties.

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

## 4.1 Other project outputs template

*Each record of project outputs will contain the following information.*

**Service / Solution Name:**

*[Specify here which service or solution this output is processed in.]*

**Unique identifier:**

*[Each solution must be marked with a unique identifier for permanent traceability.]*

**Name of the described output:**

*[Briefly and concisely name the output here.]*

**General description of the type of output and its use in the project:**

*[Here briefly describe the type of output and the purpose of its creation for the uninitiated reader. Common outputs can be described briefly. Describe new or unknown outputs in more detail.]*

**Description of the output including relevant parameters:**

*[In the case of physical output, indicate its details such as shape, material, structure, important*

*dimensions, etc., and in the case of software, the protocols used, solution architecture, computing requirements, etc.]*

**Standards followed in the creation of output:**

*[This section can include any information, for example, on compliance with technical standards or other methodological materials, but also in the case of the relevance of internal procedures such as the degree of quality control, etc.]*

**Contact person:**

*[Insert contact person.]*

**Relation to GDPR:**

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

**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 output is covered by the protection of intellectual property, protection of classified information, trade secrets or protection by special laws.]*

**Potential third parties for whom the data may be relevant:**

*[Indicate what potential third parties could use this output if published.]*

**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, recommend to publish, publish in a modified form or not to publish. Please also include a short justification.]*

# 5. Data security

To ensure data security, several measures can be implemented for data recovery, secure storage and archiving, and transfer of sensitive data. Some examples are:

1. Data backup and recovery plan: A data backup and recovery plan should be implemented to ensure that data is stored securely and quickly recovered in the event of data loss or corruption. This plan should include regular backups of data in multiple locations, testing of data recovery processes, and secure storage of backup copies.

2. Access control and authentication: Access and authentication controls should be put in place to restrict access to sensitive data to authorised personnel only. This can include the use of passwords, two-factor authentication, and role-based access control.

3. Data encryption: Sensitive data should be encrypted to prevent unauthorized access and ensure data confidentiality. This may include the use of encryption algorithms to encrypt data at rest and in transit.

4. Secure data transfer: Sensitive data should be transmitted securely to prevent it from being intercepted or tampered with. This may include using secure protocols, such as SSL/TLS or SSH, for data transmission.

5. Physical security: Physical security measures should be in place to protect the physical storage media on which data is stored. They may include secure storage devices with controlled access and environment control to ensure data integrity.

6. Data retention policy: A data retention policy should be established to ensure that data is retained for the required period and safely destroyed when it is no longer needed.

Overall, a comprehensive data security plan should be developed and implemented to ensure that data is securely stored, protected and transmitted throughout its lifecycle. This plan should be regularly reviewed and updated to ensure that it remains effective and in line with the latest safety best practices.

Long-term data storage is a complex and not always easy-to-solve issue. Trusted repositories can be institutional or proprietary and should follow established standards and best practices for data management, storage, and curation. These standards may include the use of metadata standards, data formats, and protocols for data transmission, storage, and backup. In addition, trusted repositories should have a sustainability plan that ensures that data can be retained and maintained over time, also taking into account the development of technology and storage media.

It is important to ensure that a trusted repository is accredited and meets the required standards and regulations for the specific data stored. In addition, appropriate security measures should be implemented to protect data against unauthorised access or loss, and procedures should be established for the recovery of data in the event of loss or damage.

Overall, the use of trusted repositories for long-term preservation and curation can provide assurance that data will be retained and made available for future research and that it will be protected and managed in accordance with established standards and best practices.

# 6. 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.

To address these issues, it is important to conduct an ethical review as part of the project planning process. The ethics review should assess potential ethical and legal issues related to data sharing and develop appropriate measures to address them. This may include developing informed consent procedures, anonymizing data, implementing security measures, or obtaining legal advice to ensure compliance with applicable laws and regulations.

In the context of a research project, ethical outputs and a chapter on ethics can provide guidance on how to address these issues. These documents can outline the ethical aspects and requirements of the funding agency and provide guidance on how to ensure compliance with relevant ethical and legal standards.

Where personal data are collected as part of a research project, it is important to obtain informed consent from research participants to share and store the data on a long-term basis. Informed consent is the process by which individuals are provided with information about a research study and allow to make an informed decision about whether to participate in the study.

The informed consent process should include a discussion of the potential risks and benefits of participating in the study, as well as the implications of data sharing and long-term storage. Subscribers should be provided with information on how their data will be used, who will have access to the data and how the data will be stored and stored over time.

Questionnaires can be an effective way of obtaining informed consent from research participants, but it is important to ensure that questions are clear and concise and that participants have the opportunity to ask questions and ask for clarifications if necessary. In addition, it is important to ensure that participants can refuse to share data or store it for a long time if they so choose.

In some cases, it may be necessary to obtain additional consent to share or retain data for a long time beyond the initial informed consent obtained from research participants. This may be necessary if the data will be shared with external collaborators or if the data will be used for future research beyond the original study.

Overall, obtaining informed consent to share and retain data for a long time is an important ethical consideration in research involving personal data and should be included as part of the research planning process.

# 7. Conclusion

In conclusion, DMP describes the recording and handling of a heterogeneous set of data and associated outputs within the EDIH CTU project.

The document describes the areas that need to be accounted for and what effect they have on the processed data. The responsibilities for recording and deciding on the possible publication of these data and other outputs are also defined.

Considering the nature of the EDIH CTU project, which aims to create a library of thrusted solutions and services, it was necessary to create a DMP able to work with a heterogeneous set of data and other project outputs in changing environment. Many rules and principles are therefore formulated to be able to account for the range of data types and situations.

It is recommended to update the DMP during the project and specify selected principles as the availability of information about the services offered increases.