EDULAB

European Youth in Transition to Education and Labour

Data Management Plan (Year 1)

DMP-1

Deliverable number:	*D 7.2*
Due date:	*30.06.2025*
Nature:	*R = Report, P = Prototype, D = Demonstrator, O = Other*
Dissemination Level:	*PU*
Work Package:	*WP7*
Lead Beneficiary:	*Economica*
Contributing Beneficiaries:	
Authors:	*Stanislav Savov*

Funded by the European Union

A Horizon Europe collaborative project (2025-2027)
Grant Agreement #101177428

Purpose and scope of the deliverable

The goal of D7.2, the Data Management Plan, is to outline the types of data the EDU-Lab project will gather and process, its potential uses or availability for validation and further use, and the strategies for its management and preservation following the project's conclusion.

Document history

Version	Date	Description
0.1	*04.02.2025*	*Initial DMP outline drafted; structure and main sections defined*
0.2	*03.05.2025*	*Additional subsections drafted and expanded*
0.3	*21.05.2025*	*Draft of DMP prepared for internal review*
0.5	*12.06.2025*	*Revised and restructured by Alexander*
0.6	*26.062025*	*Final review by the coordination team*
1.0 (final)	*23.06.2025*	*Approved by PMB and the Coordinator*

Abbreviation	*Description*
AI	Artificial Intelligence
ASR	Automatic Speech Recognition
DAC	Data Access Committee
DMP	Data Management Plan
DOI	Digital Object Identifier
DPIA	Data Protection Impact Assessment
EIG	Ethics Issues Group
FAIR	Findable, Accessible, Interoperable, Reusable
GDPR	General Data Protection Regulation
ICF	Informed Consent Form
ISP	Information Sheet for Participants
JCA	Joint Controllership Agreement
NEET	Not in Education, Employment, or Training
PC	Project Coordinator
PIAAC	Programme for the International Assessment of Adult Competencies
PISA	Programme for International Student Assessment
SD	Structured Diary
WP	Work Package

Contents

1 Introduction

1.1 Purpose of the document

Data Management Plan (DMP) is a fundamental component of effective data management in EDU-LAB project. It outlines the data management lifecycle for data that will be collected, processed, and/or generated within a Horizon Europe project. To ensure research data is findable, accessible, interoperable, and reusable (FAIR), this DMP details the following aspects of data collecting, processing and storage:

- How research data will be managed during and after the project;
- The types of data to be collected, processed, or generated;
- The methodologies and standards to be applied;
- Whether the data will be shared or made open access;
- How the data will be curated and preserved, including beyond the project's completion;

1.2 Project overview

EDU-LAB is a Horizon Europe collaborative research project (Grant Agreement No. 101177428) running from 2025 to 2027. The project aims to address systemic challenges related to the transition of young people (aged 15–30) from education to the labour market, with a strong focus on equity, inclusion, and evidence-based policy reform.

EDU-LAB explores how policies, practices, and institutional arrangements across different education and training systems—particularly in general education (GE) and professional/vocational education and training (PVET)—can be optimized to better support youth in accessing quality learning pathways, completing education, and transitioning into employment.

Key project dimensions include:

- Examining early school-leaving, work-based learning, and regional disparities
- Assessing the policy and investment landscape in the European Education Area
- Identifying and promoting effective and inclusive practices at the local, national, and EU levels.

The project brings together a multidisciplinary consortium of universities, research institutions, and policy-oriented organisations from across Europe. Through its work packages, EDU-LAB will deliver a combination of empirical research, policy recommendations, and practical toolkits designed to inform and support stakeholders in education, governance, and labour market development.

The project's ambition is not only to generate knowledge, but also to facilitate its uptake through collaborative stakeholder engagement, strategic communication, and targeted dissemination, ensuring that its outputs are accessible, relevant, and impactful at all levels of policy and practice.

2 EDU-LAB Data Summary

2.1 Overview of data collected, processed and stored in EDU-LAB project

In EDU-LAB project we collect, process and store various statistical, socio-economic and personal data in the course or research programs per specific work packages; these data are of different nature as summarized in Table 1 below.

Table 1. Types data collected, processed and stored in EDU-LAB project

Work Package	Structured Instruments	Tabular Data (Quant.)	Code / Scripts	Text / Transcripts	Audio	Video	Visual / Outputs
WP1: Modelling pathways and transitions in GE and PVET and to the labour market	✓	✓		✓			
WP2: Efficiency assessment of policies and investments in GE and PVET	✓	✓		✓			
WP3A: Targeted analysis of quantitative data in EEA	✓	✓	✓	✓			
WP3B: International Online Survey of Experts and Stakeholders (OSES)	✓	✓	✓	✓			
WP4: Hearing young people's voices: Qualitative Case Studies	✓	✓		✓	✓		
WP5: Research synthesis, integration and recommendations for politics and practice		✓		✓			✓
WP6: Frameworks for the dissemination, exploitation and outreach activities		✓	✓	✓	✓	✓	✓
WP7: Project and consortium coordination, management and ethics issues	✓			✓			

2.2 Data Collection Methods in EDU-LAB

Data collection in the EDU-LAB project employs a mixed-methods approach, integrating both quantitative and qualitative techniques to gather comprehensive data:

- **Quantitative Methods:** Surveys and standardized assessments are utilized to collect structured data from participants. This includes leveraging existing datasets like PISA and PIAAC, which provide valuable benchmarks and longitudinal insights that enhance the project's analytical capacity. These methods are deployed mainly in WP1, WP2 and WP3A.
- **Qualitative Methods:** To capture the depth of individual experiences and social interactions, the project employs focus groups, participatory workshops, and structured interviews. These methods are particularly crucial in Work Package 4 (WP4 – Case Studies), as well as in WP3B (OSES-Delphi survey and preceding exploratory interviews), where understanding the nuanced choices and transitions of young people is key. Focus groups facilitate dynamic discussions, participatory workshops engage participants in reflective activities, and structured interviews provide detailed personal narratives.

Each method is carefully chosen based on its suitability to the research questions addressed in different work packages. This ensures that the data collected are not only comprehensive but also directly aligned with the project's research objectives.

- **Adaptability and Integration:** The methodologies described are flexible and adaptable, allowing for tailored approaches across various contexts within the project. This adaptability ensures that the data collection strategies remain effective and relevant throughout the project's lifecycle.

2.3 Types of data operated by EDU-LAB

As in all social studies, we distinguish ***primary*** and ***secondary*** data, as explained in the follow-up sections. A special emphasis is given to the collection, processing and storage of sensitive data.

2.3.1 Primary data

Primary data include all data collected directly by the EDU-LAB consortium partners as part of the research activities. These datasets are generated through standardized instruments and qualitative methods implemented in the case study countries.

The most important primary data types are:

Socio-demographic survey responses

Collected via LimeSurvey and providing structured, numerical datasets on:

- Age, gender, education level, employment status, nationality, migration background
- Self-perception of agency, future expectations and well-being
- Experience with education and training systems

Storage in CSV or XLSX format with associated codebooks and variable descriptions. Responses are recorded anonymously or pseudonymized depending on consent and legal requirements.

Structured diaries

Participants write entries over several weeks in the diary templates developed by WP4

These include:

- Quantitative data
- Qualitative narratives on everyday experiences, challenges and identity development

Submission digitally or as a scanned paper version. Diaries carry pseudonymous participant IDs and are saved in DOCX, PDF or TXT formats. All direct identifiers are removed before analysis or sharing.

Focus groups and participatory workshops

Conducted online (ex. Zoom) or on site.

These sessions will generate:

- Audio or video recordings, provided consent has been obtained
- Full or partial transcripts (DOCX/TXT) in the source language
- English summaries and thematic coding
- Structured Diaries (SDs)
- Participant Background Information
- Group-Generated Problem Definitions and Solutions
- Within-Case and Cross-Case Analytical Summaries
- Plain Language Reporting Outputs

These datasets explore group dynamics, institutional engagement and life course narratives from young people's perspectives.

Semi-structured interviews

Semi-structured interviews will be, among others, conducted with:

- Educators
- Policy makers
- Youth advisors
- NGO and labour market stakeholders

Verbatim transcripts, audio recordings and observational notes by the interviewers are collected. Transcripts are translated and analysed thematically. Identifiable data remains securely stored and is only shared within the institution of origin and the project coordinator.

Researcher notes and observation data

Researchers may document during data collection:

- Contextual information (e.g. framework conditions, participants' reactions)
- Reflective observations
- Adjustments to the protocol

These notes are stored securely, labelled according to location and used internally for interpretation and transparency.

2.3.2 Secondary Data

Secondary data are data sets from external sources that were collected independently of the EDU-LAB project. They are used in WP3 and WP4 in particular for comparison, contextualization or exploratory analysis and are always used in accordance with the respective licensing and ethical requirements.

International education and skills datasets

EDU-LAB makes extensive use of publicly accessible datasets, including

- PISA (Program for International Student Assessment)
- PIAAC (Program for the International Assessment of Adult Competencies)
- TIMSS and TALIS (where applicable)

These sources provide standardized measures of academic achievement, skills acquisition and socio-demographic background of young people. They are used to benchmark local case study results and to analyse national or regional developments.

Labor market and socio-economic data

Also included:

- EU-LFS (Labor Force Survey)
- Eurostat indicators on youth employment and education
- National censuses and household surveys (e.g. on NEET rates, educational transitions or ethnic background

These datasets support policy classification, especially for analyzing structural factors that influence young people's educational and career paths.

Access and licensing

- All reused data originate from
- Open Access repositories
- Official statistical offices
- Sources with an explicit reuse license (e.g. CC BY, public domain)

No personal data from secondary sources is used. If necessary, variables are harmonized with EDU-LAB structures to enable integrated analyses.

Data formats and documentation of secondary data

- Secondary data is usually available in the following formats
- CSV, XLSX or XML for tabular data
- SAV (SPSS), DTA (Stata) or RData for analysis-ready files

Each dataset is stored together with a README file, citation and license information as well as notes on preprocessing and variable transformation.

2.3.3 Purpose of Data Generation or Re-use

The data generated and re-used will critically inform the project's objective to map and understand the education-to-labour market transitions within the EEA. Newly generated qualitative data will provide in-depth insights into individual experiences, decisions, and contextual factors influencing pathways and transitions. In contrast, re-used aggregated datasets (e.g., PISA, PIAAC) will offer benchmarks and broader comparative insights to contextualize these individual narratives against established trends.

This dual approach allows EDU-LAB to provide a holistic and nuanced understanding of the factors influencing educational and professional pathways, directly supporting the project's goals of enhancing policy and practice across Europe.

Data collection and processing occur under lawful grounds defined by GDPR (Art. 6(1)(e) and Art. 9(2)(j)), explicitly for purposes of scientific research in the public interest, as outlined in the project's DPIA.

2.3.4 Overview and Reasons for Re-use of secondary quantitative data

Re-use of Existing Data: EDU-LAB plans to re-use existing data from international datasets such as PISA and PIAAC, which provide comprehensive insights into educational and training outcomes and labor market integrations. These datasets are pivotal for benchmarking and contextualizing our newly collected data against established trends and metrics.

Reasons for Re-use: The re-use of these datasets will enhance the robustness of our analysis by allowing for longitudinal comparisons and deeper insights into the systemic transformations over time. Additionally, re-using these datasets ensures a cost-effective approach by leveraging pre-existing resources to maximize the project's analytical scope.

Consideration and Discarding of Data Re-use: Some potential datasets were considered but ultimately not utilized due to limitations in their scope or the specificity of data required by EDU-LAB that these datasets did not address, such as detailed micro-level transitions specific to the EEA region post-COVID-19.

Data Minimization and Purpose Limitation: Data collection will adhere to GDPR principles of data minimization and purpose limitation, ensuring that only data necessary for the defined purposes of the EDU-LAB project are collected.

2.3.5 Sensitive and Special Category Data

Sensitive / special categories of personal data (Art. 9 GDPR) may be collected as part of WP4's fieldwork (e.g. focus groups, structured diaries, socio-demographic surveys), including

- Ethnicity and cultural background
- Health status and disability
- Migration or refugee status
- Political or institutional trust
- Audio/video recordings with biometric or facial features

The DPIA confirms that this data will only be processed if necessary and only with explicit, informed consent; special safeguards apply to minors and vulnerable participants (abbreviated in the DPIA).

Legal and ethical basis for processing

- Art. 6 para. 1 lit. a GDPR - explicit, informed consent
- Art. 9 para. 2 lit. j GDPR - processing for scientific research purposes
- Requirements from DPIA, Privacy Policy and Joint Controllership Agreement (JCA)

Double consent is obtained for minors (15-17 years): Consent of the minor(s) and consent of the legal guardian (Ethics Plan D7.1).

Storage, protection and transmission

- Local storage on institutional servers of the collecting partner organization
- Encryption at rest and during transmission

- Access only for authorized team members
- Pseudonymization or anonymization before internal transfer or analysis
- No transfer outside the EU/EEA without appropriate safeguards (e.g. standard contractual clauses)

Audio and video files are stored separately from transcripts and are provided in raw format only to the coordinator. Identifiable content is not shared beyond the originating partner and the coordinator.

Documentation and supervision

All sensitive data processing is:

- documented and stored securely;
- logged under institutional GDPR compliance records;
- overseen by the local data protection officers (DPOs);
- subject to internal audits by the Ethics Issues Group (EIG) as needed.

2.4 Special note on data collected in the course of Case Studies (WP4)

No data will be reused in the WP4 Case Studies. New data will be collected and generated through multiple qualitative methods of data collection. The study will collect and generate mostly qualitative data.

In each Case Study location, the data will be collected using Focus Groups, Participatory Workshops, and Structured Diaries. Before the Focus Groups and after securing consent to study participation, the study will also collect socio-demographic information from the participants (e.g., gender, age, education level and track, ethnic, migration and socioeconomic background, employment status) using the Sociodemographic Survey.

The study involves collecting and processing qualitative personal data, including socio-demographic information and individual experiences, to explore young people's choices, pathways and transitions in education and training and from education and training to the labour market.

Preference will be given to online data collection, but some data may be collected in person. All data collection will adhere to the highest ethical standards and General Data Protection Regulation (GDPR) guidelines. Data collection will adhere to principles of data minimization and purpose limitation, ensuring that only data necessary for the defined purposes of the EDU-LAB project are collected. Special consideration will be given to ensuring the privacy and data protection, particularly for minors or adolescents under the minimum age for autonomous consent to study participation, and other vulnerable populations (e.g., youth not in education, employment or training, school dropouts).

The following software platforms will be used to collect the data: (1) videoconferencing platforms (e.g., Zoom, MS Teams) for conducting the Focus Groups and the Participatory Workshops; (2) online interactive collaboration platforms Full or partial transcripts Full or partial transcripts for collecting data and facilitating group dynamics during the Participatory Workshops; (3) software platforms (e.g., LimeSurvey) for collecting data with the Structured Diaries and the Sociodemographic Survey. Automatic recording and transcription of the Focus Groups and Participatory Workshops (e.g., Otter.ai, Zoom Auto-Transcription) will also be carried out.

All WP4 qualitative data (focus groups, participatory workshops, and structured diaries) follow a standardized file naming and folder structure per case study location. File versions include original AI-generated transcripts, checked transcripts, pseudonymised transcripts, and English summaries. Each version is securely stored and labelled accordingly (e.g., FG3_pseud_LocationName, PW2_checked_LocationName). These conventions ensure traceability, compliance, and consistency across teams.

3 Types and Formats of Data in EDU-LAB project

3.1 Data Formats

Quantitative data formats

Used for structured survey responses and numeric data sets:

- CSV, XLSX - standard table formats for raw and cleaned data
- SPSS (.sav), R (.rds), Stata (.dta) - formats for statistical analysis
- XML - for structured data that requires tagging or cross-platform integration
- PDF/A - for archive versions of survey results and data summaries

Quantitative data sets are usually accompanied by a codebook containing variable descriptions, scales and transformation rules.

Qualitative data formats

Used for interview transcripts, diary entries, open survey responses and research notes:

- DOCX, TXT, PDF - for transcripts and text-based documentation.
- ODT - open format for compatibility with open-source software
- JSON - if structured annotations are available (e.g. thematic coding results)

Qualitative files are organized in folders by data type, pseudonymized participant ID and date and are accompanied by a README or data log file. Versioning distinctions (e.g., raw, checked, pseudonymised) are reflected in file names and folder structures as part of internal data workflows, especially in WP4.

Audio/video and multimedia

Used for recordings of interviews, focus groups and workshops:

- MP3, WAV, M4A - audio formats for voice recordings
- MP4, AVI, MKV - video formats for visual recordings

The files are stored separately from transcripts, stored in encrypted form and only stored until the conclusion of the project on the 31st of December 2027.

No biometric data (e.g. facial recognition, voiceprints) is extracted from the recordings.

Graphics, visualizations and presentation materials

- JPEG, PNG - raster graphics
- PDF, PPTX - slides, reports and annotated visualizations
- SVG - scalable vector graphics

These files are stored in structured folders that are clearly assigned to the respective deliverables and publication results.

Script and software files

Used for reproducibility of the analysis:

- R scripts, SPSS syntax (.sps), Markdown files (.md; .markdown) Python scripts (.py) - internal scripts for data cleaning and analysis.
- They are only shared publicly if they are licensed accordingly and contain documentation on dependencies and usage instructions.

All datasets are stored using a standardized folder structure distinguishing between raw, processed, and anonymised data.

Table 2. Reference table of data formats in EDU-LAB project

Format	Type	Use Case	Data Category
CSV / XLSX	Tabular	Survey results, diary metrics	Quantitative (Primary)
SPSS / R / Stata	Statistical	Analysis and model outputs	Quantitative (Primary/Secondary)
DOCX / TXT / PDF	Textual	Interview and diary transcripts	Qualitative
MP3 / WAV / M4A	Audio	Interviews, focus groups	Qualitative (Raw)
MP4 / AVI / MKV	Video	Workshops, participatory activities, Interviews, focus groups	Qualitative (Raw)
JPEG / PNG / SVG	Visual	Dissemination, participatory outputs	Outputs / Dissemination
XML / JSON	Structured metadata or coded files	Exportable metadata, annotations	Cross-type
PDF/A	Archival	Finalised reports, summaries	Preservation
.py / .r / .sps	Code/scripts	Analysis pipelines	Internal (Reproducibility)

3.2 Software and Platforms Used in EDU-LAB

The EDU-LAB project uses a combination of open source and institutionally supported software for secure data collection, storage and processing. The selected platforms meet GDPR-compliant requirements, are interoperable with common data formats and can be used by all partner institutions.

Survey and diary tools

- LimeSurvey - is used to distribute socio-demographic surveys; supports multilingual questionnaires and structured export formats (CSV, XLSX).
- Individual diary templates (DOCX/PDF) - completed by participants digitally or in paper form and then uploaded to secure storage systems.

Tools for qualitative data

- Commonly available word processing software like MS Word / Nextcloud Office / Google Docs - for transcript editing and annotation.
- Zoom, MS Teams or similar – for online interviews, focus groups and participatory sessions.
- Transcripts and pseudonymised outputs from these sessions are imported into qualitative analysis software (e.g., MAXQDA or similar), with resulting coded datasets stored securely in their native format and in exportable CSV/XLSX files

Quantitative data analysis

- SPSS, R, Stata - for statistical analysis.
- Excel / Nextcloud Office - for basic data cleaning and formatting.
- Analysis scripts are stored internally as .r, .sps or .py (only shared publicly if required).

Storage and file exchange

- Nextcloud (self-hosted) - central internal system for secure file exchange, exchange of pseudonymized data sets and access-controlled collaboration.
- Institutional servers - used for long-term storage of raw and sensitive data under the responsibility of the respective partners.
- Zenodo - certified public repository for long-term archiving and publication of anonymized datasets; FAIR-compliant and supports DOI allocation.

Metadata and documentation

- Dublin Core / DataCite templates - for assigning metadata before uploading to Zenodo.
- README and Codebook templates - Standardized README and Codebook templates are currently under development and will be implemented across all WPs during the first round of data collection. These will ensure internal consistency and support future data sharing.

Exploratory tools and future infrastructure

The consortium is exploring the use of local Large Language Models (LLMs) and Automatic Speech Recognition (ASR) systems to support data processing workflows, in particular:

- WhisperX AI (open source) - will be used for local, offline transcription and translation of audio recordings (e.g. interviews, focus groups) to reduce reliance on cloud services and strengthen privacy and GDPR compliance.
- Locally hosted LLMs - Llama3.3 or Llama4 will being used in terms of summarization, initial thematic coding or metadata generation, where use increases efficiency without compromising privacy or ethics standards.

These tools would be deployed in a secure offline environment in full compliance with the project's DPIA, Ethics Plan and Joint Controllership Agreement. Their integration will be re-evaluated during a consortial review and approval process and documented in future versions of the DMP (e.g. DMP-2).

3.3 Data Retention and Deletion

This chapter describes principles and rules for data retention and deletion in EDU-LAB project.

Data Retention:

- **Direct Identifiers**: Personal data containing direct identifiers will be retained only for as long as necessary for the purposes of validation and analysis, not exceeding two years post-project conclusion, as aligned with the GDPR and project-specific privacy regulations.
- **Pseudonymized and Anonymized Data**: These datasets may be retained in GDPR-compliant repositories such as Nextcloud and Zenodo for at least five years following the project's completion. These datasets may be made publicly available when appropriate and when all privacy protections are in place.

Deletion Procedures:

- After the two-year retention period, all non-anonymized personal data and sensitive data (e.g., health, ethnicity) will be securely deleted or anonymized. This includes irreversible deletion from both active storage and backups. Audio and video recordings containing identifiable information will be deleted unless anonymization renders individuals unidentifiable, in which case retention may be extended.
- Encryption of sensitive data, both in storage and transit, will be applied throughout the retention period as mandated by the DPIA and GDPR.

Participant Rights and Exemptions:

- Participants may request deletion or anonymization of their personal data at any time. However, in accordance with GDPR Article 17(3)(d), if the data is necessary for scientific research and cannot be anonymized, it may be retained in pseudonymized form and used for research purposes, even if the participant withdraws.

Compliance:

- All retention and deletion practices are carried out in compliance with GDPR, national data protection laws, and the EDU-LAB Privacy Notice, and follow the shared responsibility structure defined in the Joint Controllership Agreement (JCA).

Retention and Deletion Justification:

- These measures ensure the project aligns with ethical standards and legal requirements, maintaining the integrity of the research while protecting participant privacy. Data that is no longer necessary for scientific inquiry or that could compromise participant anonymity will be responsibly phased out of active use and securely disposed of.

3.4 Further conventions on data management in EDU-LAB (size of data, provenance, data exchange)

3.4.1 Expected Size of Data

The expected size of the data is still being assessed. The volume will depend on several factors, including the number of qualitative sessions conducted across case study sites, the scale of survey distribution, the resolution and duration of the audio and video materials etc.

Early estimates suggest that the total volume of data (including raw, processed, and re-used datasets) may range from a hundred gigabytes to no more than one terabyte.

This includes:

- Raw qualitative data, including audio and video recordings, transcripts and structured diaries from 12 case study sites in seven countries.
- Quantitative survey data collected via online platforms (e.g. LimeSurvey) and processed statistical outputs (CSV, XLSX, SPSS, R).
- Reused external datasets, including downloaded subsets of international datasets (e.g. PISA, PIAAC) and national statistical data stored locally and integrated for benchmarking and comparative analyses.

A more precise estimate will be provided in DMP-2 once data collection progresses and storage needs are fully assessed. Current infrastructure (Servers using Nextcloud providing accessible internal storage, Zenodo for open access) is expected to be fully sufficient for managing this volume.

3.4.2 Data Origin and Provenance

The data will be collected from twelve regions or cities across seven European countries, ensuring a diverse and comprehensive dataset that captures Europe's geographical and socio-economic diversity, as well as diverse welfare regimes. The seven countries and twelve locations in these countries where the data will be collected are: Austria (Vienna and Graz region), Finland (Tampere and Helsinki region), Italy (Bologna region), Kosovo (Pristina region), Poland (Warsaw and Lodz), Portugal (Porto and Santarém), and the UK (Birmingham and Worcestershire).

Participant recruitment will be purposive, ensuring diversity in age (15-30 years), educational background and employment situation of the Case Studies Participants within and across the twelve locations in seven European countries.

Recruitment will take place in a variety of settings (e.g., schools, VET institutions, universities, universities of applied sciences, employment agencies, young professionals' organizations, NGOs and other youth organizations), which may also be used for collecting the data.

Data will be gathered through direct participant engagement, primarily via online platforms, with accessibility and flexibility in data collection procedures being ensured by the research teams involved in the WP4 Case Studies.

In addition to newly collected data, the project will incorporate data re-used from established international databases (eg. PISA and PIAAC). These datasets are curated and maintained by renowned organizations, including the OECD, and provide valuable longitudinal insights into educational and training outcomes and workforce skills development.

For both newly collected and re-used data, provenance will be documented, including details about the source, method of collection, responsible partner, and any processing steps. This will ensure traceability, validation, and compliance with GDPR and ethical requirements. Each partner is responsible for maintaining accurate and up-to-date records for the data they manage.\

In addition to newly collected data, the project will incorporate data re-used from established international databases (e.g., PISA and PIAAC).

3.4.3 Utility of Data for External Parties

Anonymized datasets and associated research outputs generated by EDU-LAB will be of use to a broad range of external stakeholders:

- Educational researchers and policymakers interested in the dynamics of education and training systems and labour markets.
- Social scientists studying youth transitions, resilience, and socio-economic outcomes.
- Non-governmental organizations focused on educational equity and access.
- European Union bodies and local government agencies devising educational, training and labour market policies.
- Academic institutions and libraries compiling data on educational outcomes and workforce integration.
- Organizations like student unions, NGOs representing young people and all sort of civic organisations working with and for youth.

4 Principles of FAIR Data policies in EDU-LAB

4.1 Making Data Findable, Including Provisions for Metadata

4.1.1 Discoverability and Identification of Data

All datasets selected for open access in the EDU-LAB project will be assigned Digital Object Identifiers (DOIs) and accompanied by standardized metadata to ensure they are easily identifiable, citable, and locatable.

DOIs will be assigned by the repository (e.g., Zenodo), enabling referencing in publications and online platforms.

Only datasets that are at least pseudonymized and approved for sharing will be made publicly discoverable through repositories and linked project outputs.

To improve accessibility, particularly for users with visual impairments, alternative text descriptions will be provided for images and figures associated with shared datasets and publications, in line with inclusive research dissemination practices.

4.1.2 Naming conventions

The project will Implement a clear and efficient naming convention to facilitate easy file identification, organization, and retrieval.

Guidelines:

- **Structure**: [ProjectAcronym]_[Year]_[Region]_[DataType]_[Descriptor]_[Version]_[Initials]
- **Example**: EDU-LAB_2025_UK_Survey_v1_AH

Principles:

- **Clarity**: Use camel case for readability without spaces (e.g., SurveyData).
- **Simplicity**: Keep names short, using only alphanumeric characters and underscores.
- **Sorting**: Include dates in a YYYY-MM-DD format within parts of the filename to support systematic sorting.
- **Version Control**: Clearly denote version information (e.g., _v1 for version 1).

Benefits:

- **Predictable Sorting**: Enables systematic organization and chronological tracking.
- **Quick Identification**: Reduces time spent searching for specific files, ensuring efficient data management.

This streamlined convention ensures that all project data is systematically named, promoting consistency and ease of access across the project's lifecycle.

Additionally, a participant ID and personal code system will be used to manage case study data. These codes will be part of the project's pseudonymization framework to protect personal data in compliance with GDPR.

While the general naming convention outlined above applies project-wide, specific work packages such as WP4 use a simplified, task-oriented structure (e.g., FG3_pseud_Warsaw, PW1_checked_Lodz) that emphasizes data type, version status, and collection site. This ensures operational clarity and consistency across local teams handling qualitative data.

4.1.3 Search keywords

Metadata will include specific search keywords, such as "education transitions," "vocational training," and "youth employment," "social inclusion" etc. to enhance discoverability and re-use and to ensure a level of standardisation across data from different partners.

Additional terms may be added during the project as it progresses. Partners will review and update keywords prior to dataset deposition to ensure consistency. Controlled vocabularies, such as those used in EuroVoc or the ERIC Thesaurus may be used and consulted to support interoperability and standardisation across the platforms.

4.1.4 What metadata will be created?

Metadata will be created to support the discoverability, accessibility, understanding, and reuse of all datasets selected for open access.

While the exact schema is still not set in stone and is currently under development, the following types of metadata will most likely be generated for each dataset:

- Descriptive metadata (e.g. title, author, date, language, keywords, abstract/description)
- Administrative metadata (e.g. file formats, licensing, access rights, version history)
- Provenance metadata (e.g. source, data collection context, responsible partner)
- Structural metadata (e.g. folder organization, relationships between files)

Widely used standards such as Dublin Core or the DataCite metadata schema will probably be adopted, particularly for datasets published via platforms like Zenodo, which supports these standards by default.

Final metadata will be reviewed before public deposition to ensure completeness and consistency across all shared datasets.

4.2 Making Data Openly Accessible

4.2.1 Data Access and Rectification Rights

Participants will be provided with clear mechanisms to access their personal data and to request the deletion, of this data, in accordance with their rights under GDPR.

These rights only apply to personal and identifiable data. Once data has been anonymised or pseudonymised, it is no longer subject to the rights of data subjects under the GDPR.

Requests are processed by the designated data protection contact person at each participating organisation. The procedure for submitting a request is explained in the informed consent form and in the privacy notice provided to participants.

In accordance with Article 17(3)(d) GDPR, data that cannot be anonymised and is still required for scientific research purposes may be stored in pseudonymised form even after revocation, provided that suitable safeguards are in place.

4.2.2 Data Availability

All public project results, including deliverables, publications, research findings, and reports on survey results will be made openly accessible via suitable platforms, such as the EDU-LAB website, Zenodo, and academic journals.

Anonymized datasets that do not contain sensitive or proprietary information are made available via open-access repositories after undergoing appropriate review and curation.

Data sets containing sensitive personal data, biometric content (e.g., **identifiable** video or audio recordings), or confidential information are not shared publicly. Only fully anonymized and ethically approved data is disseminated.

The availability of datasets depends on project timelines and publication deadlines. Data linked to specific deliverables is generally released within six months of European Commission approval or in accordance with journal publication requirements.

4.2.3 Data Accessibility Methods

Preliminary and working data will be stored securely on the organisation's self-hosted servers using Nextcloud, which will provide encrypted access and version control. Access is restricted to authorised consortium members and permissions are managed internally.

For public dissemination and long-term archiving, finalised and anonymised datasets will be deposited in an open access repository such as Zenodo, in line with Horizon Europe's Open Science policy. Publicly

accessible datasets contain the relevant metadata and documentation to facilitate discoverability and reuse.

In addition, selected research results and curated datasets can be made available via the EDU-LAB project website, which will provide a dedicated public access portal for downloading reports and anonymised data products.

4.2.4 Tools and Software Required for Data Access

The anonymised data collected through focus groups, participatory workshops, surveys and interviews are accessible with common productivity software (e.g. Microsoft Office - Word, Excel - and PDF readers). All data sets intended for public access are provided in widely compatible formats.

The data can be accessed with any standard web browser. For example, all materials stored in Zenodo are accessible in accordance with the platform's established protocols and access guidelines.

If proprietary formats are used (e.g. SPSS, R), open-format equivalents (such as CSV) are also provided to ensure wider accessibility. Where possible, open-source software is favoured for the distribution and viewing of materials.

4.2.5 Documentation of Access Software

No specific software requirements are currently expected to access the data intended for public sharing. The project relies on commonly available tools such as Microsoft Office applications, common PDF viewers and web browsers as described in section 4.2.4.

During the research process, some internal datasets may be processed using proprietary or specialised software (e.g. SPSS, R). However, all publicly released datasets are accompanied by open format equivalents (e.g. CSV, TXT) to ensure accessibility for users regardless of the software available.

Should specialised software still be required to access certain results, the project will provide accompanying documentation describing the tool, version and file format used.

This approach will be reviewed during the course of the project and updated in future versions of the DMP.

4.2.6 Inclusion of relevant software (e.g. open-source code)

The project is committed to transparency and reproducibility, and all software solutions currently used are open source.

Should relevant software, scripts or digital tools be developed or adapted in the course of the project, these can be shared publicly – provided that licensing and ethical framework conditions permit this.

Such software or code would be made available via open repositories such as GitHub or Zenodo and comprehensively documented to ensure problem-free use. The decision to publish specific software resources will be reviewed in the second phase of the project and recorded in the updated DMP.

4.2.7 Data and Metadata Depository

The complete datasets, including metadata and documentation will be initially stored and managed on the organization's Nextcloud platform, enabling secure access for project participants. the final and anonymized dataset will be transferred to the Zenodo repository to ensure long-term preservation and public accessibility.

Each dataset deposited in Zenodo will be accompanied by structured metadata and assigned a Digital Object Identifier (DOI) to ensure persistent referencing and discoverability.

The EDU-Lab project maintains a dedicated website (www.edu-lab-project.eu) outlining the project's objectives, methods, and current progress. The site features a public downloads section for reports and publications, with open access to all visitors. While most content is freely accessible, a private area is reserved for consortium members.

4.2.8 Arrangements with Identified Repositories

Arrangements with the Zenodo repository are being explored. Ensuring our data sharing practices align seamlessly with Open Research Europe (ORE) to comply with European open science mandates is a priority.

4.2.9 Access Restrictions

The data deposited in Zenodo will be freely accessible, and no restrictions are expected for datasets that have been fully anonymised and ethically cleared for public release.

However, certain project datasets - particularly those containing sensitive personal data, identifiable audio/video content or information subject to ethical restrictions - will not be made publicly available. These datasets remain on secure internal storage systems with restricted access limited to authorised consortium members according to predefined roles and responsibilities.

Only fully anonymised datasets that have been reviewed and approved for public access are stored in Zenodo or made available via the project website.

4.2.10 Data Access Committee

A Data Access Committee (DAC) is not currently deemed necessary for the EDU-LAB project because data access is controlled through secure authentication mechanisms on the Nextcloud server, with access restricted to authorized project team members and stakeholders. Sensitive data is managed under strict GDPR compliance, with predefined access controls and encryption measures ensuring data confidentiality.

However, if future assessments indicate a need for more stringent review processes for sensitive or restricted data, the establishment of a DAC will be considered. In that case, the DAC would be responsible for reviewing and approving access requests, ensuring compliance with ethical and legal standards. This decision will be revisited periodically as the project progresses.

4.2.11 Conditions for Access

The conditions for access to project data are defined by the principles enshrined in the consortium agreement and the Joint Controllership Agreement (JCA). These documents define roles, responsibilities and access rights of the EDU-LAB partners and ensure that data exchange within the consortium complies with both internal governance structures and GDPR.

Access to project data stored on our local Nextcloud server is restricted to authorised consortium members and permissions are assigned according to each partner's role in the project. Access is authenticated and monitored according to defined internal protocols.

Only datasets that have been fully anonymised and ethically approved for sharing will be made publicly available - primarily via Zenodo and the EDU-LAB website, and published under appropriate open access licensing conditions.

For data stored in the ZENODO repository, access will be facilitated through clear instructions and a comprehensive user manual. Where sensitive data, including personal data or other protected information, is involved, access will be managed in accordance with privacy and confidentiality obligations in line with the GDPR and any other applicable data protection laws. In these cases, specific access conditions will be explicitly defined and communicated as required.

Personal and sensitive data will not be shared outside the consortium. Within the project, access to such data is restricted to persons who are explicitly responsible for the processing in accordance with the provisions of the JCA.

4.2.12 Data Transfer Outside the EU:

The EDU-LAB project currently has no plans to transfer personal data outside the EU or the EEA, with an exception of two consortium partners located in the UK and Kosovo. Should such a transfer become necessary, it will be carried out strictly in accordance with Articles 44-49 of the General Data Protection Regulation (GDPR).

In this case, appropriate safeguards will be applied, including the use of Standard Contractual Clauses (SCCs) approved by the European Commission. Transfers will be limited to anonymised or pseudonymised data sets where possible; any transfer of personal data will be accompanied by a transfer impact assessment where necessary.

All data processing partners remain jointly responsible for ensuring that international data transfers comply with the principles laid down in the Joint Controllership Agreement (JCA) and the relevant national laws.

4.2.13 Verification of Data Access Identity

The datasets provided via Zenodo are fully anonymised and made publicly available under open licence terms. Zenodo does not require individual identity verification to access open data and no additional authentication is required for these datasets.

For internal data access (e.g. provisional or sensitive datasets stored on the internal Nextcloud servers), identity verification is performed via institutional login credentials and secure authentication procedures. Access is only granted to authorised personnel with specific project roles, in accordance with the principles set out in the Joint Controllership Agreement (JCA).

Access to personal or sensitive data outside the consortium is not granted.

4.3 Making Data Interoperable

4.3.1 Data Interoperability

The data will be formatted and managed using standard data and metadata vocabularies to ensure interoperability. Our internal management on the Nextcloud server will adhere to these standards to facilitate seamless integration and transfer to broader data platforms like Zenodo.

The EDU-LAB project will structure and store the data in widely used, non-proprietary formats to ensure long-term accessibility and interoperability across platforms and research contexts.

- Tabular and text data will be stored in CSV, XLSX and TXT formats.
- Audio and video data are archived as MP3, MP4, M4A, AVI, MKV.
- Transcripts, structured diaries and related files are provided in DOCX and PDF with standardised formatting and markup.
- Multilingual data is processed using a standardised versioning and translation approach; where possible, English summaries are attached to the original languages. In many instances, translations would be done with the help of locally hosted AI tools as previously mentioned (section 4.2).

Standardised naming conventions and metadata practices promote semantic coherence. Where appropriate, controlled vocabularies and ontologies (e.g. EuroVoc, ERIC-Thesaurus) are used to standardise keywords, topics and classifications.

These measures ensure compatibility with repository requirements (e.g. Zenodo) and increase the potential for data integration and reuse in educational and social science research contexts.

4.3.2 Data and Metadata Vocabularies, Standards, or Methodologies

The EDU-LAB project will use recognised vocabularies and standards to improve the interoperability, findability and reusability of its datasets. These will be selected and harmonised across the consortium to ensure consistent classification and metadata generation.

The following options are currently being considered:

- Metadata standards: Dublin Core and DataCite, both of which are supported by Zenodo and widely used in open research repositories.
- Thematic vocabularies: EuroVoc and the ERIC Thesaurus to align dataset descriptions with established classifications in education, labour market and policy.

The final selection of vocabularies and standards will be reviewed and validated prior to publication of the public datasets to ensure compatibility with the metadata requirements of the hosting platforms and compliance with the FAIR principles.

4.3.3 Use of Standard Vocabularies for Interdisciplinary Interoperability

To support interdisciplinary research and integration in areas such as education, labour market research, youth research and social sciences, the EDU-LAB project will adopt standardised vocabularies and classification systems. These will enable consistent interpretation of data across disciplines, institutions and infrastructures.

Standard vocabularies currently under consideration are:

- **EuroVoc** for European institutional and policy descriptors
- **ERIC Thesaurus** for terminology in education and pedagogy
- **CESSDA** vocabularies, insofar as they are relevant for indexing in the social sciences

In addition to thematic vocabularies, EDU-LAB datasets are provided in widely compatible formats (e.g. CSV, TXT, XML, PDF) to ensure their accessibility across platforms and software environments. Where necessary, the datasets are accompanied by a readme file in plain text (ASCII) format containing instructions on how to access or convert the files, including links to suitable freely available open-source tools.

4.3.4 Mappings to Common Ontologies

Where relevant, the EDU-LAB project will map key data elements and thematic concepts to widely recognised ontologies to support semantic clarity and interoperability with other datasets and infrastructures.

Potential ontologies include:

- OECD Fields of Science and Technology (FOS) classification
- European Skills, Competences, Qualifications and Occupations (ESCO) framework
- ISCED (International Standard Classification of Education) for educational levels
- CEDS (Common Education Data Standards), where applicable for structured educational descriptors)

These categorisations are made selectively - especially for variables and topics that are intended for dissemination, retrievability or reuse. Ontological harmonisation takes place during data curation and, if supported by the hosting repository (e.g. Zenodo), is integrated into machine-readable metadata. These measures increase the visibility, comparability and reusability of EDU-LAB datasets in broader research contexts.

4.4 Data Quality Assurance Processes

4.4.1 Data Collection and Input Validation

Data will be collected using standardized and validated instruments, including quantitative surveys, qualitative interviews, and focus groups. Electronic data collection systems will include validation checks to minimize data entry errors. For manual data collection, cross-validation and double-checking methods will be applied. The data management team will ensure that data collection tools are consistent with the project's objectives and comply with GDPR requirements.

In order to ensure the accuracy, consistency and reliability of the data collected, standardised survey instruments and procedures are used in all partner institutions and research locations. All research teams use harmonised protocols for focus groups, participatory workshops, structured diaries and surveys. These procedures are described in the internal project documents and regulated by the roles and responsibilities defined in the Joint Controllership Agreement (JCA).

Input validation measures include:

- Real-time validation in digital survey platforms (e.g. mandatory fields, format checks)
- Manual consistency checks for qualitative data entries (e.g. diaries, transcripts)

- Multilingual coordination using standardised templates and translation protocols for local language versions

Training and written guidelines are provided to all researchers involved in data collection to ensure compliance with protocols. Any deviations will be monitored and corrected as part of ongoing project oversight.

Each partner is responsible for ensuring the accuracy of the data at the time of collection, in accordance with its role in the Joint Controllership Agreement (JCA). These responsibilities will be implemented through internal validation measures. The data management team coordinates the collection tools and ensures that they meet the project objectives and comply with the GDPR and other relevant data protection standards.

These procedures are designed to ensure high-quality, interoperable data that is suitable for integration and analysis in different project contexts.

4.4.2 Data Cleaning and Verification

To ensure the accuracy, reliability and comparability of the data between the partners, all data collected is subjected to a structured cleaning and verification process. This includes:

- Checking quantitative survey data for missing values, duplicate entries and inconsistent responses
- Validation of qualitative data - such as interview transcripts and structured diaries - for accuracy, completeness and compliance with format specifications
- Standardisation and validation of coding schemes and classifications across all data sets

When qualitative data is collected in local languages, English summaries and translations are checked to ensure fidelity and consistency across sites. Clean-up work will be carried out by the local research teams and supported by central coordination structures defined in the consortium agreement, including work package leads and interdisciplinary coordination groups.

The consortium agreement defines mechanisms to ensure quality and consistency of project results, including roles for coordination, oversight and internal review. These responsibilities are carried out during data cleansing to ensure compliance with common methodologies and ethical obligations.

All personal or sensitive data processed during cleansing is handled in compliance with GDPR principles - in particular data minimisation, integrity and confidentiality.

Shared documentation, templates and cleansing protocols are used across the consortium to ensure consistency and traceability, thereby promoting the integrity and reusability of EDU-LAB datasets.

4.4.3 Data Harmonization and Standardization

Harmonization protocols will be implemented to standardize data formats, coding schemes, and variable naming conventions across all datasets. This ensures consistency, especially when integrating data from multiple sources or countries. International standards for data interoperability will be followed to facilitate data reuse and comparability.

This process will include:

- Harmonisation of the survey instruments between the partners
- Ensuring consistency in variable naming, coding schemes and data structures
- Standardisation of translated materials to ensure semantic equivalence across different languages
- Application of common definitions and classifications to qualitative and quantitative data

Harmonisation will be managed by the work package leaders and the coordination bodies defined in the consortium agreement to ensure fidelity to the original data and enable integrated analyses across different research contexts.

The harmonisation process will be documented and version controlled and all procedures will be shared with the consortium partners to ensure transparency, traceability and replicability of the harmonised datasets.

4.4.4 Responsibility and Oversight

Responsibility for ensuring data quality is shared within the consortium and coordinated by clearly defined project structures. The work package leaders (WPLs) oversee all data-related activities in their respective areas - from collection, cleansing, verification and harmonization to documentation.

These tasks are supported by coordination mechanisms in accordance with the consortium agreement, including technical coordination groups and frameworks for co-operation between the partners. Where necessary, feedback and clarification loops between the local research teams and the central coordination units will be maintained to ensure consistent implementation of quality assurance procedures.

Data managers and lead partners in each work package monitor compliance with protocols, maintain documentation and resolve issues as they arise throughout the course of the project. Regular review meetings, internal audits and peer consultations support adherence to common data management practices and compliance with ethical standards and project objectives.

In the event of unresolved issues or inconsistencies, an escalation to the Project Management Board (PMB) can be initiated for a decision and clarification.

4.4.5 Documentation and Metadata Management

To ensure a well-organized, consistent and reusable database throughout the EDU-LAB project, a project-wide documentation and metadata strategy will be implemented. This strategy applies to all data generated or reused in the work packages - including quantitative surveys, qualitative research results and external data sets that are integrated for comparison or contextual purposes.

General principles for all data:

- All datasets are provided with clear documentation describing collection methods, data structures, variable names and any data transformations or cleansing applied.
- Metadata follows recognized standards such as Dublin Core or DataCite to ensure compatibility with repository platforms (e.g. Zenodo).
- Each dataset contains a README file that lists the following:
    - Purpose and scope of the data
    - File structure and naming conventions
    - Conditions of access or use
- For non-standardized formats, a text-only (ASCII) file explains how the data can be opened or converted with free tools.

Work package-specific implementation:

- WP4 Case Studies
    - Structured folder hierarchies differentiate between data types (e.g. focus groups, participatory workshops, structured diaries).
    - Documentation includes detailed protocols for each survey event, including date, tools used and team responsible.
    - Multilingual data is supported by English summaries and harmonized metadata templates.
    - Cross-reference files link participant IDs to data sets and ensure GDPR-compliant anonymity.
- WP3 and other work packages
    - For large-scale quantitative surveys or reused datasets (e.g. PISA, PIAAC) contain the metadata:
    - Source and licensing of external data
    - Pre-processing or integration steps
    - Description of variables and data structure
- Harmonized documentation templates ensure comparability between the work packages.

Work package leaders and data managers are responsible for coordinating documentation efforts in their area of responsibility; overall supervision lies with the project's central data management team. All metadata

and documentation practices support the FAIR principles and are in line with EDU-LAB's data governance structure.

Where documentation includes information related to participant IDs or pseudonymization schemes, the process will be reviewed by the Data Protection Officers (DPOs) in partnering teams to ensure compliance with data protection standards and ethical obligations.

4.5 Increase Data Re-use

4.5.1 Licensing for Data Re-use

All anonymized datasets intended for public dissemination will be made available under appropriate open access licenses to promote transparency, reproducibility and a wider impact.

The default license for most datasets will be ***Creative Commons Attribution (CC BY),*** which allows sharing, redistribution and editing with attribution. If legal, ethical or contextual considerations require it, alternative licenses such as ***CC BY-NC*** (non-commercial use only) or ***CC BY-ND*** (no derivatives) may be used.

The license decision is based on:

- The nature and sensitivity of the dataset
- The intended audience and scope of reuse (e.g. science, politics, education)
- Compliance with the GDPR and research ethics obligations

To comply with Horizon Europe requirements, all data and metadata will be made openly accessible through trusted repositories like Zenodo. Metadata will be licensed under CC0 (Public Domain Dedication) to ensure unrestricted use. Specific licensing details and any restrictions will be outlined in the data management plan.

All license terms are clearly stated in the metadata and documentation of the dataset, according to the requirements of the repository (e.g. Zenodo). The choice of license is checked by the data management team and the responsible work package leaders before publication.

4.5.2 Availability Timing for Data Re-use

Each dataset suitable for open access will be made publicly available within six months of the approval of the relevant deliverable by the European Commission to allow sufficient time for publication.

Only datasets that have been anonymized and reviewed in accordance with internal project governance - including the provisions of the Joint Controllership Agreement, the Privacy Statement and relevant ethics clearances - will be made publicly available. Where appropriate, embargo periods or staggered publication dates may be set, depending on ethical, legal or publication-related considerations.

This timeline ensures compliance with Horizon Europe's Open Science guidelines while meeting data protection obligations under the GDPR.

4.5.3 Usability of Data by Third Parties

Data intended for dissemination is fully anonymized and provided in common, machine-readable formats such as CSV, XLSX, TXT, PDF and MP3/MP4, M4A, WAV (depending on data type). These formats are compatible with open source and standard productivity tools and facilitate use by researchers, educators, policy makers and other stakeholders.

The metadata follows recognized standards (e.g. Dublin Core, DataCite) to ensure interoperability and improve findability in repositories such as Zenodo.

These measures ensure that third parties can effectively reuse the datasets, even if they were not previously involved in the EDU-LAB project.

4.5.4 Duration of Data Re-usability

The EDU-LAB project ensures that anonymized public datasets remain accessible and usable for at least five years after publication. This is in line with Horizon Europe's Open Science requirements and supports long-term research impact and data reuse.

Long-term archiving is achieved through storage in trusted repositories such as Zenodo, which provide persistent identifiers (DOIs), stable access and versioning. Each dataset is accompanied by comprehensive metadata and documentation to ensure long-term usability and interpretability.

These measures ensure that the data remains findable, accessible, interoperable and reusable (FAIR) beyond the duration of the EDU-LAB project.

5 Allocation of resources

5.1 Costs Associated with Making Data FAIR

The costs for implementing the FAIR principle (Findable, Accessible, Interoperable, Reusable) are anchored in the overall project budget and embedded in the work package structures and the responsibilities of the partners.

These costs include, among other things

- Creation of metadata and structured documentation
- Data cleansing, standardization and anonymization
- Maintenance of the internal data infrastructure
- Time and effort for data curation, coordination and quality assurance

A separate cost centre exclusively for FAIR activities is not identified; instead, these tasks are distributed across the work packages and personnel budgets as foreseen in the financial and operational project plan.

5.2 Coverage of Costs

The costs associated with data management activities - including cleaning, documentation, anonymization, metadata creation and repository storage - are covered by the overall project budget allocations. These activities are usually embedded in the implementation of the work packages and are carried out within the personnel time allocated to the partners, rather than through separate budget items.

No separate budget is foreseen for specific data management infrastructures such as repository subscriptions, as platforms such as Zenodo are available free of charge and meet Horizon Europe's requirements for open data and long-term archiving. The consortium may consider alternative or complementary repositories in the future, depending on accessibility, subject-specific standards or evolving requirements.

The responsibility for long-term archiving has not yet been assigned centrally and can be adapted depending on the formal requirements of the European Commission. If necessary, the consortium members will jointly take appropriate measures.

Financial oversight and coordination - including the resources required for FAIR data - will be the responsibility of the Project Coordinator (PC), the Project Management Board (PMB) and the General Assembly (GA) according to the governance structure defined in the consortium agreement.

5.3 Responsibility for Data Management

Responsibility for data management is distributed within the consortium in accordance with the governance structure defined in the consortium agreement.

Specific roles are assigned to:

- Work Package Leaders (WPLs) who oversee the implementation of data collection, documentation and quality control within their respective work packages;
- the Project Coordinator (PC), who monitors the timely delivery of data-related deliverables, ensures compliance with Horizon Europe requirements and reports to the Commission;
- The Project Management Board (PMB), which provides project-wide oversight and handles escalated data quality, access and compliance issues.

Responsibilities for data protection and GDPR compliance are shared among the partners in accordance with the Joint Controllership Agreement (JCA). Each partner organization is responsible for collecting, processing and storing the data it generates or processes in a way that complies with project protocols, ethics clearances and legal obligations.

This distributed but coordinated model ensures that data management across the project is consistent, compliant and in line with the FAIR principles and Horizon Europe's Open Science policy

5.4 Resources for Long-Term Preservation

The project will use existing infrastructures such as Zenodo to ensure the long-term archiving and accessibility of datasets made available for public re-use. Zenodo is a trusted, FAIR-compliant repository supported by the European Open Science Cloud (EOSC), providing persistent identifiers (DOIs), metadata support and open access features.

Only fully anonymized datasets that meet the ethical and legal standards of the project will be archived in this way. No specific resources or budget items are currently assigned exclusively for the long-term archiving infrastructure. However, the consortium may consider additional solutions, including institutional repositories if project requirements change.

The archiving strategy will be reviewed as part of the next DMP update and adapted to project developments and Horizon Europe's expectations.

6 Data security, ethics aspects of data management and other Data Management Procedures in EDU-LAB

6.1 Data Security Provisions & Breach Notification

The EDU-LAB project ensures the security of research data through a combination of technical and organizational measures that comply with the responsibilities defined in the Joint Controllership Agreement (JCA) and the project's internal data protection framework.

Data is stored on secure institutional servers and the project's internal Nextcloud platform. Access restrictions, password protection and encryption prevent unauthorized access and only authorized personnel within the consortium can access raw or identifiable data. Access rights are based on role and project requirements.

In the event of a data breach involving personal data, all partners must immediately inform their institutional Data Protection Officers (DPO) and the project coordinator. Data subjects and competent authorities will be notified within 72 hours in accordance with Article 33 GDPR and national law. The JCA establishes this joint accountability structure so that breaches are reported and dealt with appropriately across institutions.

In the event of a data breach, all partners must immediately inform their institutional Data Protection Officers (DPO) and the project coordinator. The JCA requires that all relevant parties are notified immediately and that reports are made to the relevant data protection authorities and data subjects within 72 hours in order to comply with Article 33 GDPR and national laws. In addition, the JCA emphasizes that cybersecurity incidents such as data breaches, must be handled in accordance with GDPR and other applicable laws and that controllers and processors bear full responsibility for breaches within their scope of activity.

Each partner organization is required to implement its own internal security policies and data protection procedures, in particular when handling sensitive or identifiable data such as audiovisual content, transcripts or socio-demographic profiles. These procedures comply with the safeguards described in the Privacy Policy and Data Protection Impact Assessment (DPIA).

The project's data management team, in collaboration with the Project Management Board, oversees the implementation and regular review of security procedures throughout the project cycle.

6.2 Safety and Certification of Repositories for Data Storage

Non-anonymised, raw data will be securely stored locally, in each partner’s host organisations’ safe server throughout the lifetime of the EDU-LAB project and retained as required by local or national regulations. WP4 Case Studies employ secure storage solutions on both physical and cloud platforms, with restricted access controls to protect data integrity and confidentiality. Additionally, backup policies and technical and organizational measures will be implemented following the Joint Controllership Agreement (Annex F of the EDU-LAB Consortium Agreement) to ensure the confidentiality, integrity, and availability of project data.

Anonymized project data will be securely archived in the certified repository ZENODO for a minimum of five years after the project's completion, ensuring long-term accessibility and compliance with preservation standards.

Each dataset deposited in ZENODO will include structured metadata and a persistent Digital Object Identifier (DOI), in line with FAIR data principles and Horizon Europe’s Open Science requirements.

6.3 Impact of Ethical and Legal Issues on Data Sharing

The EDU-LAB project collects and processes personal and sensitive data, including data from minors aged 15-17. Accordingly, ethical and legal considerations govern the entire data sharing framework - always in full compliance with the General Data Protection Regulation (GDPR), national legislation and Horizon Europe's ethical standards.

Compliance with the GDPR and data protection legislation

All data processing operations follow the GDPR (Regulation (EU) 2016/679), Regulation (EU) 2018/1725 and relevant national laws. The Consortium implements the principles enshrined in Article 5 GDPR - lawfulness, purpose limitation, data minimization, transparency, integrity and accountability. Legal basis: Article 6(1)(a) and Article 9(2)(j) GDPR (explicit, informed consent for scientific research purposes). Data minimization will ensure that only the data necessary for the project objectives is collected, stored and used.

Protection of minors and vulnerable participants

As young people (15 - 17 years) are involved, increased ethical safeguards apply:

- Dual consent (from minors and their legal representatives)
- Information sheets in the respective national languages
- Right of withdrawal at any time, exercisable by participants or legal guardians
- Anonymization or pseudonymization of sensitive data (opinions, socio-demographics, audio/video).

These precautions are anchored in the Ethics Issues Management Plan (Deliverable D7.1) and ensure voluntariness, risk minimization and the comprehensive protection of minors.

Joint controllership and shared responsibility

A Joint Controllership Agreement (JCA) in accordance with Article 26 GDPR defines the roles and obligations of all partners:

- Specifying access and usage rights
- Clarifying responsibilities for the protection of participant data
- Defining reporting mechanisms for data protection violations

Each partner institution ensures that the data processed on site is stored securely, anonymized (and/or pseudonymized as feasible) if necessary and only made accessible to authorized personnel. Compliance is monitored by the respective data protection officers (DPOs).

Restrictions on data sharing

- Only fully anonymized, ethically approved data sets are made publicly available (e.g. via Zenodo)
- No sharing of personal or identifiable data outside the consortium.
- Sensitive content (biometric audio/video recordings, qualitative transcripts) remains on internal servers under strictly controlled conditions.
- International data transfers outside the EEA will only take place, if necessary, with appropriate safeguards (e.g. standard contractual clauses).

Embargo periods or staggered releases may be necessary to comply with publication schedules or data protection concerns.

Oversight and review

- The Ethics Issues Group (EIG) oversees all ethical aspects of data collection, processing and disclosure.
- In cooperation with the DPOs, the EIG reviews consent processes, participant protection and anonymization in compliance with data protection regulations.
- Regular internal audits and - if necessary - consultations with national data protection authorities ensure that the project always complies with EU and national law.

- Results and measures are incorporated into deliverables D7.1 (Ethics Plan) and D7.2 (Data Management Plan).

This integrated approach guarantees that EDU-LAB meets both the FAIR principles and the high ethical and legal requirements of Horizon Europe.

6.4 Informed Consent for Data Sharing and Preservation

Informed consent is a fundamental requirement for all data collection activities involving personal data in the EDU-LAB project. It ensures that participants fully understand the purpose, usage, sharing, and long-term preservation of their data, and that their rights under GDPR are protected, including the right to access, rectification, erasure, and withdrawal of consent at any time without consequences.

The informed consent process includes the following measures:

- **Comprehensive Informed Consent Forms (ICF):** Provided in local languages, these forms explain the purpose, usage, storage, sharing, and long-term preservation of personal data. Special provisions are made for minors, requiring parental or guardian consent to ensure legal compliance and ethical integrity. An initial contact form (accent form) gauges participant interest before seeking guardian consent.
- **Transparency and Participant Rights:** Information Sheets for Participants (ISP) ensure participants understand the study's objectives, scope, data management procedures, and their rights. Consent forms specify a data retention period of at least five years post-project completion and inform participants about the potential for data to be preserved in anonymized form for long-term research.
- **Ethical Oversight and Documentation:** The Ethical Issues Group (EIG) monitors and documents all informed consent processes to ensure GDPR compliance and alignment with Horizon Europe's ethical guidelines. The Ethics Issues Management (EIM) document serves as a reference, ensuring consistency with the ethics chapter in the Description of the Action (DoA).

This approach underpins EDU-LAB's commitment to ethical data management and participant rights protection, as outlined in Ethics Plan D7.1.

6.5 Use of Other Data Management Procedures

The EDU-LAB project is governed by the national and sector-specific data management policies in place at each participating institution, provided they are compatible with the standards and responsibilities of this data management plan and the FAIR principles.

Partners may apply their institutional procedures for:

- Data storage and backup (e.g. institutional servers or certified repositories)
- Ethics clearance and oversight
- Documentation, file versioning and long-term archiving
- Data protection protocols in accordance with GDPR and national legislation

The partners' internal procedures are subject to coordination and review by the project's data governance structure, including the Ethics Issues Group (EIG) and the partners' Data Protection Officers (DPOs). Where institutional systems are used, each partner remains responsible for ensuring that its practices support compliance with the GDPR, Horizon Europe requirements and the Joint Controllership Agreement (JCA).

This approach allows for institutional flexibility while ensuring a consistent and secure framework for data management across the EDU-LAB consortium.

Annex – Overview of Data Management Roles and Responsibilities

Role / Body	Responsibilities
Work Package Leaders (WPLs)	Coordinate all data-related activities in their WPs (e.g. collection, cleaning, consent implementation, documentation, and reporting). Provide quarterly status reports and support deliverables.
Project Coordinator (PC)	Oversees overall data governance. Coordinates DMP updates, chairs PMB and GA, transmits reports to EC, ensures quality of deliverables. Central contact for data issues escalated beyond WP level.
Project Management Board (PMB)	Exercises executive control over project execution, monitors deliverable quality and timelines, ensures ethical coordination, and mediates on methodological/data-related issues.
General Assembly (GA)	Top decision-making body for structural, budgetary, and legal matters (including consortium composition, ethics bodies, and policy changes). May define high-level data strategy or respond to escalated disputes.
Ethics Issues Group (EIG)	Monitors compliance with ethical and data protection standards (GDPR, informed consent, anonymisation). Oversees ethical risks in fieldwork and EU/non-EU data cooperation.
Institutional Data Protection Officers (DPOs)	Ensure local GDPR compliance, especially in cases of personal data processing, breach notification, and consent storage. First point of contact for data protection queries.
Case Studies Coordination Group (CSCG)	Ensures uniformity of WP3–WP4 fieldwork implementation, conducts quality control, and coordinates reporting across national teams.
Integration and Dissemination Panel (IDP)	Coordinates dissemination of anonymised datasets and research outputs. May advise on accessibility and public sharing formats.
Team Leaders (TLs)	Act as institutional points of contact. Responsible for ensuring their team complies with assigned tasks, including administrative, scientific, and data-related duties.