AI4ED

TOWARDS AN AI DRIVEN EDUCATIONAL PROCESS INTEGRATING MODERN CAREERS IN THE EDUCATIONAL SYSTEM

Deliverable

D3.2 – Data Management Plan

Deliverable Lead: UBremen
Deliverable due date: 31/07/23
Actual submission date: 30/08/23
Dissemination level: PU
Version: COMPLETED

Co-funded by the European Union

This project has received funding from the European Union's Erasmus + programme under grant agreement No 101087543
Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

The information contained in this report is subject to change without notice and should not be construed as a commitment by any members of the AI4ED Consortium. The information is provided without any warranty of any kind.

Document Control Page	
Title	Data Management Plan
Creator	UBremen
Description	The deliverable “Data Management Plan” outlines how the research data collected or generated will be handled during the project AI4ED. It describes which standards and methodology for data collection and generation will be followed, and whether and how the data will be shared.
Contributors	All partners
Creation date	04/04/2023
Type	Report
Language	English
Audience	☒ public ☐ confidential
Review status	☐ Draft ☐ Assigned Reviewer accepted ☒ Coordinator accepted
Action requested	☐ to be revised by the Assigned Reviewer ☐ for approval by the Project Coordinator ☐ for acknowledgement by Partners

Revision history

Version	Author(s)	Changes	Date
V 0.8	UBremen	Draft creation Heike Thöricht	05/04/2023
V 0.9	UBremen	Document reviewed and adapted by Vivian Harberts	20/04/2023
V 1.0	UBremen	Document reviewed by consortium and adapted by UBremen	08/08/2023

Table of Contents

List of figures

List of tables

Glossary

API	Application Programming Interface
CCO	Creative Commons
CSV	Comma- separated value
DOI	Digital Object Identifier
GDPR	General Data Protection Regulation
DGA	The European Data Governance Act
DSA	Data Service Act
MFA	Multi-factor Authentication
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting

EXECUTIVE SUMMARY / ABSTRACT

Abstract	This document is the first version of the data management plan (DMP) for the project AI4ED. The DMP provides descriptions of the data created and used in the project and provides an outline of how the data is managed, shared, and preserved. This DMP describes the steps undertaken by the Consortium to make the data generated by AI4ED to be findable, accessible, interoperable and re-usable (FAIR). The DMP is framed in accordance to the “Guidelines on FAIR Data Management in Horizon 2020”, which lists the following points that are to be addressed by a DMP that promotes FAIR data: • the handling of research data during and after the end of the project, • what data will be collected, processed and/or generated • which methodology and standards will be applied • whether data will be shared/made open access, and • how data will be curated and preserved (including after the end of the project). As a DMP is a living document, the document will be updated regularly until the end of the project by its authors.
Keywords	AI, data lifecycle, data structure, data management

1 Introduction

The data management plan ('DMP') follows the template provided by the European Commission. This document was established in April 2023 and will be as a living document regularly updated until the end of the project AI4ED by the author of the DMP (e.g. in case of new data, changes in consortium policies, changes in consortium composition and external factors).

2 Data Summary

Research data will be collected for four use cases in work package 5 “Toolkit Development and Use Case Implementation” in order to develop AI models. This will be realized by trainings in the learning management systems Moodle and eAssistant and pre- and post-surveys of students by the consortium partners IMH, UBREMEN, CENFIM and SCSKZ. Additionally, IMH will use administration data (e.g. for the enrolment of the student, previous studies or qualifications), and will collect ratings about student’s abilities in different competences in interviews and results from psychometric tests.

Key performance indicators (KPIs) have been defined by the consortium partners in work package 2 “Definition of the AI strategy in educational processes” upfront, such as consumed content in the learning management system, percentage of completed assignments, attendance to the sessions, students’ satisfaction, grades obtained, student/apprentice's personal data (previous education, grades).

The raw data will be collected by the above mentioned consortium partners, anonymized and transferred to ALCHEMYML which is in charge of processing the data and developing models. The data collecting consortium partners are required to respect the data protection regulations of the own country because regulations can differ and they will collect informed consent from the participants of the trainings.

For harmonization of the datasets and as preparation for the data transfer from research institutions to ALCHEMYML, the company provided an xlsx file for discussion with information for each variable. For example:

FIELD	**educ_level**
DESCRIPTION	Which is the student's last educational level?
INPUT TYPE	String
INPUT OPTIONS	not_formal | primary | lower_secondary | upper_secondary | vocational | undergraduate | postgraduate | doctorate | other | NULL

On base of the pre-processed data there will be three AI models developed:

1. Active Learning

   This model recommends new educational actions that individuals can enrol in, adapting to the individual needs of the students.

2. Personalised Tutoring

   This model is a content recommendation system that personalizes students' learning experiences.

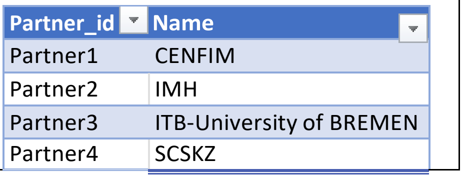
3. Drop-out prevention

   This model aims to calculate the risk of student dropout, allowing educational institutions to intervene early and provide additional support and retention strategies to at-risk students.

In this section, issues about the type, format, origin, and size of the data are addressed as well as on aspects of data re-use and utility. Detailed information on the different data is provided by each partner in Table 1 below. Given information from ALCHEMYML will be displayed differently to the rest of the partners from the use cases, as their role is different and hence their data management looks different.

Table 1 - Summary of the type, format, origin, and size of data for each beneficiary

Data set number	Consortium partner	Title of data	Short description of data
1	IMH	Personal data of students in IMH Campus	Data collected from students that is needed to be collected for student's enrolment
2	IMH	Data collected from student's interview	The interviewer numerically rates each student's abilities in the different attitudinal competencies (soft skills)
3	IMH	Data collected from the psychometric test performed by the students in IMH Campus	Psycho-technical test that measures certain competences of the student
4	IMH	Previous studies qualifications	The marks of the last training you have completed prior to the completion of the current training will be collected.
5	IMH	Data generated in Moodle	Data on different elements: • Surveys of different types: Pre-training survey (interests, expectations, academic level, motivation...); Survey on the evaluation of the contents by the trainees of the contents and of the training; Survey on the evaluation of the level of active learning existing in the training post-training survey • Grades obtained by the students in the different activities carried out.
6	UBREMEN	Pre-training survey data of students of vocational education and training (VET) schools in Bremen	Socio-demographic survey data of the participating VET students
7	UBREMEN	Post-training survey data of students of VET schools in Bremen	Evaluation survey data of participating students after the training, e.g. about satisfaction with structure and content of learning platform, user friendliness
8	UBREMEN	Data generated in Moodle	Data about students' navigation patterns within the learning management system, drop out points and time frames for working on modules (key performance indicators)
9	CENFIM	Personal data from students during enrolment	General personal data collected from LMS platform for enrolment purposes

<table>
<tr><td>10</td><td>CENFIM</td><td>Survey/ interview data of students</td><td>The interviewer measures each student's knowledge, technical skills and attitudes/behaviours.</td></tr>
<tr><td>11</td><td>CENFIM</td><td>Data generated Moodle</td><td>Grades obtained by the students in the different activities carried out.</td></tr>
<tr><td>12</td><td>CENFIM</td><td>Evaluation survey data</td><td>Post-training survey from Moodle</td></tr>
<tr><td>13</td><td>SCSKZ</td><td>Data generated in eAssistant and Teams</td><td>Personal data, data on class attendance, grades, praise... will be obtained from eAssistant, and data on school work in certain modules - groups from MS Teams.</td></tr>
<tr><td>14</td><td>ALCHEMYML</td><td>Data of the different partners and models</td><td>Data comes from different sources of information (partner id), but these data will be sent in a standardized way. In terms of their typology, data will consist of data collected from survey data and data from the Learning Management System (LMS) that contain anonymized information about the students.

<table>
<tr><th>Partner_id</th><th>Name</th></tr>
<tr><td>Partner1</td><td>CENFIM</td></tr>
<tr><td>Partner2</td><td>IMH</td></tr>
<tr><td>Partner3</td><td>ITB-University of BREMEN</td></tr>
<tr><td>Partner4</td><td>SCSKZ</td></tr>
</table></td></tr>
</table>

3 Data collection tables of information of each dataset

In order to collect all relevant information for each dataset, the following table was provided to the consortium partners. The topics addressed in Table 2 include all relevant information, such as dataset responsible partner, definition of data formats, provisions for making the data FAIR, security and ethical aspects.

Table 2 - Template to collect relevant information on data to be used

[number of dataset]– [title of data]	
[Responsible partner]	
[Data ownership]	
Data formats	
Data size (estimated)	
Data collection: start – end Data processing & analysis: start – end	
Personal data protection: are they personal data? If so, have you gained (written) consent from data subjects to collect this information?	
Data sharing, re-use, distribution, publication (How?)	
Providing documentation needed to validate data analysis and facilitate data re-use (How?)	
Which tools, software, technologies or processes are used to process and analyse the data?	
Is documentation about relevant software needed to use the data?	
Data storage - Where? For how long? How to back up?	

Next tables show the answers from each partner.

3.1 Fabrikazio Aurreratuaren eta Digitalaren Campusa – IMH

Table 3 - IMH - Personal data of students in IMH Campus

1 – Personal data of students in IMH Campus	
Responsible partner: IMH Campus	
Data ownership: IMH Campus	
Data formats	Structured data will be exported in tables, i.e. CSV
Data size (estimated)	50Kb
Data collection: start – end Data processing & analysis: start – end	July 2023 – September 2024 September 2023 – October 2024
Personal data protection: are they personal data? If so, have you gained (written) consent from data subjects to collect this information?	Data contains personal data and a consent will be collected.
Data sharing, re-use, distribution, publication	Anonymized data will be published on Zenodo

(How?)	for reuse of data
Providing documentation needed to validate data analysis and facilitate data re-use (How?)	Yes, codebook
Which tools, software, technologies or processes are used to process and analyse the data?	Analyse of the survey data will be executed by AI applications from ALCHEMYML and Moodle application with "learning analytics" plug in for data gathering
Is documentation about relevant software needed to use the data?	No
Data storage - Where? For how long? How to back up?	Data will be stored on a local server of IM Campus for 10 years. Backups are replicated on other servers, that are protected by MFA authentication and hardened by firewall rules.

Table 4 - IMH - Data collected from student's interview

2 – Data collected from student's interview	
Responsible partner: IMH Campus	
Data ownership: IMH Campus	
Data formats	Structured data will be exported in tables, i.e. CSV
Data size (estimated)	100Kb
Data collection: start – end Data processing & analysis: start – end	July 2023 – September 2024 September 2023 – October 2024
Personal data protection: are they personal data? If so, have you gained (written) consent from data subjects to collect this information?	Data contains personal data and a consent will be collected.
Data sharing, re-use, distribution, publication (How?)	Anonymized data will be published on Zenodo for reuse of data
Providing documentation needed to validate data analysis and facilitate data re-use (How?)	Yes, codebook
Which tools, software, technologies or processes are used to process and analyse the data?	Analyse of the survey data will be executed by AI applications from ALCHEMYML and Moodle application with "learning analytics" plug in for data gathering
Is documentation about relevant software needed to use the data?	No
Data storage - Where? For how long? How to back up?	Data will be stored on a local server of IM Campus for 10 years. Backups are replicated on other servers, that are protected by MFA authentication and hardened by firewall rules.

Table 5 - IMH - Data collected from the psychometric test performed by the students in IMH Campus

3 – Data collected from the psychometric test performed by the students in IMH Campus	
Responsible partner: IMH Campus	
Data ownership: IMH Campus	
Data formats	Structured data will be exported in tables, i.e.

	CSV
Data size (estimated)	100Kb
Data collection: start – end Data processing & analysis: start – end	July 2023 – September 2024 September 2023 – October 2024
Personal data protection: are they personal data? If so, have you gained (written) consent from data subjects to collect this information?	Data contains personal data and a consent will be collected.
Data sharing, re-use, distribution, publication (How?)	Anonymized data will be published on Zenodo for reuse of data
Providing documentation needed to validate data analysis and facilitate data re-use (How?)	Yes, codebook
Which tools, software, technologies or processes are used to process and analyse the data?	Analyse of the survey data will be executed by AI applications from ALCHEMYML and Moodle application with "learning analytics" plug in for data gathering
Is documentation about relevant software needed to use the data?	No
Data storage - Where? For how long? How to back up?	Data will be stored on a local server of IM Campus for 10 years. Backups are replicated on other servers, that are protected by MFA authentication and hardened by firewall rules.

Table 6 - IMH - Previous studies qualifications

4 – Previous studies qualifications	
Responsible partner: IMH Campus	
Data ownership: IMH Campus	
Data formats	Structured data will be exported in tables, i.e. CSV
Data size (estimated)	100Kb
Data collection: start – end Data processing & analysis: start – end	July 2023 – September 2024 September 2023 – October 2024
Personal data protection: are they personal data? If so, have you gained (written) consent from data subjects to collect this information?	Data contains personal data and a consent will be collected.
Data sharing, re-use, distribution, publication (How?)	Anonymized data will be published on Zenodo for reuse of data
Providing documentation needed to validate data analysis and facilitate data re-use (How?)	Yes, a Codebook is going to be provided as data is anonymized.
Which tools, software, technologies or processes are used to process and analyse the data?	Analyse of the survey data will be executed by AI applications from ALCHEMYML and Moodle application with "learning analytics" plug in for data gathering
Is documentation about relevant software needed to use the data?	No
Data storage - Where? For how long? How to back up?	Data will be stored on a local server of IM Campus for 10 years. Backups are replicated

	on other servers, that are protected by MFA authentication and hardened by firewall rules.

Table 7 - IMH - Data generated in Moodle

5 – Data generated in Moodle	
Responsible partner: IMH Campus	
Data ownership: IMH Campus	
Data formats	Structured data will be exported in tables, i.e. CSV
Data size (estimated)	5Mb
Data collection: start – end Data processing & analysis: start – end	July 2023 – September 2024 September 2023 – October 2024
Personal data protection: are they personal data? If so, have you gained (written) consent from data subjects to collect this information?	Data contains personal data and a consent will be collected.
Data sharing, re-use, distribution, publication (How?)	Anonymized data will be published on Zenodo for reuse of data
Providing documentation needed to validate data analysis and facilitate data re-use (How?)	Yes, codebook
Which tools, software, technologies or processes are used to process and analyse the data?	Moodle application with “learning analytics” plug in for data gathering
Is documentation about relevant software needed to use the data?	No
Data storage - Where? For how long? How to back up?	Data will be stored on a local server of IM Campus for 10 years. Backups are replicated on other servers, that are protected by MFA authentication and hardened by firewall rules.

3.2 University of Bremen – UBREMEN

Table 8 - UBREMEN - Pre-training survey data of students of vocational education and training (VET) schools in Bremen

6 – Pre-training survey data of students of vocational education and training (VET) schools in Bremen	
Responsible partner: ITB, University of Bremen	
Data ownership: ITB, University of Bremen	
Data formats	Structured data will be exported in tables, i.e. CSV
Data size (estimated)	30kb
Data collection: start – end Data processing & analysis: start – end	June 2023 – September 2024 September 2024 – October 2024
Personal data protection: are they personal data? If so, have you gained (written) consent from data subjects to collect this information?	The data is fully anonymized from the beginning. Each user of the learning platform is assigned an anonymous generated e-mail address and a generated password. The users can then log in to the Moodle site with this data and fill in the survey. However, for reasons of school law, written consent is required and obtained from the individuals concerned. In addition, the survey must be approved by the Institute for Quality

	Development in the State of Bremen (IQHB). To ensure compliance with all data protection regulations, the survey is also sent in advance to the deputy data protection officer of the Legal Service and the Legal Department of the University of Bremen.
Data sharing, re-use, distribution, publication (How?)	Anonymized data will be published on Zenodo for reuse of data
Providing documentation needed to validate data analysis and facilitate data re-use (How?)	Yes, codebook
Which tools, software, technologies or processes are used to process and analyse the data?	The analysis of the survey data is performed with the help of AI applications from ALCHEMYML. Likewise, the data that is automatically tracked when using the Moodle learning platform is used for data collection. The plug-in "Learning Analytics" can be integrated additionally for data collection if necessary.
Is documentation about relevant software needed to use the data?	No
Data storage - Where? For how long? How to back up?	Data will be stored on a local server of the ITB for 10 years. Back-up on another server in encrypted format is provided.

Table 9 - UBREMEN - Post-training survey data of students of VET schools in Bremen

7 – Post-training survey data of students of VET schools in Bremen	
Responsible partner: ITB, University of Bremen	
Data ownership: ITB, University of Bremen	
Data formats	Structured data will be exported in tables, i.e. CSV
Data size (estimated)	30kb
Data collection: start – end Data processing & analysis: start – end	June 2023 – September 2024 September 2024 – October 2024
Personal data protection: are they personal data? If so, have you gained (written) consent from data subjects to collect this information?	The data is completely anonymized from the start. Each user of the learning platform is assigned an anonymously generated e-mail address and password. With this data, users can then log in to the Moodle site and fill in the survey. However, for reasons of school law, written consent from the data subjects is required and must be obtained. In addition, the survey must be approved by the Institute for Quality Development in the State of Bremen (IQHB). To ensure that all data protection regulations are complied with, the survey is also sent in advance to the deputy data protection officer of the Legal Service and the Legal Department of the University of Bremen.
Data sharing, re-use, distribution, publication (How?)	Anonymized data will be published on Zenodo for reuse of data
Providing documentation needed to validate data analysis and facilitate data re-use (How?)	Yes, codebook
Which tools, software, technologies or processes are used to process and analyse the data?	The survey data is evaluated with the help of AI applications from ALCHEMYML. Likewise, the data

	that is automatically tracked when using the Moodle learning platform is used for data collection. The plug-in "Learning Analytics" can be additionally integrated for data collection if required. The final evaluation survey is intended to identify possible improvements to the learning unit, both in terms of content and handling. The satisfaction of the users in dealing with the learning unit should also become apparent.
Is documentation about relevant software needed to use the data?	No
Data storage - Where? For how long? How to back up?	Data will be stored on a local server of the ITB for 10 years. Back-up on another server in encrypted format is provided.

Table 10 - UBREMEN - Data generated in Moodle

8 – Data generated in Moodle	
Responsible partner: ITB, University of Bremen	
Data ownership: ITB, University of Bremen	
Data formats	Structured data will be exported in tables, i.e. CSV
Data size (estimated)	2-5 MB per user, if compressed size will be significantly less
Data collection: start – end Data processing & analysis: start – end	November 2023 – September 2024 September 2024 – October 2024
Personal data protection: are they personal data? If so, have you gained (written) consent from data subjects to collect this information?	No personal data, as all data is anonymized. Each user of the learning platform is assigned an anonymously generated e-mail address and password. With this data, users can then log in to the Moodle site and work on the learning units. However, due to school laws, a written consent from the data subjects is necessary and will be collected. Apart from that the survey needs approval from the Institute for Quality Development in the State of Bremen (IQHB).
Data sharing, re-use, distribution, publication (How?)	Data will be shared with partners and published on Zenodo for re-use and distribution
Providing documentation needed to validate data analysis and facilitate data re-use (How?)	Yes, codebook
Which tools, software, technologies or processes are used to process and analyse the data?	AI software of ALCHEMYML and “learning analytics” plug in of Moodle
Is documentation about relevant software needed to use the data?	No, data will be provided with standard software like Excel
Data storage - Where? For how long? How to back up?	Data will be stored on a local server of the ITB for 10 years. Back-up on another server in encrypted format is provided.

3.3 Centro de Formacao professional da industria metalurgica e metalomecanica - CENFIM

Table 11 - CENFIM - Personal data from students during enrolment

9 – Personal data from students during enrolment	
Responsible partner: CENFIM	
Data ownership: CENFIM	
Data formats	Structured data will be exported in tables, i.e. CSV
Data size (estimated)	30 kb
Data collection: start – end Data processing & analysis: start – end	Sep 23 – Sep 24 Sep 23 – Oct 24
Personal data protection: are they personal data? If so, have you gained (written) consent from data subjects to collect this information?	Personal data will only be collected after consent has been given.
Data sharing, re-use, distribution, publication (How?)	Only anonymized data will be published on Zenodo.
Providing documentation needed to validate data analysis and facilitate data re-use (How?)	Yes, codebook
Which tools, software, technologies or processes are used to process and analyse the data?	Analysed survey data will be executed by AI applications from ALCHEMYML
Is documentation about relevant software needed to use the data?	No
Data storage - Where? For how long? How to back up?	Data will be stored on a CENFIM server for 10 years and replicated to another internal backup server on a different location.

Table 12 – CENFIM - Data collected from student's interview

10 – Data collected from student's interview	
Responsible partner: CENFIM	
Data ownership: CENFIM	
Data formats	Structured data will be exported in tables, i.e. CSV
Data size (estimated)	30 kb
Data collection: start – end Data processing & analysis: start – end	Sep 23 – Sep 24 Sep 23 – Oct 24
Personal data protection: are they personal data? If so, have you gained (written) consent from data subjects to collect this information?	Personal data will only be collected after consent has been given.
Data sharing, re-use, distribution, publication (How?)	Anonymized data will be published on Zenodo for reuse of data
Providing documentation needed to validate data analysis and facilitate data re-use (How?)	Yes, a Codebook is going to be provided as data is anonymized.
Which tools, software, technologies or processes are used to process and analyse the data?	Analyse of the survey data will be executed by AI applications from ALCHEMYML and Moodle application with "learning analytics" plug in for data gathering
Is documentation about relevant software needed to use the data?	No

Data storage - Where? For how long? How to back up?	Data will be stored on a local server of CENFIM for 10 years and replicated to another internal backup server on a different location.

Table 13 - CENFIM - Data generated in Moodle

11 – Data generated in Moodle	
Responsible partner: CENFIM	
Data ownership: CENFIM	
Data formats	Structured data will be exported in tables, i.e. CSV
Data size (estimated)	30 kb
Data collection: start – end Data processing & analysis: start – end	Sep 23 – Sep 24 Sep 23 – Oct 24
Personal data protection: are they personal data? If so, have you gained (written) consent from data subjects to collect this information?	Personal data will only be collected after consent has been given.
Data sharing, re-use, distribution, publication (How?)	Anonymized data will be published on Zenodo for reuse of data
Providing documentation needed to validate data analysis and facilitate data re-use (How?)	Yes, codebook
Which tools, software, technologies or processes are used to process and analyse the data?	Analyse of the survey data will be executed by AI applications from ALCHEMYML and Moodle application with “learning analytics” plug in for data gathering
Is documentation about relevant software needed to use the data?	No
Data storage - Where? For how long? How to back up?	Data will be stored on a local server of CENFIM for 10 years and replicated to another internal backup server on a different location.

Table 14 - CENFIM - Final Evaluation

12 – Final evaluation	
Responsible partner: CENFIM	
Data ownership: CENFIM	
Data formats	Structured data will be exported in tables, i.e. CSV
Data size (estimated)	30 kb
Data collection: start – end Data processing & analysis: start – end	Sep 23 – Sep 24 Sep 23 – Oct 24
Personal data protection: are they personal data? If so, have you gained (written) consent from data subjects to collect this information?	Personal data will only be collected after consent has been given.
Data sharing, re-use, distribution, publication (How?)	Anonymized data will be published on Zenodo for reuse of data
Providing documentation needed to validate data analysis and facilitate data re-use (How?)	Yes, codebook

Which tools, software, technologies or processes are used to process and analyse the data?	Analyse of the survey data will be executed by AI applications from ALCHEMYML and Moodle application with "learning analytics" plug in for data gathering
Is documentation about relevant software needed to use the data?	No
Data storage - Where? For how long? How to back up?	Data will be stored on a local server of CENFIM for 10 years and replicated to another internal backup server on a different location.

3.4 Šolski center Slovenske Konjice-Zreče – SCSKZ

Table 15 - SCSKZ - Data generated in eAsistent and Teams

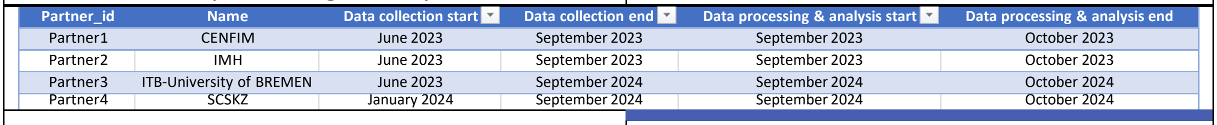
13 – Data generated in eAsistent and Teams	
Responsible partner: SCSKZ	
Data ownership: SCSKZ	
Data formats	Structured data will be exported in tables, i.e. CSV
Data size (estimated)	35 kB
Data collection: start – end Data processing & analysis: start – end	September 2023 – October 2024 October 2024- October 2025
Personal data protection: are they personal data? If so, have you gained (written) consent from data subjects to collect this information?	Data will be anonymized from beginning on. However, due to school laws, a written consent from the data subjects is necessary and will be collected.
Data sharing, re-use, distribution, publication (How?)	Anonymised data will be published on Zenodo for reuse of data
Providing documentation needed to validate data analysis and facilitate data re-use (How?)	Yes, a Codebook is going to be provided
Which tools, software, technologies or processes are used to process and analyse the data?	Analyse of the surveys will be executed by AI applications from ALCHEMY ML
Is documentation about relevant software needed to use the data?	No
Data storage - Where? For how long? How to back up?	Data will be stored on a local server for 10 years and replicated to another internal backup server on a different location

3.5 Alchemy Machine Learning - ALCHEMYML

Table 16 - Partner_id survey data

14 – Partner_id survey data	
Responsible partner: ALCHEMYML	
Data ownership: Partner_id	
Data formats	JSON, CSV
Data size (estimated)	Survey1: 3KB per observation (per student) Sample of 300 students Total estimated: 1,5MB. Survey2:

	3kb per observation (per student) Sample of 300 students Total estimated: 1,5MB. LMS_data: One course data
Data collection, processing & analysis: start – end	



Partner_id	Name	Data collection start	Data collection end	Data processing & analysis start	Data processing & analysis end
Partner1	CENFIM	June 2023	September 2023	September 2023	October 2023
Partner2	IMH	June 2023	September 2023	September 2023	October 2023
Partner3	ITB-University of BREMEN	June 2023	September 2024	September 2024	October 2024
Partner4	SCSKZ	January 2024	September 2024	September 2024	October 2024

Personal data protection: are they personal data? If so, have you gained (written) consent from data subjects to collect this information?	We will work with anonymized data that has the consent of the individuals involved and whose generation and extraction are approved by the competent authorities.
Data sharing, re-use, distribution, publication (How?)	Anonymized data will be published on Zenodo for reuse of data.
Providing documentation needed to validate data analysis and facilitate data re-use (How?)	Datasheets for datasets (Gebru; Morgenstern; Vecchione; Wortman Vaughan; Wallach; Daumé; Crawford. 2018) & model cards for the models as documentation (Mitchell; Wu; Zaldivar; Barnes; Vasserman; Hutchinson; Spitzer; Deborah Raji; Gebru. 2019)
Which tools, software, technologies or processes are used to process and analyse the data?	Basic office tools such as Excel, Word, or Acrobat Reader will be used to open and inspect the content of the most basic documents provided by each partner. For data analysis, Python will be used as the programming language along with its most well-known libraries for data analysis such as: Pandas, Stats models, Matplotlib, Seaborn, etc. Regarding the construction of predictive models, the Scikit-learn library along with PyCaret will be used to outline the models. The developed models will be exposed through an API with the following endpoints: • An endpoint that collects the current status data of a course on a weekly basis and makes predictions about the risk of dropout. • An endpoint that collects the current status data of a course on a weekly basis and makes predictions to recommend content. • An endpoint to evaluate the performance of the models and retrain them if necessary.
Is documentation about relevant software needed to use the data?	The documentation needed is related to the API to be built so that ALCHEMYMLs system can communicate with others partners data. In fact, the partners should have the knowledge to use the API.

Data storage - Where? For how long? How to back up?	Data will be stored on a private cloud hosted server for 10 years. We will backup our projects code in Bitbucket. Data will be backed up in both the server used to develop the project and our storage system (a NAS).

4 FAIR data

4.1 Making data findable, including provisions for metadata

Open data will be preserved and shared in the repository Zenodo.org. Zenodo registers DOIs (via DataCite) for all deposited records. The repository applies the DataCite Metada scheme. All metadata in Zenodo may be freely used under the CC0 waiver and can be exported via OAI-PMH and harvested.

The bibliographic metadata include all of the following:

- the terms “Erasmus+”;
- the name of the project “Towards an AI driven educational process integration modern careers in the educational system”, acronym “AI4ED” and grant number “101087543”;
- the publication date, and length of embargo period if applicable
- a persistent identifier

For AI4ED, the following deposition metadata fields are mandatory:

- Creators: Main researchers involved in producing the data (including their ORCID id)
- Title: Title of the deposition
- Description: Abstract or description for deposition
- Upload type: Type of the deposition from a controlled vocabulary (publication, dataset, software, ...).
- Publication date: Date of publication in ISO8601 format (YYYY-MM-DD).
- License: License for the reuse of data
- Keywords: Free form keywords for the deposition. If applicable, keywords from the CESSDA ELSST Thesaurus. For example: e-learning, learning, artificial intelligence, training, learning management system
- Related identifiers: Persistent identifiers of related publications, datasets and software
- Language: The primary language of the resource

4.2 Making data accessible

This section aims at clarifying which AI4ED research data are made fully accessible and/or at justifying if any restriction to data accessibility is deemed necessary, including the rationale behind it. It describes the data that are made accessible, the methods or software tools needed to access them, where to find the data, and other relevant issues.

In each country access control to information is the responsibility of the individual organizations involved in conducting the data collection studies.

Open data from all partners will be preserved and shared in the repository Zenodo.org which is registered in the Directory of Open Access Repositories (OpenDOAR) and is one of four repository which fulfils all requirements for Europe Horizon 2020 projects (Jahn, Najko, Laakso, Mikael, Lazzeri, Emma, & McQuilton, Peter (2023). Zenodo is open to all research outputs from all fields of science regardless of funding source and registers DOIs (via DataCite) for all deposited records. All metadata in Zenodo may be freely used under the CCO waiver. The items and the metadata will be retained for the lifetime of the repository.

However, apart from publishing project results at Zenodo, different rules apply for data within each institution:

Table 17 - Data rules for each beneficiary

IMH	All anonymized data will be openly available as necessary written consents have been obtained before data collection and processing and anonymization will take place from the beginning on, hence, no personal data will be published and no legal reasons for not publishing apply. Structured data will be in a table (i.e. CSV), there will be no other access protocol due to security reasons which are hacking prevention, misuse of datasets, lack of citations of data origin. Data retrieval is mediated without specialised or proprietary tools or communication methods. There is no need for a data access committee.
UBREMEN	All anonymized data will be openly available as necessary written consents have been obtained before data collection and processing and anonymization will take place from the beginning on, hence, no personal data will be published and no legal reasons for not publishing apply. Structured data will be in a table (i.e. Excel), there will be no other access protocol due to security reasons which are hacking prevention, misuse of datasets, lack of citations of data origin. Data retrieval is mediated without specialized or proprietary tools or communication methods. There is no need for a data access committee.
CENFIM	All anonymized data will be openly available as necessary written consents have been obtained before data collection and processing and anonymization will take place from the beginning on, hence, no personal data will be published and no legal reasons for not publishing apply. Structured data will be in a table (i.e. CSV or Excel), there will be no other access protocol due to security reasons which are hacking prevention, misuse of datasets, lack of citations of data origin. Data retrieval is mediated without specialised or proprietary tools or communication methods. There is no need for a data access committee.
SCSKZ	All anonymized data will be publicly available, as the necessary written consents will be obtained before the collection and processing of data and anonymization from the beginning, so personal data will not be published. Structured data will be in a table (i.e. CSV; imports from E assistant or MS Teams tools). Data retrieval is mediated without specialised or proprietary tools or communication methods. There is no need for a data access committee.
ALCHEMYML	The data we will receive will be anonymous, so it's possible for the data to be publicly available. Access to the data, will be given by uploading the final versions of the data, models and documentation to Zenodo as well as to Github. For uploaded data on Zenodo, there is no need to set up additional restrictions to access the data. Given that the data to be stored is anonymous, there should be no sensitive information nor the need for a data access committee.

4.3 Making data interoperable

This section aims at describing how the data produced in the project are made interoperable, i.e. how data exchange and re-use between researchers, organizations, countries, etc. is enabled, for instance, by using standard formats and/or data compliance with open software applications.

To ensure interoperability, all project members will follow the following practices (whenever applies and possible):

- Open standards: The data will be made available in widely accepted interoperable data formats such as JSON, XML, and CSV to facilitate data exchange and integration between different educational AI systems and applications.

- ONNX (Open Neural Network Exchange) and PMML (Predictive Model Markup Language): These formats are used to share or upload the AI models created in the project to public platforms. These options, widely accepted in the machine learning community, enable the dissemination and deployment of models in various environments and platforms, that is, the interoperability and portability of the AI models are guaranteed.

4.4 Increase data re-use

If possible, the data will be licensed under an Open Access license (e.g. CC BY). However, this will depend on the level of privacy, and the Intellectual Property Right (IPR) involved in the data primarily by consortium partners and project implementors. A period of embargo will only be necessary if a data set contains exploitable results that will justify an embargo. Therefore, the data will be licensed to permit the widest reuse possible when no limitations are identified by the key stakeholders. The intention of the project is to make as much data as possible reusable for third parties. Restriction will only apply when privacy, IPR, or other exploitation grounds are in play.

All data sets will be cleared of bad records, with clear naming conventions (Structure: [Projectabbrevation]_[Workpackage]_[TypeOfData]_[Title]_[yyyymmdd]_[version]), and with appropriate metadata.

All data generated and collected in AI4ED will undergo a quality check by the project coordinator in order to analyse its individual plausibility and consistency, making sure that others can directly use it to perform assessments and validate the results produced by the project.

Metadata and documentation will be provided in a codebook for the survey data sets and in model cards and datasheets for datasets will accompany the machine learning assets for an increase of reuse in appropriate research contexts.

5 Other research outputs

In addition to the management of data, beneficiaries will also consider and plan for the management of other research outputs that may be generated or re-used throughout their projects. Such outputs can be either digital (e.g. software, workflows, protocols, models, etc.) or physical (e.g. new materials, antibodies, reagents, samples, etc.). In the case of AI4ED the following research outputs are planned at this point of time:

- Report which describes the programme for training AI4Ed transformation skills (Deliverable 4.1)
- MOOCs for AI4Ed training programmes (Deliverable 4.2)
- Toolkit on AI4Ed solutions for reliable AI in education (Deliverable 5.1)
- Model Implementation Report on EDEHub (Deliverable 7.2)

6 Allocation of resources

For all partners: Preserving and sharing the data in the repository Zenodo are free of charge. Items will be retained for the lifetime of the repository.

Table 18 - Allocation of resources per beneficiary

IMH	n/a
UBREMEN	n/a
CENFIM	n/a
SCSKZ	n/a
ALCHEMYML	The costs associated with making data or other research outputs FAIR (Findable, Accessible, Interoperable, and Reusable) in our project will include both direct and indirect costs. Some aspects to consider are: Storage and archiving: We will need to have adequate storage resources to store the data we receive from our collaborators and the created models. This may involve costs related to the acquisition and maintenance of servers or cloud storage services. Infrastructure for the API: To create the API that will allow access to the data and consume the model, a server or platform will be required to host and manage the API. These costs will include the rental of hardware, configuration, and maintenance of software, and possibly expenses related to security and API management. It would be possible to consider the consolidation of the two previously mentioned servers into a single one. To ensure long-term preservation of the data, we will consider the following aspects: Backup and data recovery: We will implement policies and procedures for regular backup of stored data and models, ensuring that up-to-date and accessible backups are available in case of data loss or corruption. Documentation and metadata: We will maintain comprehensive and up-to-date documentation about the data, including relevant metadata that describe the nature of the data, its origin, its structure, and any important considerations for its use and reuse.

7 Data security

With regard to the personal data, only anonymized data is allowed to be placed on the selected platform Odoo for sharing it with consortium partners. The raw confidential data will be securely stored by individual organization which will explain below which access measures do apply. Every consortium partner collects personal data which will not be shared with other partners. These personal datasets will be stored on the responsible partner's storage system for the duration of the project. Every partner is responsible to ensure that the data are stored safely, securely and in full compliance with European Union data protection laws. Collected anonymized data, will be stored on password-protected database of the respected partner to which only ALCHEMYML and own involved personnel can have access on a need-to-know basis (Figure 1).

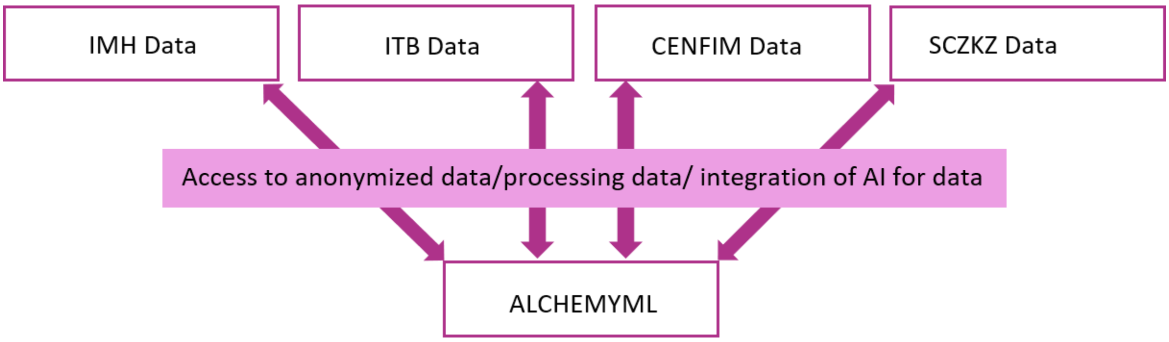

IMH Data
ITB Data
CENFIM Data
SCZKZ Data
Access to anonymized data/processing data/ integration of AI for data
ALCHEMYML

Figure 1: ALCHEMYML data access

Figure 1 - ALCHEMY data access

The anonymized project data will be stored in the trusted repository Zenodo for long term preservation and curation. Within the institutions following data security measures apply:

Table 19 - Data security measures per beneficiary

IMH	Data will be saved on local backup servers. These servers are accessible only by admin staff. Backups are replicated on other servers, that are protected by MFA authentication and hardened by firewall rules. Also, the server has failure proof hardware e.g. redundant power, raid hard-disk system. Responsible for data security: Ignacio Aguayo Simó from Dataprev (https://www.dataprev.es/)
UBREMEN	Data will be saved on local hosted cloud server. Hard-disks and therefore data is encrypted with 4096bit encryption. Data is only accessible by selected project members. Back-up on another server in encrypted format is provided. Server has failure proof hardware e.g. redundant power, raid hard-disk system. Responsible for security: Alex Seedorf
CENFIM	Data will be saved on local backup servers and replicated to another internal backup server on a different location. These servers are accessible only by admin staff. Also, the server has failure proof hardware e.g. redundant power, raid hard-disk system.

	Responsible for data security: Rita Lima
SCSKZ	The data in eAsistent is stored and protected under the protocols of the external school documentation provider and complies with GDPR protocols. The data that we will directly process during the project will be stored on a local server in the cloud. Hard drives and thus data are encrypted. The data is accessible only to selected members of the project. Hard-disks are in raid mode, so if one disk has a failure, the data are stored on another disk. Hard drives and thus data are encrypted. The data is accessible only to selected members of the project. Responsible for data security: Aleš Ribnikar
ALCHEMYML	Secure storage: We will use secure storage systems with appropriate access controls to manage the data and the models such as firewalls to protect against unauthorized access. Data transfer protocols: We will use secure and encrypted protocols, such as HTTPS, to protect data and model's consumption. This helps prevent unauthorized interception or access to the data and the models during transfer. If access control and authentication is necessary for data consumption a robust access control mechanism will implement to ensure that only authorized individuals or systems have access to it. This may involve user authentication methods, role-based access control, and permissions management to restrict data access based on user roles and responsibilities. Responsible: Mikel Armendariz

8 Ethics

In line with the European Strategy for Data 2020 (Data Economy 2023) and the recent post-GDPR proposals for regulation, which include the proposed regulations on Data Governance, DGA (European Commission 2022), the EU Artificial Intelligence Act (European Parliament 2023), the proposals for the Digital Services Act, DSA (European Commission 2023) as well as the Ethics Guidelines for Trustworthy AI (European Commission 2022), this section outlines the ethical blueprint of AI4ED. It presents an ethical framework thataims to establish the necessary means, rules, and structures for the implementation of AI4ED.

This project directs to utilize artificial intelligence in developing an educational model that supports students throughout their academic journey, starting from enrolment, progressing through various course phases, and concluding with final grades. The system will incorporate key performance indicators (KPIs) to monitor students' progress, offer recommendations, evaluate potential dropouts, and personalize the learning experience. By leveraging AI's adaptivity, both teachers and students can benefit from enhanced educational and learning processes.

The development of the model and hence the courses for students will adhere to the recommendations outlined above, ensuring compliance with transparency and ethical standards. Specifically, the system will follow and implement the criteria presented in the document "Artificial Intelligence and the Future of Teaching and Learning" (Office of Educational Technology 2023) whenever feasible. Apart from these criteria which apply for the whole project consortium, the key requirements for trustworthy AI stated in the ethical guidelines on the use of artificial intelligence (AI) and data in teaching and learning for Educators (European commission 2022) are taken into consideration:

“Human agency and oversight including fundamental rights, children’s rights, human agency, and human oversight.

Transparency including traceability, explainability and communication.

Diversity, non-discrimination, and fairness including accessibility, universal design, the avoidance of unfair bias, and stakeholder participation, which allows use regardless of age, gender, abilities, or characteristics - with a particular focus for students with special needs.

Societal and environmental wellbeing including sustainability and environmental friendliness, social impact, society, and democracy.

Privacy and data governance including respect for privacy, quality and integrity of data, and access to data” (European Commission 2022, page 18)

They mostly overlap with the criteria stated in Artificial Intelligence and the Future of Teaching and Learning" (Office of Educational Technology 2023), but are more global and include the issue of privacy and data governance.

With respect to WP5 privacy is an important ethical concern. Details on personal data regarding to each data set can be found in the relevant data collection tables. When the data is anonymized, the potential ethical and legal issues that can impact data sharing are significantly reduced. Therefore, the project consortium aims to avoid collecting any personal data from participants or is going to anonymize data at the very beginning of data collection. Anonymization helps protect the privacy and confidentiality of individuals whose data is involved. However, the consortium ensures that informed consent for data sharing, data processing and long-term preservation takes place. The project procedure and aims will be clearly stated andexplained in the informed consent, even if the data is anonymized.

By obtaining informed consent, individuals are made aware of the purpose, scope, and potential risks associated with data sharing and long-term preservation. This ensures transparency and allows individuals to make an informed decision about their data. Additionally, including consent in survey data helps establisha legal basis for data processing and sharing activities. Therefore, the project consortium considers it as good practice to include informed consent provisions in survey data to address any potential ethical or legal considerations and to ensure compliance with relevant regulations and guidelines.

9 Other issues

Besides the Erasmus+ procedures, the consortium partners will use these procedures for research data management:

Table 20 - Other Issues per beneficiary

IMH	N/A
UBREMEN	Open Access Policy: https://www.uni-bremen.de/fileadmin/user_upload/forschung/Open_Science/Open_Acces_Policy_DE_002_.pdf Recommendations for research data management: https://www.uni-bremen.de/fileadmin/user_upload/forschung/Divers/Empfehlungen_zum_Umgang_mit_Forschungsdaten.pdf
CENFIM	Internal requirements only
SCSKZ	Yes, as a school we follow the guidelines and legal requirements on the protection of personal data, regulations, as on link: https://www.sc-konjice-zrece.si/index.php/predstavitev/varstvo-osebnih-podatkov
ALCHEMY ML	CSIC: Plan de Gestión de datos. Bernal. Valladolid University: Plan de Gestión de Datos - Datos de Investigación - Biblioguías at Universidad de Valladolid Biblioteca (uva.es)

10 References

European Commission (2022). The European Data Governance Act (DGA). https://digital-strategy.ec.europa.eu/en/policies/data-governance-act

European Commission (2023). The Digital Services Act (DSA). https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package

Data Economy (2023). EU Data Strategy 2020. http://dataeconomy.eu/eu-data-strategy-2020/#page-content

European Commission (2022). Ethics guidelines for trustworthy AI. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

European Commission (2023). A European Strategy for data. https://digital-strategy.ec.europa.eu/en/policies/strategy-data

European Parliament (2023). Artificial intelligence act. https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/698792/EPRS_BRI(2021)698792_EN.pdf

Jahn, Najko, Laakso, Mikael, Lazzeri, Emma, & McQuilton, Peter (2023). Study on the readiness of research data and literature repositories to facilitate compliance with the Open Science Horizon Europe MGA requirements (Version 1.0). Zenodo. https://doi.org/10.5281/zenodo.7728016).

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru (2019). Model cards for model reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency. 220–229. https://dl.acm.org/doi/pdf/10.1145/3287560.3287596

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé, and Kate Crawford (2018). Datasheets for Datasets. http://arxiv.org/abs/1803.09010