

D-2024-3108: Explanations in Artificial Intelligence to strengthen Learning Analytics for higher education

Research Data Summary

List and describe all datasets or research materials that you plan to generate/collect or reuse during your research project. For each dataset or data type (observational, experimental etc.), provide a short name & description (sufficient for yourself to know what data it is about), indicate whether the data are newly generated/collected or reused, digital or physical, also indicate the type of the data (the kind of content), its technical format (file extension), and an estimate of the upper limit of the volume of the data.

Dataset name / ID	Description	New or reuse	Digital or Physical data	Data Type	File format	Data volume	Physical volume
ALPACAS	The ALAPCAS data set resulted from the ALAPACAS educational innovation project and holds information on student activity on the learning platform and academic attendance. (Case study 1a)	E	D	N+T	csv and txt	<100GB	KU Leuven one drive
CPS-peer evaluation-PenO	The dataset contains the mid-term and final peer evaluation of students in the collaborative problem solving course of PenO - engineering science bachelor (Case study 2a) - collaborative problems solving	N	D	N+T	csv and xlsx	<1GB	
CPS-peer evaluation-EduMa	The dataset contains the mid-term and final peer evaluation of students in the collaborative problem solving course of the EduMa - engineering science bachelor (Case study 2b) - collaborative problems solving	N	D	N+T	csv and xlsx	<1GB	
EssayFeedback	The dataset contains the short essays of students and the feedback they received (when available) from the teachers/teaching assistants (Case study 1b)	N	D	N+T	txt, csv and xlsx	<100GB	
User-centered-design-Case1	WP2.1c - iterative user centered design of dashboard for predictive advising case. The different phases of the iterative user-centered design are audio-recorded and transcribed, as a basis for qualitative data analysis	N	D+P	S+N+T	Audio-recorded and transcribed into MS Word, Nvivo and Excel	<100GB	
User-centered-design-Case2	WP2.1c - iterative user centered design of dashboard feedback. Semi-structured interviews with educators	N	D	S+N+T	Audio-recorded and transcribed into MS Word, Nvivo and Excel	<100GB	
Literature-overview	WP3.1 - database of literature	N	D	T	csv - excel	<100GB	
Impact-WP3.2a	WP3.2a- Case study 2 - Impact measurement on teachers	N	D	N+T	For qualitative data: Audio-recorded and transcribed into MS Word, Nvivo and Excel For quantitative data: csv and excel	<100GB	
Impact-WP3.2b	WP3.2b - Case study 2 - Impact measurement on students	N	D	N+T	For qualitative data: Audio-recorded and transcribed into MS Word, Nvivo and Excel For quantitative data: csv and excel	<100GB	

Impact-WP3.2c	WP3.2c - Case study 1 - Impact measurement on teachers	N	D	N+T	For qualitative data: Audio-recorded and transcribed into MS Word, Nvivo and Excel For quantitative data: csv and excel	<100GB	
Impact-WP3.2d	WP3.2d - Case study 1 - Impact measurement on students	N	D	N+T	For qualitative data: Audio-recorded and transcribed into MS Word, Nvivo and Excel For quantitative data: csv and excel	<100GB	

If you reuse existing data, please specify the source, preferably by using a persistent identifier (e.g. DOI, Handle, URL etc.) per dataset or data type:

The ALPACAS dataset originates from the ALPACAS educational innovation project. A pseudonomized version is used after obtaining permission from SMEC or the KU Leuven Student and Learning data committee.

Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? If so, refer to specific datasets or data types when appropriate and provide the relevant ethical approval number.

- Yes, human subject data (Provide SMEC or EC approval number below)

For all above data permission will be requested with SMEC. An overall SMEC approval has been obtained: G-2024-8446.
For each of the subproject (different phases of the case studies, ...) separate approval with SMEC will be obtained.

Will you process personal data? If so, please refer to specific datasets or data types when appropriate and provide the KU Leuven or UZ Leuven privacy register number (G or S number).

- Yes (Provide PRET G-number or EC S-number below)

For all above data permission will be requested with SMEC. An overall SMEC approval has been obtained: G-2024-8446.
For each of the subproject (different phases of the case studies, ...) separate approval with SMEC will be obtained.

Does your work have potential for commercial valorization (e.g. tech transfer, for example spin-offs, commercial exploitation, ...)? If so, please comment per dataset or data type where appropriate.

- No

Do existing 3rd party agreements restrict exploitation or dissemination of the data you (re)use (e.g. Material or Data transfer agreements, Research collaboration agreements)? If so, please explain in the comment section to what data they relate and what restrictions are in place.

- No

Are there any other legal issues, such as intellectual property rights and ownership, to be managed related to the data you (re)use? If so, please explain in the comment section to what data they relate and which restrictions will be asserted.

- No

Documentation and Metadata

Clearly describe what approach will be followed to capture the accompanying information necessary to keep data understandable and usable, for yourself and others, now and in the future (e.g. in terms of documentation levels and types required, procedures used, Electronic Lab Notebooks, README.txt files, codebook.tsv etc. where this information is recorded).

We will provide documentation in the following way:

1. **Quantitative data - log data, outcome of surveys** : the documentation will accompany the collected data, containing the date of data collection, the population, the context, variable information, etc.
2. **Qualitative data - interviews, focus-group discussions, think-alouds** : each original and transcribed audiorecorded file will be accompanied by a standardized overview sheet containing the following elements: date, setting, participants, informed consent, implemented guidelines, etc..
3. **Qualitative data - thematic analyses** : the coding book will accompany the data

This documentation will be kept in the same folder as the data. T

We will anonymise our data where possible and use pseudonymization where anonymization is not possible. The file with the subject ID's in the case of pseudonomized data will be in the appropriate folder.

Will a metadata standard be used to make it easier to find and reuse the data?

If so, please specify which metadata standard will be used.

If not, please specify which metadata will be created to make the data easier to find and reuse.

- No

Currently we do not plan to use a metadata standard.

Data Storage & Back-up during the Research Project

Where will the data be stored?

- Other (specify below)
- Shared network drive (J-drive)
- Sharepoint online
- Personal network drive (I-drive)

The main storage will be the project's teams-site with the following additional measures:

- strict management of access rights to the datasets
- all project collaborators use the multi-factor authentication provided by KU Leuven
- for pseudonymized data:
 - the code list is kept on the project's j-drive or on the personal network drive of the project promotor and responsible researcher.
 - the data is additionally encrypted. The key is kept on the project's j-drive or on the personal network drive of the project promotor and responsible researcher.

How will the data be backed up?

- Standard back-up provided by KU Leuven ICTS for my storage solution

Is there currently sufficient storage & backup capacity during the project?

If no or insufficient storage or backup capacities are available, explain how this will be taken care of.

- Yes

Yes, the KU Leuven offer is expected to be sufficient. If not, additional storage from KU Leuven will be bought. All solutions are scalable, allowing capacity to be expanded as needed.

How will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?

This is done through strict management of access rights to datasets. At least yearly, an additional check will be done to check if

- only the right people have access to the data (need-to-have)
- these people only have access to the data they strictly need (need-to-know)

What are the expected costs for data storage and backup during the research project? How will these costs be covered?

Due to the limited size of the data no additional costs are expected. If additional costs occur, they can be paid with the obtained funds, or with reserve-funds from the main promotor.

Data Preservation after the end of the Research Project

Which data will be retained for 10 years (or longer, in agreement with other retention policies that are applicable) after the end of the project?

In case some data cannot be preserved, clearly state the reasons for this (e.g. legal or contractual restrictions, storage/budget issues, institutional policies...).

- All data will be preserved for 10 years according to KU Leuven RDM policy

Where will these data be archived (stored and curated for the long-term)?

- Shared network drive (J-drive)
- KU Leuven RDR

What are the expected costs for data preservation during the expected retention period? How will these costs be covered?

No costs are expected due the limited data size.

Data Sharing and Reuse

Will the data (or part of the data) be made available for reuse after/during the project?
Please explain per dataset or data type which data will be made available.

- Yes, as open data
- No (closed access)

We will strive for opening the data. The actual decision on to open the data will be available on the decision of the SMEC/PRET application.

If access is restricted, please specify who will be able to access the data and under what conditions.

Either completely open or closed.

Are there any factors that restrict or prevent the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions)?

Please explain per dataset or data type where appropriate.

- Yes, ethical aspects

We will strive for opening the data. The actual decision on to open the data will be available on the decision of the SMEC/PRET application.

Where will the data be made available?

If already known, please provide a repository per dataset or data type.

- KU Leuven RDR (Research Data Repository)

When will the data be made available?

- Upon publication of research results

Which data usage licenses are you going to provide?

If none, please explain why.

- GNU GPL-3.0 (code)
- CC-BY 4.0 (data)

We intend to use the CC-BY-NC-SA and GNU GPL-3.0 licenses.

Do you intend to add a persistent identifier (PID) to your dataset(s), e.g. a DOI or accession number? If already available, please provide it here.

- Yes, a PID will be added upon deposit in a data repository

What are the expected costs for data sharing? How will these costs be covered?

The size of the data is limited and no additional costs are expected.

Responsibilities

Who will manage data documentation and metadata during the research project?

The PhD students working with the data.

Who will manage data storage and backup during the research project?

The project's main promotor supported by the PhD students.

Who will manage data preservation and sharing?

Promotors and ICTS

Who will update and implement this DMP?

The main promotor