Plan Overview

A Data Management Plan created using DMPonline

Title: MSCA-IF n. 101030581, "STEMCo" (v1.2.0)

Creator: Andrea Renee Leone-Pizzighella

Principal Investigator: Andrea Renee Leone-Pizzighella

Contributor: Elias Telser

Affiliation: Other

Funder: European Commission

Template: Horizon Europe Template

ORCID iD: 0000-0002-0585-8407

Project abstract:

Stances Toward Education in Multilingual Contexts (STEMCo) transforms findings from ethnographic engagement with multilingual school communities into research-driven and policy-informed teacher-education materials. Ten months of ethnographic and participatory action research in middle schools—one in South Tyrol, a historically multilingual context, and the other in Veneto, a newly multilingual context—will support data-driven teacher reflection on classroom practices (WP1). These data are then analyzed in light of communicative repertoires, multilingualism, pedagogy, and language ideologies and findings are disseminated to academic stakeholders (WP2). Via short visits and a secondment with key national and international partners, these findings are then exploited for the professional development of teachers throughout Europe via a sustainable, Open Access online resource (WP3). The overall aim of STEMCo is to develop data-driven and research-based approaches, and associated training materials, for accommodating community multilingualism and individual plurilingualism in school communities throughout Europe.

ID: 140740

Start date: 04-04-2022

End date: 08-06-2025

Last modified: 19-12-2023

Grant number / URL: https://cordis.europa.eu/project/id/101030581

Copyright information:

The above plan creator(s) have agreed that others may use as much of the text of this plan as they would like in their own plans, and customise it as necessary. You do not need to credit

the creator(s) as the source of the language used, but using any of the plan's text does not imply that the creator(s) endorse, or have any relationship to, your project or proposal

MSCA-IF n. 101030581, "STEMCo" (v1.2.0)

Data Summary

Will you re-use any existing data and what will you re-use it for?

No, I will not re-use any existing "raw" data from other repositories.

What types and formats of data will the project generate or re-use?

This project will generate five types of data:

- Type A: audio and video recordings of one-on-one interviews and classroom discourse
- Type B: audio and video recordings of teacher reflection sessions
  - A & B will be collected via three digital audio recorders placed on students' and teachers' desks, and via one digital video camera.
- Type C: transcripts of Type A and B audio and video files
  - Transcripts will be produced with the assistance of CLAN software
- Type D: field notes and artifacts (e.g., photos of class materials, student work, etc.)
  - Fieldnotes and artifacts will be converted into digital format (.doc or .pdf)
- Type E: surveys
  - 1 anonymous survey was conducted on paper and then scanned.

Pseudonymization:

- A pseudonym key was created which transforms participants' names into other names that share the same distribution in the local population but which are not similar to the original name. A numeric code has also been established for each participant, based on the first two letters of the participant's first and last names, their sex, and the region where they live. For instance:
  - 39 = South Tyrol; 37 = Veneto
  - Male = 1, Female = 2
  - First name Maria = MA = 1301 (M is 13th letter of alphabet, A is 1st letter)
  - Last name Rossi = RO = 1815 (R is 18th letter, O is 15th letter
    - Code for Maria Rossi, female, from South Tyrol = 39213011815
- Type C, D, E (textual) data will be pseudonymized using this key via a simple find and replace function in Word or the appropriate software. Any scans of data in pdf format will be redacted and re-labeled with the correct pseudonym.
  - All Type D data have been pseudonymized. Type E data was collected anonymously and does not require pseudonymization. The creation of Type C data is underway [as of 12.2023].
  - The pseudonym key has been saved to a local hard disk which only the PI has access to.
- Type A and B data (audiovisual) will be pseudonymized in Adobe Premiere Pro by (1) tagging points in the associated audio/video clips which contain participants' names, (2) bleeping out the relevant tagged points which contain names, and (3) via the application of a visual filter (or series of filters) which obscure the identity of the participants. After the application of this visual filter, it will still be possible to see where participants are positioned in the classroom, which direction they are looking in, and whether or not they are speaking, but their distinguishing features will not be visible.
  - Type A and B data are being pseudonymized via visual and audio filters in Adobe Premiere Pro. Audiovisual data imported into CLAN will already have been pseudonymized.

Anonymization:

- Type E data will be used both for information about specific participants and for aggregated information about the class, school, and study overall. Whenever information culled from surveys risks revealing the identity of a participant, it will only be presented as part of aggregated and therefore anonymized data. For instance, a student's dyslexia will not ever be associated with his/her actual identity or pseudonym, but will only be presented in aggregated format (e.g., of the 25 students in the study, 2 had certificates for additional support by teachers according to Law 104).
  - All survey data were collected anonymously
- Type C data can be anonymized in certain cases but will more often be pseudonymized. In the event that Type C data does not accompany Type A/B data, it can be anonymized by limiting the surrounding contextual information.

Data storage and sharing:
The data storage process will be as follows:

1. Upon initial collection, due to lack of internet access at the field sites, Type A and B data were stored on SD disks and/or in digital audio recorders. Type D data were stored either in a physical notebook/paper format or on PI's password-protected computer.
2. As soon as possible, within 48 hours, these data (A, B, D) were transferred (directly, not via the internet) to the PI's computer

and immediately to Eurac Research's secure server in a "to be processed" file folder and deleted from the original sources or, in the case of paper copies, shredded.
    1. Data collection concluded in June 2023 and no original copies of data remain outside of the secure server.
3. As soon as possible, within one month from the date of original data collection, pseudonyms were applied to all textual data (Type D). A copy of these data was transferred to NVivo for analysis (coding, transcription, etc.). All non-pseudonymized copies of the data have been deleted. NVivo files will be shared on Eurac's secure server between members of the STEMCo team (the PI and three interns, who have all taken required trainings in data protection and ethics).
4. Due to technical requirements that surpassed what was expected, Types A and B data have not been entirely pseudonymized. They are stored in the secure server and will be fully pseudonymized by the end of 2024.
5. Type C data will be produced via CLAN software from initial analyses of Types A, B, D data and will be saved directly to Eurac's secure server. These data will then be uploaded to NVivo for analysis.
6. Type E data were gathered anonymously and then digitized, not via SurveyMonkey as anticipated.
7. Data will be securely stored on Eurac Research's server for the time that it takes to complete the analysis (approximately 36 person months, extending through the end of 2025 or beginning of 2026). Before this period ends:
    1. The pseudonym key will be destroyed.
    2. Selections of Type A and B data (pseudonymized) will be made available with accompanying pseudonymized Type C data (transcripts) on the Open Access repository TalkBank (specifically ClassBank and/or BilingBank).
        1. For more information about TalkBank: https://talkbank.org/
    3. Selections of Type D data (anonymized where possible, or otherwise pseudonymized) will be made available on the TalkBank and/or CLARIN repositories.
        1. For more information about CLARIN: https://www.clarin.eu/content/repositories
    4. Type E data will be aggregated and anonymized and made available on the CLARIN repository.

A Data Access Agreement has been created in accordance with a former intern who is using selections of anonymized data for their thesis.

Naming conventions for data:

- Type A: audio and video files will have the same names and will only be differentiated by file extensions (e.g., .mp3 vs .mp4)
    - interview audio/video recordings: YYYYMMDD_PN*_INT_Pseudonym
        - *PN = place name abbreviation
    - classroom audio/video recordings:
        - for whole class recording: YYYYMMDD_PN_SUB_TPseudo, and if necessary 01, 02, ...
            - SUB examples: ART, GER, HIS, ITA, ENG…
        - for student/group recording: YYYYMMDD_PN_SUB_StuPseudo and if necessary 01, 02, ...
            - REV 29.06.2023: All original copies of the files were tagged as ORIG in the file name so as not to confuse them with the pseudonymized copies.
            - REV 29.06.2023: In case of no clear identification of who speaks (students) the files will be named as whole class recordings YYYYMMDD_PN_SUB_TPseudo and if necessary 01, 02, ... + name of recorder.
                - e.g. YYYYMMDD_PN_SUB_TPseudo 01 RecName ORIG
- Type B: audio and video files will have the same names and will only be differentiated by file extensions (e.g., .mp3 vs. .mp4)
    - REV 29.06.2023: Recordings of WORKSHOPS are labeled YYYMMDD_BZ_WKP1 01, 02, etc. + RecName (date followed by city initials, number of workshop in series, and recorder label)
- Type C: transcripts will have the same names as their associated audio/video recordings
    - REV 26.06.2023: e.g., YYYYMMDD_PN_SUB_TPseudo/StuPseudo and if necessary 01, 02,...
    - REV 29.06.2023: transcripts that combine all recordings from a single discourse event will be labeled as: YYYY.MM.DD PN 2A science TPseudo
- Type D: field notes and digitized artifacts will be labeled as follows:
    - REV 29.06.2023: YYYYMMDD_PN ORIG (for fieldnotes)
    - REV 29.06.2023: YYYYMMDD_PN ORIG with 01, 02, 03 etc extensions where necessary (for all other artifacts)
- Type E: survey results will be aggregated in an Excel sheet
    - REV 29.06.2023: survey results have been digitized and saved as YYYYMMDD_PN_SURVEYS

What is the purpose of the data generation or re-use and its relation to the objectives of the project?

Data generation is associated with the work packages as follows:

WP1: Engage teachers in South Tyrol and Veneto via PAR in the ethnographic exploration of their multilingual classrooms over one academic year.

The four phases of participatory action research (PAR) used in STEMCo (observation, reflection, planning, implementation) are associated with different types of data.

In the observation and implementation stages, Type A data and associated Type D data are collected via participant observation in classrooms and schools. These are then used in the generation of Type C data.

In the reflection and planning stages, pseudonymized Type A-C-D data are shared with teachers at the school where the data were collected in order to stimulate reflection on teacher practice and plan potential modifications to teaching methods, approaches, classroom management, and interactional elements of classroom life. Type C data is then generated based on the recordings of workshop sessions with teachers.

WP2: Analyze classroom discourse, didactics, artifacts, and PAR interviews and workshops to identify communicative patterns in these classrooms and to develop pedagogical strategies for teaching multilingual students.

The data generation phase of STEMCo (WP1) has the purpose of involving teachers in the exploration of discourse in their own schools (WP2). In combination, WP1 and WP2 facilitate a participatory research approach that involves teachers in reflection on the ways they and their students communicate, and provides opportunities for them to learn from themselves and their colleagues via linguistic ethnographic methods. The reflections and modifications have a cumulative effect over the course of the academic year, facilitating not only a means of managing and accommodating the linguistic diversity of teachers' current classes, but also a means of doing the same in the future.

WP2 also includes the analysis of data by the research team with the aim of making scientific contributions to empirical and theoretical studies in the linguistic anthropology of education and associated disciplinary areas.

WP3: Exploit findings from WP1 and WP2 to design, promote, and launch open-access, research-driven and policy-informed teacher education materials for teachers who (will) work in multilingual communities in Europe.

After data collection and subsequent engagement with the fieldsites is complete, WP3 involves the creation of teacher training materials for pre-service and in-service teachers working in linguistically diverse settings. This phase draws on pseudonymized and, when possible, anonymized Type A-B-C-D-E data in the generation of Open Access materials and courses, and will be facilitated by short visits and a secondment with collaborating research centers and teacher training programs. Courses and repositories of materials for teachers working with multilingual students and/or in linguistically diverse settings will be made available in an Open Access format that is yet to be determined, but may resemble the mode of presentation used in learninghowtolookandlisten.com. Pseudonymized or anonymized audio, video, and transcript data will be available in Open Access format for researchers in linguistics and education via the TalkBank repository (talkbank.org). Metadata will be generated for all Type C and E data and will be made available on CLARIN.

What is the expected size of the data that you intend to generate or re-use?

I intend to record approximately 10-15 hours of discourse per week (both in the form of classroom discourse and interviews). Over the course of the 2022-2023 academic year, this amounts to approximately 350-500 hours of recordings. Type A & B data will be reviewed and matched up with one another so that classroom events and interviews can be coded qualitatively, compared with Type D data, and then transcribed to produce Type C data. It is unlikely that all recorded discourse will be transcribed over the course of the project. Instead, the researcher will draw on personal observations of the day's events, compare to fieldnotes, and use the fieldnotes as a guide to identify moments of talk that fall within the scope of the project. These moments will be transcribed and will then be coded further so as to identify themes to pursue in analysis. Type E data will be collected via 2-3 short surveys of maximum 5 questions each from approximately 90-100 participants.

REV 29.06.2023: To date, I have 1.3TB of data stored. This will fluctuate as originals are transformed into pseudonymized/anonymized copies.

What is the origin/provenance of the data, either generated or re-used?

The discourse data was obtained via recordings of classes in session and interviews with teachers at two middle schools. Transcripts and codes will be produced by the researcher and other team members. Surveys were gathered from student participants.

To whom might your data be useful ('data utility'), outside your project?

The data will be useful to two groups of users:

1. Applied linguists (for discourse analysis, interactional sociolinguistics, second language development, etc.), linguistic anthropologists (enregisterment, speech chains, etc.), anthropologists of education (academic discourse socialization, student identity development, etc.)

2. Teachers and teacher trainers (in pre-service teacher education courses at universities, in professional development initiatives at schools and school boards for in-service teachers, for practitioner-researchers in their own classrooms)

FAIR data

2.1. Making data findable, including provisions for metadata: Will data be identified by a persistent identifier?

Each transcript and media file in TalkBank is assigned a Permanent ID via the Handle System (www.handle.net), and each corpus has an ISBN and DOI (digital object identifier) number. Data deposited in CLARIN will be assigned handles.

2.1. Making data findable, including provisions for metadata: Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed? In case metadata standards

do not exist in your discipline, please outline what type of metadata will be created and how.

Rich metadata will be provided to allow for discovery of Type A-B-C-E data. The DDI standards will be followed in order to maximize data discoverability (https://ddialliance.org/). LRMI schemas will be followed in the labeling of data archived in TalkBank (https://www.dublincore.org/specifications/lrmi/concept_schemes/). TEI standards for ensuring findability of texts deriving from transcribed speech will be followed wherever possible for Type C data (https://tei-c.org/release/doc/tei-p5-doc/en/html/TS.html). The vocabulary used for creating metadata descriptions will follow CESSDA standards (https://www.cessda.eu/Tools/Vocabulary-Service).

2.1. Making data findable, including provisions for metadata: Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use?

The indexing and registration of materials in TalkBank have been adapted to align with OLAC (Online Language Archives Community) at www.language-archives.org and VLO (Virtual Language Observatory) at https://vlo.clarin.eu. As is stated on the TalkBank website, "These systems allow researchers to search for whole corpora or single files, using terms such as *Cantonese, video, gesture,* or *aphasia.* In order to publish or register TalkBank data within these systems, we create a 0metadata.cdc file at the top level of each corpus in TalkBank. Some of the fields in this metadata file are designed for indexing in OLAC and some are designed for the CMDI system used by VLO and the related facility called The Language Archive (tla.mpi.nl). Because of the highly specific nature of the terms and the software used for regular harvesting and publication of these data, we do not require users to create the 0metadata.cdc files. The following table explains what keywords are expected within each field of these files. The first fields listed are for OLAC and the later ones are for CMDI. For CMDI, the values *unknown* and *unspecified* are also available for most of the fields. ... We use a CLAN program that takes the information from the 0metadata.cdc files and from the header lines in each transcript."
For more information about the specific fields, please see the table in Section 6.2:
https://talkbank.org/manuals/CHAT.html#_Toc107417263

2.1. Making data findable, including provisions for metadata: Will metadata be offered in such a way that it can be harvested and indexed?

Please see above.

2.2. Making data accessible - Repository: Will the data be deposited in a trusted repository?

Yes, data will be deposited in TalkBank's repositories ClassBank and/or BilingBank, as well as in the Eurac node of CLARIN.

2.2. Making data accessible - Repository: Have you explored appropriate arrangements with the identified repository where your data will be deposited?

For TalkBank, STEMCo audiovisual and transcript data has been provisionally accepted. The final decision will be made on the basis of the filters used for pseudonymization of audiovisual materials, since TalkBank does not typically accept "doctored" audiovisual materials. In the event that I am not able to deposit edited video recordings in TalkBank, I will either (a) submit only transcripts and accompanying audio data or (b) submit transcripts, accompanying audio data, and a visual representation (a diagram) of the seating arrangement during the speech event.
The Eurac node of CLARIN does not currently host audiovisual files. I will therefore only deposit textual data and survey results (Types C, D, E).

2.2. Making data accessible - Repository: Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?

Yes. See above.

2.2. Making data accessible - Data:
Will all data be made openly available? If certain datasets cannot be shared (or need to be shared under restricted access conditions), explain why, clearly separating legal and contractual reasons from intentional restrictions. Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if opening their data goes against their legitimate interests or other constraints as per the Grant Agreement.

Not all data will be made openly available because of the ethnographic nature of the project and the possibility of personal data and special categories of personal data appearing in Type A-B-C data, as well as the fact that the majority of participants are minors. In

order to gain consent to gather data, the participants were informed that no data (or ensembles of data) would be made openly available if there were a risk of the data being traced back to them or if there were a risk that personal and/or sensitive details be exposed. Thus, only selections of pseudonymized Type A-B data and of pseudonymized/anonymized Type C data will be made available on TalkBank.

All aggregated Type E data will be made openly available on CLARIN. Selections of Type D and C data that do not risk compromising the confidentiality of participants' will also be made available on CLARIN.

Note: All data made available via TalkBank is by default accessible via CLARIN. However, data made available on CLARIN is not necessarily accessible via TalkBank. (TalkBank is a "node" of the global CLARIN network).

2.2. Making data accessible - Data:

If an embargo is applied to give time to publish or seek protection of the intellectual property (e.g. patents), specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

There is no embargo.

2.2. Making data accessible - Data:

Will the data be accessible through a free and standardized access protocol?

Data stored on TalkBank are openly accessible to all. There is no login required and no password protecting data. Data stored on CLARIN are accessible to anyone affiliated with a university or research center via their institutional email address.

2.2. Making data accessible - Data:

If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?

N/A

2.2. Making data accessible - Data:

How will the identity of the person accessing the data be ascertained?

For TalkBank, this does not apply. For CLARIN, via institutional email address.

2.2. Making data accessible - Data:

Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)?

No.

2.2. Making data accessible - Metadata:

Will metadata be made openly available and licenced under a public domain dedication CC0, as per the Grant Agreement? If not, please clarify why. Will metadata contain information to enable the user to access the data?

Yes.

2.2. Making data accessible - Metadata:

How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available?

Long-term data and metadata preservation for TalkBank is guaranteed by Carnegie Mellon University's KiltHub, as described here: https://www.talkbank.org/info/CMU_Support.pdf

The Eurac node of CLARIN is linked to the broader CLARIN network which ensures long-term preservation of data and metadata should the Eurac node close.

2.2. Making data accessible - Metadata:

Will documentation or reference about any software be needed to access or read the data be included? Will it be possible to include the relevant software (e.g. in open source code)?

No documentation is necessary to access or read the data.

2.3. Making data interoperable:

What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data interoperable to allow data exchange and re-use within and across disciplines? Will you follow community-endorsed interoperability best practices? Which ones?

Please see above, **Section 2.1. Making data findable, including provisions for metadata**

2.3. Making data interoperable:

In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining or extending them?

Yes.

2.3. Making data interoperable:

Will your data include qualified references[1] to other data (e.g. other data from your project, or datasets from previous research)?

[1] ***A qualified reference is a cross-reference that explains its intent. For example, X is regulator of Y is a much more qualified reference than X is associated with Y, or X see also Y. The goal therefore is to create as many meaningful links as possible between (meta)data resources to enrich the contextual knowledge about the data. (Source: https://www.go-fair.org/fair-principles/i3-metadata-include-qualified-references-metadata/)***

Yes.

2.4. Increase data re-use:

How will you provide documentation needed to validate data analysis and facilitate data re-use (e.g. readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.)?

For both repositories, readme files will be provided with information on methods for data collection and analysis, including softwares and conventions used for data processing and transcription.

2.4. Increase data re-use:

Will your data be made freely available in the public domain to permit the widest re-use possible? Will your data be licensed using standard reuse licenses, in line with the obligations set out in the Grant Agreement?

Yes.

2.4. Increase data re-use:

Will the data produced in the project be useable by third parties, in particular after the end of the project?

Yes.

2.4. Increase data re-use:

Will the provenance of the data be thoroughly documented using the appropriate standards?

Yes.

2.4. Increase data re-use:

Describe all relevant data quality assurance processes.

Quality assurance for ethnographic data comes at several points:

1. The PI is employing the assistance of interns throughout the research and analysis phases of STEMCo in order to check any unconscious bias that she has in processing and analyzing data. The STEMCo team meets regularly to share emergent coding schemes, discuss emerging overall themes, transcription conventions, data collection methods, etc.

2. The processing of data done by multiple people and/or over a long period of time may lead to shifts or errors in labeling or organization of data. Software such as Open Refine is meant to identify weaknesses and inconsistencies in the cleaning and categorization of data and to correct them. For more information: https://openrefine.org/

2.4. Increase data re-use:

Further to the FAIR principles, DMPs should also address research outputs other than data, and should carefully consider aspects related to the allocation of resources, data security and ethical aspects.

STEMCo with both publish findings for the academic community and create research-based materials for teachers and future teachers. Any publications in peer-reviewed journals, books, or otherwise will be made open access. Any materials for teachers will also be made open access (entirely free and accessible, with no login or registration requirement).

Other research outputs

In addition to the management of data, beneficiaries should 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.).

Other research outputs in addition to data will include teacher training materials in the form of an open access teacher training course (or a series of independent modules), as well as digital materials that can be used online or printed out and used in class.

Beneficiaries should consider which of the questions pertaining to FAIR data above, can apply to the management of other research outputs, and should strive to provide sufficient detail on how their research outputs will be managed and shared, or made available for re-use, in line with the FAIR principles.

Open access publications will be indexed with appropriate metadata based on the publisher's conventions, with persistent identifiers, and will be findable via Google Scholar, university library searches, and other conventional means of accessing research materials.

Teacher materials will be made available at an accessible URL (no password, no email address, no login information required). In order to avoid any type of membership, registration, or subscription, the costs of hosting the materials for at least ten years will be paid up front in order to guarantee their availability even without a DOI. Once the materials have been created and are in the piloting phase, the PI will seek out a permanent "home" for them (e.g., European Centre for Modern Languages, a university, Eurac Research, or another body of the European Commission). This, however, cannot be determined until the final product of STEMCo has been drafted.

Allocation of resources

What will the costs be for making data or other research outputs FAIR in your project (e.g. direct and indirect costs related to storage, archiving, re-use, security, etc.) ?

There is no direct cost for hosting data on TalkBank or the Eurac node of CLARIN.

Open Access publications that come out of STEMCo will likely cost approximately €2500-3000 each, and there may be as many as five or six of these. Some will be covered by the grant awarded by the European Commission and others of these costs will be covered by Eurac Research.

I anticipate hiring an external collaborator to assist with metadata creation and other technical aspects of data management for approximately €8000-10.000 total.

Hosting the teacher materials in an entirely open format for at least 10 years will also likely cost approximately €1500-3000.
I anticipate spending approximately 3-6 person months on various aspects of the FAIRification of data for my project.

How will these be covered? Note that costs related to research data/output management are eligible as part of the Horizon Europe grant (if compliant with the Grant Agreement conditions)

These costs will be covered either by the MSCA grant itself, by the host institute (Eurac Rsearch), or by other funds made available by third parties.

Who will be responsible for data management in your project?

The project leader/PI, Andrea Leone-Pizzighella, is responsible for data management during the project. TalkBank and CLARIN are responsible for long-term preservation once the data from STEMCo have been accepted.

How will long term preservation be ensured? Discuss the necessary resources to accomplish this (costs and potential value, who decides and how, what data will be kept and for how long)?

Long-term data and metadata preservation for TalkBank is guaranteed by Carnegie Mellon University's KiltHub, as described here: https://www.talkbank.org/info/CMU_Support.pdf
The Eurac node of CLARIN is linked to the broader CLARIN network which ensures long-term preservation of data and metadata should the Eurac node close.
After the 36 person months of the project have ended, the original data will have already been destroyed, and only the pseudonymized and anonymized data will be retained (in the above two repositories and the teacher training website).

Data security

What provisions are or will be in place for data security (including data recovery as well as secure storage/archiving and transfer of sensitive data)?

All data will also be stored in the OneDrive/Sharepoint at Eurac Research. The folders on the PI's desktop and on the server will be encrypted, and the passwords for encryption management will be stored securely by the PI and by the ICT department of Eurac. All files on the PI's desktop are automatically backed up to Eurac's secure server as soon as the computer is connected to the internet and the VPN. Eurac is ISO certified -- Quality Management System (9001) and Information Security Management (27001)

Ethics

Are there, or could there be, any ethics or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).

The raw discourse data (audio, video) cannot be shared because the ethnographic nature of the project may reveal sensitive personal data about participants. The survey data that is made publicly available will only be published in aggregated form. Selections of audio and video, pending the functionality of audiovisual filters, may be made available in a public repository. Transcripts of discourse data will be pseudonymized and made publicly available in excerpted form so as to ensure that no sensitive data is shared.

Will informed consent for data sharing and long term preservation be included in questionnaires dealing with personal data?

The consent forms for all participants include the following simplified description of the GDPR, as well as the relevant aspects of the GDPR in its original form:

Analysis of data and presentation of results for academic audiences and the general public

Anonymized data from (1), (2), (3), and (4) will be analyzed for academic conferences and publications. Selections of pseudonymized data will also be made available via an Open Access database for teachers and language researchers, such as TalkBank. Selections of anonymized data will also be made available via an Open Access repository such as CLARIN. Small selections of anonymized data may also be presented at school community events and at venues for the general public, such as Eurac's website, blog, or social media accounts. See table on page 5 for further details on pseudonymization and anonymization.

- For more information about Open Access science: https://open-research-europe.ec.europa.eu/
- For more information about TalkBank: https://talkbank.org/
- For more information about CLARIN: https://www.clarin.eu/content/repositories

Adaptation of data for creation of teacher training materials

Pseudonymized data from (1), (2), (3), and (4) will also be adapted for use in teacher training materials such as online courses, seminars, and workshops. The aim of these trainings is to facilitate accommodation of multilingualism and communicative differences in middle school communities.

Please see the digital version of complete consent form here: https://eu.surveymonkey.com/r/DLS8TX9

Other issues

Do you, or will you, make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones (please list and briefly describe them)?

N/A