COST Action CA21114:
Data Management Plan

CLIL Network for Languages in Education: Towards bi- and multilingual disciplinary literacies

Contents

1. Introduction

Castelli et al. (2021) maintain that "consistent data management is of paramount importance for the acceleration of science" for all stakeholders involved in research from funders, to researchers to those in industry or professions whose practices are shaped by their findings. The key to good data management is a clear, transparent set of guidelines to direct that research process to ensure the long term curation of data. This comes in the form of a Data Management Plan (DMP), a key requirement now of ERC (European Research Council) and Horizon projects. The reason for this is that not only does the storing of data related to published discoveries encourage and make possible accountability and transparency but it also allows other researchers to build on "previously generated research data from a trustworthy source". As a part of CA21114, our Memorandum of Understanding places great importance on the creation of deliverables that not only support and further our own research agenda but also those who wish to pursue further research in the field of multi/bilingual disciplinary literacies beyond the life of the Action.

The guiding principles that guide the creation of this DMP are those related to FAIR: that is the "storing, sharing, publication and preservation of generated data" (Castelli et al., 2021) that complies with FAIR principles. We aspire to ensure that our data is:

- **Findable** - can be sourced and cited from a reliable, long-term repository
- **Accessible** - can be accessed openly by those who need to access it
- **Interoperable** - can be used and analysed in a multitude of ways with the minimum of intervention.
- **Reusable** - it can be used by many researchers in the years to come to generate new ideas.

As such, our ambition is to create a DMP that serves 3 purposes:

1. An operational tool that guides the creation of the data when research questions and instruments are devised to ensure that FAIR principles can be adhered to in in a timely manner
2. A cohesive tool that helps unify the approaches to data adopted by the ever-increasing group of researchers from across Europe involved in CA21114
3. A demonstrative tool for how a DMP can facilitate good data management practices within COST Actions for other researchers involved in similar projects.

The template that we will follow for our DMP is that created by Science Europe, as this is seen as being a well-respected and formulated example to describe the storage, sharing, publication and generation of a wide range of qualitative and quantitative data. However, as Brand et al. (2015) show in their own DMP, this template is simply a starting point allowing us to modify its contents and structure where necessary to meet the needs of this diverse group. The data storage guidelines that we will follow are those operated by the University of Vienna for their repository *Phaidra*.

Although the creation of a DMP is not considered a requirement of our COST Action, we feel that in order for the widest range of researchers from around the world to benefit from the data we collect, it is important that they are able to access it during and after the project. Moreover, with such a variety of researchers working on different aspects of data gathering beyond their normal working hours, we feel it is important that contributors are recognised for their contributions.

Finally, the creation of this DMP is a collaborative project. The 7 members of CA21114 who contributed to the writing of this DMP all come as representatives of the 5 different working groups. While this core group is recognised as making the most significant contribution to its creation, they

are not the only custodians of it. All members of the COST Action have a right and duty to modify its contents where necessary in a timely fashion to respond to changing research plans.

The Data Management Plan Team

Dr Craig Neville (Coláiste na hOllscoile Corcaigh, Éire)

Dr Evdokia Pittas

Dr Rahime Filiz Agmaz

Dr Kübra Okumuş Dağdeler

Leah Tompkins

Dr Merita Hoxha

2. General information

In this section, we provide essential information regarding this project and general considerations made in the DMP for the collection and management of data.

- **Applicant:** Prof Julia Hüttner (Julia.huettner@univie.ac.at) (Action Chair & Grant Holder Scientific Representative)
    - **Leadership Team**
        - Action Vice Chair — Prof Ana Llinares
        - Grant Awarding Coordinator — Prof Visnja Pavcic Takac
        - Working Group 1 Leader: — Prof Tarja Nikula
        - Working Group 2 Leader: — Dr Francisco Lorenzo
        - Working Group 3 Leader: — Dr Y.L. Teresa Ting
        - Working Group 4 Leader: — Prof Ute Smit
        - Working Group 5 Leader: — Prof Yasemin Bayyurt
    - Additional Roles
        - Grant Awarding Co-Ordinator — Prof Eleni Meletiadou
        - WG1 Vice Leader — Dr Talip Gulle
        - WG2 Vice Leader — Prof Lidija Cvikic
        - WG3 Vice Leader — Dr Merita Hoxha
        - WG4 Vice Leader — Dr Craig Neville
        - WG5 Vice Leader — Dr Barbara Muszyńska
        - YRI Coordinator — Dr Ekaterina Strati
- **Project Number:** CA 21114
- **Funding Program:** COST: European Cooperation in Science and Technology
- **DMP Version:** Version 1.0 (15/09/2023)
- **Research Questions:**
    - *Aims of Action:*
        1. develop a shared conceptualisation and research agenda for the investigation of bi/multilingual disciplinary literacies in CLIL
        2. provide an accessible collection of standardised research instruments and research training
        3. identify patterns of use, development and existing good practices in terms of supporting bi/multilingual disciplinary literacies at school, focusing on grades 5-13
        4. disseminate information on supporting the development of bi/multilingual disciplinary literacies in CLIL classes primarily to educational stakeholders and within academia, but also to postsecondary and industry stakeholders and the general public
    - *Research Coordination Objectives*
        1. To develop a common understanding and shared terminology regarding the use and development of bi/multilingual disciplinary literacies within the framework

of CLIL for academic and educational stakeholders, integrating linguistic, disciplinary, and pedagogical concerns, as well as influences from informal learning through digital media.
2. To provide a co-ordinating platform for the collation, hosting and disseminating of key information on the provision and implementation of CLIL, especially with regard to bi/multilingual disciplinary literacies. This will be available freely.
3. To develop and evaluate standardised research tools and methods for the study of bi/multilingual disciplinary literacies, drawing on the Action's multi-disciplinary expertise.
4. To identify locally appropriate examples of good practice in the teaching of bi/multilingual disciplinary literacies within CLIL.
5. To disseminate information on locally sensitive, effective pedagogic practices of supporting bi/multilingual disciplinary literacies to educational and other stakeholders across Europe and beyond.

- *Capacity Building Objectives*
    1. To achieve an interdisciplinary research agenda for the study of bi/multilingual disciplinary literacies within CLIL across a range of educational contexts and so to strengthen the position of CLIL and CLIL research as key agenda within the European Research Area.
    2. To provide a platform for the mutual sharing of existing research findings on the nature and development of bi/multilingual disciplinary literacies within Key Subjects drawing on different sub-fields of CLIL and Applied Linguistics, as well as general education, subject education, digital media and multilingual schooling.
    3. To strengthen the CLIL research activity in ITCs, especially those where CLIL has only recently been established, through the co-ordination of specific research training schools and STSMs, the reservation of Action Roles for ITC members and a dedicated ITC Co-ordinator in the Action Core Group.
    4. To support Young Researchers and Innovators (YRIs) working on CLIL through the provision of timely training in appropriate research methods and tools in research training schools and through STSMs, which will be widely advertised and supported through a dedicated YRI Coordinator in the Action Core Group.
    5. To facilitate networking between experts from a range of fields involved in bi/multilingual disciplinary literacies, like educational research, applied linguistics as well as researchers involved in the multilingual and digital life-worlds of children and teenagers.
    6. To gather information on the needs from post-secondary stakeholders (workplace, Further/Higher Education) on bi/multilingual disciplinary literacies and provide information to industry (publishing) stakeholders and the general public of the role of bi/multilingual disciplinary literacies.

Data Generation: To provide a comprehensive perspective on all data generated within the project, we employ visuals, graphs, and images, elucidating the data. This section will provide general information about data collection processes, data sources, and the nature and quantity of collected data.

Data Sources: We will outline the sources from which each working group obtains their data (surveys, observations, literature reviews, etc.).

Data Collection Methods: Each working group's data collection methods will be explained.

Data Types: We will specify the types of data that each working group will collect.

Data Collection Process: We will outline the stages, timeline, and responsibilities within the data collection process for each working group. The information gathered from each working group will assist in formulating and implementing data collection strategies.

Metadata Standards: Dublin Core™ Metadata Standards will be utilized to describe and manage the data within the project. It is essential to use Dublin Core Metadata Standards to ensure that metadata is in a standard and machine-readable format and semantically meaningful.

Data storage: during collection phases, all data will be stored securely (i.e. password protected using institutional accounts) and password protected. When data has been collected and cleaned (prior to analysis), it will be transferred to the CA21114 collection in Phaidra at the University of Vienna.

Glossaries: Information will be provided on how the project's data documentation glossary is organized and located, indicating where readers can access definitions and explanations of terms.

Readme Files: "Readme" files will be included to provide important information to project users and readers, making usage more convenient. These files will clarify the types of information they contain and the standards to be followed.

3. Data description and collection or re-use of existing data

This section explains the processes involved in data production, encoding, and storage, as well as the specific data formats and types employed within the scope of the action.

Data Collection

Each WG (Working Group) focuses on distinct research topics and formulates methodologies to address their respective research questions collaboratively. In pursuit of these inquiries, different types of research approaches, including qualitative, quantitative, and mixed research methods, are employed. Notably, surveys and corpus/document analysis serve as the primary methodologies within this action.

To ensure extensive participation, surveys will be administered on a pan-European level to participants from various countries using web-based tools such as Qualtrics and Google Forms due to their capacity to reach a diverse and geographically dispersed audience effectively. All of WGs utilise survey methods as they will help determine beliefs/opinions/perceptions of learners/teachers related to Content and Language Integrated Learning (CLIL) practices and experiences.

WG1	Survey (Qualtrics)	Analysis of existing provision
WG2	Survey (Google Forms) + Corpus Analysis	Analysis of use and development of bi/multilingual literacies by students through corpus analysis
WG3	Content Analysis	Analysis of developmental trajectories of learner groups, transition from primary to secondary school, meta-language to describe disciplinary literacies.
WG4	Survey (Qualtrics)	Analysis of extramural digital practices of students; analysis of teachers digital practices in CLIL classrooms
WG5	Survey (Qualtrics)	Analysis of expectations and practices of bi/multilingual disciplinary literacies from stakeholders.

Constraints for use of existing data

The vast majority of data collected as part of new cost action will be new data given that existing data is not suitable to address the specific research questions of WGs.

Where there is a requirement for secondary data to be used, care and attention will be taken to ensure that previous permissions afforded to research participants in the collection of that data (i.e. privacy, data security, rights of ownership, data accuracy and integrity, purpose limitation, unknown quality, denormalization) will be checked to ensure that it can be re-used. To ensure clarity of provenance:

- README files accompanying any data will detail the provenance of re-used data (see below)
- Clarifications will also be provided in README files that specify parts of the re-used data set that have been used and which have been discarded with short explanations as to why this decision has been taken.

Documenting data provenance

All new and existing data will be accompanied by a README file that will be uploaded to Phaidra along with the dataset.

The README file will include general information, data and file overview, sharing and access information, methodological information and data-specific information. The minimum content required for each of these headings is given below:

(The recommended minimum content is in bold)

General information

1. **Title of the dataset**
2. **Name/institution/address/email information for Principal investigator,** Associate or co-investigators or other people responsible for administrative tasks as follows:
    i. Author/Principal Investigator/Associate Investigator Information
    ii. Name:
    iii. ORCID:
    iv. Institution:
    v. Address:
    vi. Email:
3. **Date of data collection (suggested format: YYYY-MM-DD)**
4. **Geographic location of data collection (city/region, State, Country)**
5. Keywords used to describe the data topic
6. Language information
7. Information about funding sources

Data and file overview

1. **Provide, for each filename, a short description of what the data is about**
2. Provide information about the format of the file if this is not clear from the file name
3. Indicate the relationship between files in case more than one file are included in the data set (e.g. “dataset” or “study” or “data package”)
4. **Date that the file was created (suggested format: YYYY-MM-DD)**
5. If there are more than one versions of the file, include the Date(s) that the file(s) was updated (versioned) and the reason why (if important). Remember to use “V” to denote version and “R” to denote revision.
6. Include information about other data collected that is not previously described

Sharing and access information

1. **Licenses (DECISION TO BE MADE about the Creative Commons licenses types: CC BY-ND or CC BY-NC-ND)**
2. Links to publications that cite or use the data
3. Links to other publicly accessible publicly spaces of the data

4. Recommended citation for the dataset (Include information such as the creator(s) or contributor(s), date of publication, title of dataset, publisher, identifier (e.g. Handle, ARK, DOI or URL of source, version-when appropriate, date accessed-when appropriate)

Methodological information

1. **Description of methods and procedures used for data collection or generation e.g. survey (include links or references to publications or other documentation describing the protocols used)**
2. **Description of methods used for data processing. Provide information about any steps used for data cleaning, missing values, outliers, normalization etc.**
3. People involved with sample collection, processing, analysis and/or submission

Data-specific information

(this section should be created for each dataset, folder or file, as appropriate)

1. Number of variables
2. Number of cases/rows
3. **Variable list (include variable names and descriptions)**
4. **Units of measurement**
5. **Missing data codes and definition of each code/symbol**
6. Specialized formats or other abbreviations used

WGs can add extra information if they wish.

Data types

As stated above, the WGs utilize mainly surveys and corpus/document analysis, encompassing both textual and numeric data. Thus data will be comprised of text files, spreadsheets, images and datasets (see table below). Data produced by WG2 and WG3 will also include document analysis and imagery.

	WG 1	WG 2	WG 3	WG 4	WG 5
Data Types	Text Spreadsheets images	Text Spreadsheets Images dataset	Text Spreadsheets Images dataset	Text Spreadsheets dataset	Text Spreadsheets dataset
Preferred Data Formats	.qsf (Qualtrics) .pdf .csv	.pdf .csv .txt	.pdf .txt .tiff	.qsf (Qualtrics) .pdf .csv	.qsf (Qualtrics) .pdf .csv

These file types will be used for the following reasons:

PDF: The PDF format is recommended as it is free of charge and has high accessibility. Moreover, it is suitable for long-term storage as outlined in Phaidra. This will be used for (multimodal) text documents, including surveys.

QSF: Qualtrics offers advanced analysis and integration options. By storing original survey design data in this format allows future researchers to use the survey directly in the Qualtrics platform and to see the underlying coding used to create the survey

CSV: any numerical data will be stored in this format as it is the most interoperable across all quantitative data analysis software. Data will be cleaned before being saved in this format.

TXT: any data that can be analysed using content or corpus-based analysis will be saved in txt format for the purposes of interoperability.

TIFF: any visual data will be stored in this format as individual images. Where text and visual data appear together, as in WG3, this data will be separated into PNG and TXT files for storage but will be associated through markers in the TXT document, i.e. [SEE IMAGE 1.1]

To future proof this DMP, the following file types should also be used for other types of data:

- *For Audio*: WAV, FLAC
- *For video*: AVI, MPEG-2, MKV

Data storage space required

Phaidra is a large open-sourced repository hosted by the University of Vienna and the size of file types is not really of concern. However, if recording any videos is required, reducing the quality of the video will be necessary for storage.

4. Documentation and data quality

This section refers to the metadata and documentation that will accompany the data as well as the data quality control measures that will be used. Metadata and documentation information is provided for each of the WGs and information on data quality control measures are provided for the whole Action.

Metadata

Metadata is data that provides crucial information about the data of the study. This information helps other researchers to identify, discover, understand and use the data. The metadata standards adopted for this COST action are those outlined by the University of Vienna.

Metadata standards comprise a set of guidelines that specify how information on data should be presented. All WGs will follow the Dublin Core elements that will be put into a README file. Phaidra uses the metadata standards recommended by Dublin Core.

The **process of organising and maintaining the data** created and collected by the WGs plays a crucial role in finding, understanding, and re-using the data. Based on the type of data and the nature of the study, data organisation will be organised in a number of ways (i.e., conventions, version control, folder structures).

To enable re-usability of the project and ensure research transparency, a**ll WGs will produce README files to accompany their data** (archived in PHAIDRA) as outlined in Section 3.

In addition, the following data should be provided to the Data Custodian at the University of Vienna for storage purposes:

0. Title of the dataset

1. **Name/institution/address/email information for Principal investigator,** Associate or co-investigators or other people responsible for administrative tasks as follows:
    a. Author/Principal Investigator/Associate Investigator Information
    b. Name:
    c. ORCID:
    d. Institution:
    e. Address:
    f. Email:

2. Date of data collection (suggested format: YYYY-MM-DD)

3. Geographic location of data collection (city/region, State, Country)

4. Keywords used to describe the data topic

5. Language information

6. Provide, for each filename, a short description of what the data is about

7. Indicate the relationship between files in case more than one file are included in the data set (e.g. “dataset” or “study” or “data package”)

8. Date that the file was created (suggested format: YYYY-MM-DD)

9. Licenses (DECISION TO BE MADE about the Creative Commons licenses types: CC BY-ND or CC BY-NC-ND)

10. Recommended citation for the dataset (Include information such as the creator(s) or contributor(s), date of publication, title of dataset, publisher, identifier (e.g. Handle, ARK, DOI or URL of source, version-when appropriate, date accessed-when appropriate)

Consistency and Quality of Data

The **consistency and quality of data collection** will be controlled and documented with the use of data quality control measures. In order to achieve data consistency:

1) WGs need to proceed with data entry validation (by WG members themselves), with the peer review of data so that other researchers in the field can identify any biases or inconsistencies that may have occurred during the data entry or data collection.
2) Data should be accompanied by a Data Dictionary and Glossary (see below)
3) Data must be cleaned and in its final version before being sent to Phaidra
4) Data must be saved in the agreed file format (see section 3)

Data dictionaries and glossaries

Data quality may be enhanced through controlled vocabularies that WGs have agreed upon terms and phrases such as data dictionary and glossary for survey tools and datasets. For more information, see https://www.usgs.gov/data-management/data-dictionaries.

The purpose of a **data dictionary** is to outline what key terms or concepts mean. For example, bi/multilingual literacies would be accompanied by a definition so that others can make sense of the data.

The purpose of a **data glossary** is to outline the variables that were analysed in any qualitative or quantitative data, their coded name, their measurement unit, the values that they contain and any further description. An example is provided below:

Data Dictionary Template

Data DOI:
Data Title:
Data Type:
Data Description:

Variable	Variable Name	Measurement Unit	Allowed Values	Description
Age in years	*AGE*	*Categorical*	*8-12, 13-16, 17-19*	*Participants select age range for themselves.*

Templates are available on request for this COST Action.

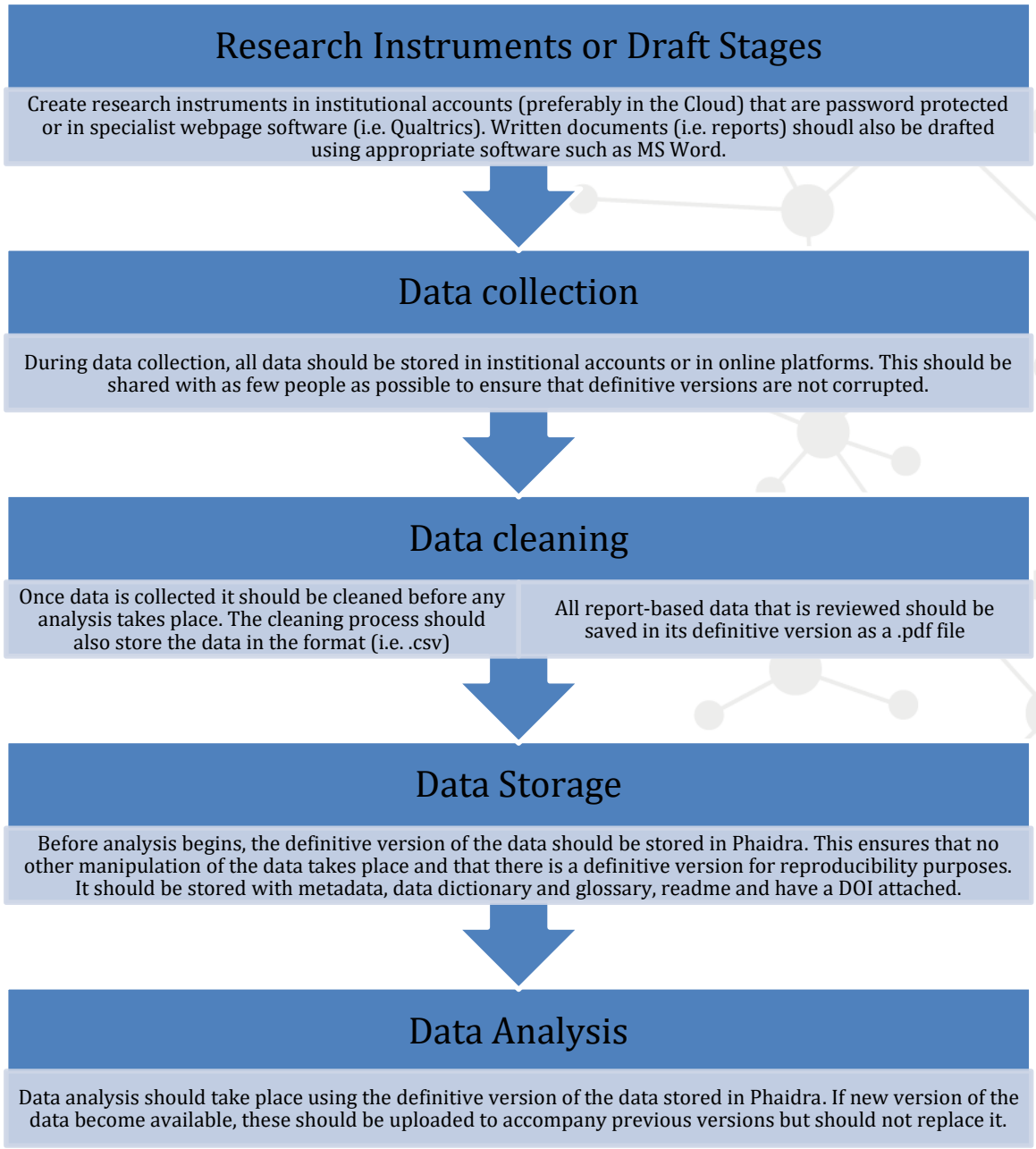
5. Storage and backup of data during research process

This section refers to the ways in which data and metadata will be stored and backed up during the research as well as how data security and protection of sensitive data will be taken care of during the research.

.

Data storage and backup

The proper storage and backup of the project's research data is of paramount importance for the long-term preservation of data and metadata. To safeguard the data and metadata during research activities, against any data loss, WGs will archive data in the following way:



By ensuring that data is always stored in a Cloud-based system or in Phaidra, we can ensure that there will always be an automatic backup available.

Data security and protection of sensitive data

Data security and protection of sensitive data should be meticulously taken care of during the research in order to ensure that data is protected and stored securely. In the event of an incident, **recovering data** is crucial. Phaidra is based on Fedora that includes the Veeam software, which can recover data in the event of an incident.

Concerning access **to the data during the research** and **how this access to data is controlled**, especially in collaborative partnerships, requires meticulous handling. In this project, WG leaders and members are the only ones with access to the data and data is controlled by restricting access to the files in Phaidra to different user groups. With respect to data protection, no personal or sensitive data will be collected and data will be anonymised at the point of collection (anonymous demographic data). Furthermore, all the servers where the data is stored are located in a room that is monitored and whose physical access is limited to authorized users. For more detailed instructions, see Section 8.

Lastly, with regard to the institutional data protection policies that are in place, see here the Data Protection Declaration of the University of Vienna.

Specific Data Storage Details for Each Working Group

Working Group	Types of Data that might be stored by each WG	Indicate how the data will be organised during the project, mentioning for example conventions, version control, and folder structures.
WG1	• blank version of the survey • blank consent form • signed consent forms • protocol for anonymisation • protocol for data collection • instructions for teachers • meta Vocabulary, • Qualtrics Survey File	• creating accessible content based on supported formats (https://datamanagement.univie.ac.at/en/about-phaidra/formats/), • using folders to sort out the files (conventions), • using descriptive and standardized naming conventions for files and folders. The filename should be short, simple and meaningful indicating what the document is about. Use underscores and hyphens instead of spaces and avoid the use of special characters. Use "V" to denote version and "R" to denote revision. Use the date format YYYYMMDD. • using README files to organise the data (conventions) • using a free version control system that records changes to a file or set of files over time (Git: https://git-scm.com/book/en/v2/Getting-Started-About-Version-Control OR Mercurial: https://www.mercurial-scm.org/).
WG2	• Existing learner corpora • the steps involved in building a corpus of language base on Sinclair Corpora standards • blank version of the survey • blank consent form • signed consent forms • protocol for anonymisation • protocol for data collection • Case Study Protocol (CSP) • meta Vocabulary, • Qualtrics Survey File	• creating a repository for centrally located storage purposes where the group can keep all of the project's corpora and other files or resources, • simplifying the secondary data used and by creating accessible content based on supported formats (https://datamanagement.univie.ac.at/en/about-phaidra/formats/).
WG3	• blank version of the survey • blank consent form • signed consent forms • protocol for anonymisation • protocol for data collection • Case Study Protocol (CSP)• meta Vocabulary, • Qualtrics Survey File	• using folders to sort out the files (conventions), by using sub-folders such as data, documentation and results (Folder structure), • using descriptive and standardized naming conventions for files and folders. The filename should be short, simple and meaningful indicating what the document is about. Use underscores and hyphens instead of spaces and avoid the use of special characters. Use "V" to denote version and "R" to denote revision. Use the date format YYYYMMDD. • using README files to organise the data (conventions), • using a free version control system that records changes to a file or set of files over time (Git: https://git-

		scm.com/book/en/v2/Getting-Started-About-Version-Control OR Mercurial: https://www.mercurial-scm.org/).
WG4	• blank version of the survey • blank consent form • signed consent forms • protocol for anonymisation • protocol for data collection • instructions for teachers • meta Vocabulary, • Qualtrics Survey File	• using folders to sort out the files (conventions), • using sub-folders such as data, documentation and results (Folder structure), • using descriptive and standardized naming conventions for files and folders. The filename should be short, simple and meaningful indicating what the document is about. Use underscores and hyphens instead of spaces and avoid the use of special characters. Use “V” to denote version and “R” to denote revision. Use the date format YYYYMMDD. • using README files to organise the data (conventions), and by using a free version control system that records changes to a file or set of files over time (Git: https://git-scm.com/book/en/v2/Getting-Started-About-Version-Control OR Mercurial: https://www.mercurial-scm.org/).
WG5	• blank version of the survey • blank consent form • signed consent forms • protocol for anonymisation • protocol for data collection • instructions for post-secondary stakeholders • meta Vocabulary, • Qualtrics Survey File	• keeping a chronological framework organizing the outputs received from the other WGs, • keeping a thematic framework organizing the outputs around a specific topic, • using folders to sort out the files (conventions), • using sub-folders such as data, documentation and results (Folder structure), •using descriptive and standardized naming conventions for files and folders.The filename should be short, simple and meaningful indicating what the document is about. Use underscores and hyphens instead of spaces and avoid the use of special characters. Use “V” to denote version and “R” to denote revision. Use the date format YYYYMMDD. • using README files to organise the data (conventions), and by using a free version control system that records changes to a file or set of files over time (Git: https://git-scm.com/book/en/v2/Getting-Started-About-Version-Control OR Mercurial: https://www.mercurial-scm.org/).

6. Legal and ethical requirements

This section delineates the legal and ethical requirements applicable to data collection, storage, processing and access, as well as to intellectual property rights.

Compliance with legislation on personal data and on security

When dealing with personal data, we must ensure compliance with data protection laws, such as the General Data Protection Regulation (GDPR) of the European Union. It applies to any data which can be linked back to the data subject (i.e. research participant), including survey data and written texts, when the data controller (i.e. data collector), data processor (i.e. data analyzer) or data subject is based in the EU. It mainly addresses issues of consent/assent, anonymization/encryption and access procedures, as will be delineated further below. National, regional, and institutional regulations may also apply, such as the Austrian Data Protection Act.

Informed consent and assent

Under the GDPR, personal data may only be processed if there is a legal basis to do so. Six such bases are identified in Article 6, of which the most relevant to our purposes is that "the data subject has given consent to the processing of his or her personal data for one or more specific purposes." This consent must be:

- unambiguous,
- specific,
- affirmative,
- and freely given by data subjects who are at least 16 years old, but member states may reduce this age requirement to as low as 13. When data subjects are underage, consent must be given by a parent or guardian.

In practical terms, this means that all data collection within the COST Action involving data subjects or their guardians must be prefaced with:

1) an information sheet detailing the purpose of the study which must include:
    a. the data to be collected,
    b. the contact details of the data controller,
    c. the period of retention for holding the data,
    d. data storage and access procedures,
    e. and participants' rights.

2) A place where data subjects or their guardians can sign a consent form either on paper or virtually (e.g., by ticking a box on an online survey item). Data subjects under 16 should also receive the information from the sheet in age-appropriate language before giving their informed assent (written or oral).

Informed consent requirements for each data type

Working group	Data name	Data subject	Consent required of adult	Assent required of minor
WG1	SURVEY: Questionnaire on CLIL practices	MC members	YES	NO
WG2	Compilation of EXISTING CORPORA on CLIL learners' productions	CLIL students	YES - SHOULD HAVE BEEN GIVEN DURING INITIAL DATA COLLECTION	YES - SHOULD HAVE BEEN GIVEN DURING INITIAL DATA COLLECTION
WG3	SURVEY: Teachers descriptions of multisemiotic items	CLIL teachers	YES	NO
WG4	SURVEY: Digital Literacies Teacher Questionnaire	CLIL teachers	YES	NO
WG4	SURVEY: Digital Literacies Student Questionnaire	CLIL students	YES	YES
WG5	SURVEY: post-secondary stakeholders	Post-secondary stakeholders	YES	NO

Examples of ways in which informed consent can be gathered are found here.

Anonymization, pseudonymization and encryption

One decision to be taken regarding the storage of the data collected by each working group is whether the data will be **anonymized, pseudonymized or encrypted.** Of these three methods, the most secure is **anonymization,** which is advisable whenever we will not need to link the data subjects' contributions to any future data they may provide.[1]

As mentioned above, the information sheet must inform participants about how their data will be stored (i.e. anonymized, pseudonymized or encrypted) and who will have access to it. If the information obtained can be linked back to the data subject, then that subject has certain privacy rights under GDPR which they must also be made aware of, including the rights:

[1] When data is sufficiently *anonymized*, it can no longer be linked back to a discrete individual by any party, including the data controller. This means that any information which would make such a connection possible (e.g., names of participants, email addresses, full dates of birth, zip codes, telephone numbers, or a combination thereof), is either permanently deleted or simply not collected. In the introduction to the GDPR, it is stated that anonymous information lies outside the scope of the principles of data protection and is thus not regulated as personal data.

If data is not anonymized, it should be *pseudonymized,* in which case "the personal data can no longer be attributed to a specific data subject *without the use of additional information,* provided that such additional information is kept separately," (Article 4 of GDPR). In *pseudonymization,* any sensitive data which would allow an individual to be identified is hidden or replaced by tokens, while the rest of the data is left visible for processing. This process is reversible, and the sensitive data should be kept separate from the data set. For example, participant names could be replaced with unique identifiers, and the key linking identifiers to the names of the participants could be stored by a trusted third party. A special case of pseudonymization is *encryption*, a mathematical process whereby all the original data is rendered unintelligible without access to the decryption key, which again would be stored separately from the dataset.

- to revoke consent to data processing at any time,
- to view their personal data and obtain a copy of it,
- to access an overview of how their data is being processed,
- to request the erasure of their data, and
- to file complaints with the Data Protection Authority.

Since access to and erasure of personal data are only possible if the data can be traced back to the subject who provided it, these rights do not apply to anonymized data, but they do apply to pseudonymized and encrypted data.

Decisions to be made by WGs regarding informed consent process:

- **WG1 must decide** whether to anonymize, pseudonymize or encrypt the data coming from MC members on CLIL provision in their context. Internally, CA21114 members will likely be able to identify the individuals providing the data, since we know who the MC members are, but WG1 must protect their identities when sharing this data outside the Action.
- **WG2 should be able** to anonymize the texts going into their corpus, as they come from existing corpora, and thus the data subjects are unlikely to produce additional data for CA21114 in the future. We recommend that members providing corpora to WG2 ensure the complete erasure of any identifying information appearing in the texts before sharing them with the WG.
- **WG3 must decide** whether to anonymize, pseudonymize or encrypt the survey data coming from teachers, taking into consideration any plans to link their responses to future contributions from the same data subjects AND the privacy rights applicable to pseudonymized and encrypted data (outlined above).
- **WG4 must decide** whether to anonymize, pseudonymize or encrypt the survey data coming from teachers and students, taking into consideration any plans to link their responses to future contributions from the same data subjects AND the privacy rights applicable to pseudonymized and encrypted data (outlined above).
- **WG5 must decide** whether to anonymize, pseudonymize or encrypt the survey data coming from post-secondary stakeholders, taking into consideration any plans to link their responses to future contributions from the same data subjects AND the privacy rights applicable to pseudonymized and encrypted data (outlined above).

Managed data access procedures

Managed access data is not publicly accessible, but rather shared with the research community via a managed access procedure which is clear and transparent. A decision should be taken regarding whether to establish a managed access procedure for each of the data sets reflected in the table below. If a managed access procedure is desired, we must then establish the terms for accessing the data. For example, we may wish to require that researchers requesting access meet specific criteria

(e.g., having authored a peer-reviewed publication in a related field, forming part of CA21114, belonging to certain Working Groups, being an MC or Core member), and agree to a set of conditions (e.g., to not pass on data, to take security measures to protect data confidentiality, to acknowledge CA21114 as data providers, etc.). They must also agree to only use the data for its intended purpose (i.e. the one specified in the Information Sheet - see 4.a.i) in order to comply with the "purpose limitation" set forth in Article 5 of the GDPR.

Management of intellectual property rights and ownership

There is a decision to be taken regarding who will be the owner of the data collected and generated by each working group. The owner will have the rights to control access to the data. It could be, for example, only the leader of each working group, the leader and vice-leader together, or also include other members of the Action.

Access conditions

Decisions to be taken: After an owner has been determined for each data output (see 4.b.), they will need to decide who can access that data. Data may be

1. open, downloadable and editable,
2. open and downloadable,
3. open and read-only online,
4. closed but available on request (see 4.a.iii for more on managed access procedures),
5. or completely closed, i.e. only available to a pre-selected group of individuals, such as core members or members, members of their working group or members of the Action.

In the case of (4), the owner must also decide on the terms for accessing the data, namely the criteria that data users must meet and/or the conditions that they must agree to before accessing it (see 4.a.iii), as well as whether data access and re-use licenses will be issued on the basis of these terms.

In the case of (5), the owner must decide on the specific individuals who may access it.

Options (4) and (5) are more appropriate for personal data which may contain sensitive or identifying information, such as survey responses and student texts. However, the tools used to generate these datasets (blank questionnaires, prompts) may be open access.

The access conditions agreed upon to date are outlined in the table on page 23,, but should be confirmed.

Intellectual property rights

Within CA21114's data collection process, intellectual property rights must be considered as regards:

1. WG3's analysis of national curricula, and
2. WG3's use of semiotic resources from published textbooks in the responses to their teacher survey

In each case, we must ensure that we do not infringe on copyright law.

In 2), WG3 members will analyze national and regional curricula. These documents may be copyrighted but openly licensed. For example, the UK national curricula falls under Crown Copyright and is available under the Open Government License, meaning that it can be copied, published, distributed, transmitted, adapted and exploited commercially and non-commercially as long as the source of the information is acknowledged. When analyzing national/regional curricula, we should identify any applicable copyright and licensing restrictions to know what we can legally do with these documents when distributing or publishing our work.

In 3), the WG3 survey asks teachers to share their favorite semiotic resource from their favorite textbook and write an ideal student text about it. The images shared by teachers are likely copyrighted with all rights reserved, and thus could not be redistributed by CA21114. There is a decision to be taken regarding when to remove copyrighted images from our data (e.g., upon receipt, after analysis, before sharing with others, etc.).

Re-use of third-party data

When reusing third-party data, we must be aware of and comply with any protections affecting that data (e.g., Creative Commons).

The only occasion in which CA21114 plans to reuse third-party data is in the creation of a "CLIL corpus:" WG2 has asked CA21114 members to share their existing learner corpora with the Action to compile a large corpus of student productions. When members share their corpora, they will need to disclose whether they are protected under any licenses, so that we know what we are legally allowed to do with that data.

Furthermore, the Australian Research Data Commons (2019) notes that "a combined dataset will adopt the most restrictive condition(s) of its component parts—unless the individual parts can be easily identified and separated," (p. 8). This means that, if any of the corpora are protected by a more restrictive license, the final corpus as a whole would be held to those same conditions. For this reason, if any of the existing corpora are copyrighted or licensed, there is a decision to be taken regarding which to incorporate in the final corpus.

All of the above issues are summarised in this table and will replace this information once decisions have be made by the Leadership Team

Working group	Data name	Ownership	Access conditions	Intellectual property rights affected?	Re-use of third-party data?
WG1	Definition of bi/multilingual disciplinary literacies	Decision to be taken	Open and downloadable	NO	NO

	SURVEY: questionnaire on CLIL practices TOOL	Decision to be taken	Open and downloadable	NO	NO
	SURVEY: questionnaire on CLIL practices DATASET	Decision to be taken	Closed, on request Decision to be taken: Terms of access (criteria, conditions); data access and re-use licenses	NO	NO
	Written Report	Decision to be taken	Open and downloadable	NO	NO
WG2	Existing Learner Corpora (External)	Decision to be taken	N/A External	NO	YES Find out about any licensing or protections
	SURVEY: Initial Questionnaire TOOL	Decision to be taken	Open and downloadable	NO	NO
	SURVEY: Initial Questionnaire DATASET	Decision to be taken	Closed, on request Decision to be taken: Terms of access (criteria, conditions); data access and re-use licenses	NO	NO
	Corpus of Student production	Decision to be taken	Closed, on request Decision to be taken: Terms of access (criteria, conditions); data access and re-use licenses	NO	YES (based on existing learner corpora) Decision to be taken: which corpora to include
	Report	Decision to be taken	Open and downloadable	NO	NO

WG3	Analysis of national/regional curricula	Decision to be taken	Open, downloadable, and editable Decision to be taken: Editable by whom? WG3 members?	YES Identify copyrighted curricula and the licenses under which they have been shared to know what we can do with them	NO
	SURVEY: teachers descriptions of multisemiotic items TOOL	Decision to be taken	Open and downloadable	NO	NO
	SURVEY: teachers descriptions of multisemiotic items DATASET	Decision to be taken	Closed, on request Decision to be taken: Terms of access (criteria, conditions); data access and re-use licenses	YES Includes semiotic resources from copyrighted books Decision to be taken: At what point do we delete these resources?	NO
	Report (Survey)	Decision to be taken	Open and downloadable	NO But respect copyright of sources when writing report	NO
	Report (National Curricula)	Decision to be taken	Open and downloadable	NO	NO
WG4	SURVEY: Digital Literacies Teacher Questionnaire TOOL	Decision to be taken	Open and downloadable	NO	NO
	SURVEY: Digital Literacies Teacher Questionnaire DATASET	Decision to be taken	Closed, on request Decision to be taken: Terms of access (criteria, conditions); data access and re-use licenses	NO	NO

	SURVEY: Digital Literacies Student Questionnaire TOOL	Decision to be taken	Open and downloadable	NO	NO
	SURVEY: Digital Literacies Student Questionnaire DATASET	Decision to be taken	Closed, on request Decision to be taken: Terms of access (criteria, conditions); data access and re-use licenses	NO	NO
	Report on DLTS	Decision to be taken	Open and downloadable	NO	NO
	Report on DLSS	Decision to be taken	Open and downloadable	NO	NO
WG5	SURVEY: post-secondary stakeholders TOOL	Decision to be taken	Open and downloadable	NO	NO
	SURVEY: post-secondary stakeholders DATASET	Decision to be taken	Closed, on request Decision to be taken: Terms of access (criteria, conditions); data access and re-use licenses	NO	NO
	Report	Decision to be taken	Open and downloadable	NO	NO

Ethical issues and codes of conduct

Ethical issues affecting how data are stored and transferred, who can see or use them, and how long they are kept

We have a responsibility to protect the identities of data subjects during the storage, access and use of personal data, i.e. survey responses and written texts (see Sections 6 and 7)

1. **Data storage:** Personal data should be stored in its anonymized, pseudonymized or encrypted version. *Pseudonymization and decryption keys must be stored separately from the data.* If we choose to pseudonymize or encrypt, rather than anonymize, personal data**,** there is a decision to be taken regarding where such keys are stored and by whom (e.g., a trusted third party).
2. **Data access and use:** Once identifying information has been removed or separated from the data, there is a decision to be taken regarding who can access the data. See 4.a.iii for guidelines on managed access procedures. Whoever uses the data must comply with the "purpose limitation" on personal data which is set forth in Article 5 of the GDPR, i.e. they must only use data for the purpose specified in the Information Sheet which was provided to the data subject before data collection.
3. **How long data are kept:** The GDPR does not specify a time limit for the storage of personal data, but states in Article 5 that they must only be "kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed." In other words, we must delete or fully anonymize personal data once it has been used to fulfill the purpose stated in the Information Sheet. There is a decision to be taken regarding the purpose to be stated in the Information Sheet and when it can be considered to have been fulfilled.

Codes of conduct, institutional ethical guidelines and ethical review

We must follow national and international codes of conduct and institutional ethical guidelines.

COST requires that Actions adhere to established rules and principles, including ethical principles listed on pp. 6-7 of the linked document, namely that COST and participants to COST activities adhere to the European Code of Conduct for Research Integrity, and respect fundamental rights as by the Charter of Fundamental Rights of the European Union.

Additionally, the University of Vienna (UNIVIE), where CA21114 is based, has its own data protection guidelines to ensure compliance with GDPR, which link to the European Commission's guidelines on Ethics for Researchers, as well as these relevant support services:

1. Advising on research data management and storage from PHAIDRA-Services at UNIVIE
2. Expertise on the preparation, archiving and provision of data, counseling on data management plans, anonymisation and data acquisition from the Austrian Social Science Data Archive

The Faculty of Philological and Cultural Studies at the University of Vienna encourages researchers to apply for an ethics review only if "the publisher or funding body requires an ethics vote." The aforementioned COST rules and principles do not mention ethical review as a requirement. However, should we need one at any point, we can get in touch with the ethics committee at UNIVIE here.

Additional information on planning research projects at UNIVIE is available here. Importantly, UNIVIE requires that researchers processing personal data "must register the project in the record of processing activities of the University of Vienna before starting the collection." Information on this can be found here when one logs in with a UNIVIE account. This applies to any personal data collected in the surveys by WGs 1, 3, 4 and 5, as well as in the corpus compiled by WG2, since they will be stored in PHAIDRA.

7. Data Sharing and long-term preservation

Long-term data preservation

Regarding data storage members of CA21114, CLILNetLe will consult two sources: the GDPR and Phaidra. Specific focus will be given to the types of data collected.

In case Working Groups are collecting personal data from participants, according to GDPR all data to which participants have not given consent (i.e. trial data) must be deleted/destroyed and may not enter the principal dataset. Only data that is collected following the informed consent and/or assent process appropriate to the country where it is collected will be retained. The working groups will then need to decide how and when the data will be stored based on the type of data they collect.

According to Art 5. GDPR 1.f (Principles relating to processing of personal data) all personal data will be processed in a manner that ensures appropriate security of the personal data, including protection against unauthorized or unlawful processing and against accidental loss, destruction or damage, using appropriate technical or organizational measures ('integrity and confidentiality').

The data will be stored in Phaidra, the data repository attached to the University of Vienna, given that the origin of this COST action is located within this institution. The repository complies with Open Access principles and, where possible, the data collected as part of this COST action will be stored here.

Regarding non personal data (reports, publications, papers, datasets, etc.) each working group will need to propose and then decide how and where this data will be stored.

Sharing of data

Once cleaned, all data will be stored as a collection of data sources in Phaidra at the University of Vienna. Other considerations include:

1. The data will be retained indefinitely at the University of Vienna
2. The gatekeeper of the data will be the Cost Action Chair Prof Julia Hüttner
3. The data will be subject to license agreements and access will be controlled according to data type or output:
    a. Research instruments will be available on request and accessible outside the University of Vienna.
    b. Data dictionaries , glossaries, README files and metadata will be freely available
    c. Datasets will be available on written request to Prof Julia Hüttner at the University of Vienna. Agreements will be signed with regards to the use of the data, how it will be credited and published.
    d. Access to the data of any type will be postponed until the end of the COST Action. This embargo is to allow the action to fulfil its dissemination objectives.

Methods to access data

All data will be stored in file formats that provide ultimate interoperability between different platforms. This also reinforces the openness of the data.

Requests to access the data will be made to the University of Vienna via Phaidra and approved by Prof Julia Hüttner.

Application of DOIs

The citation of digital objects is part of good scientific practice. Research data should be cited in the same way as other sources of information, such as articles and books. Phaidra. For this purpose, persistent identifiers (Digital Object Identifier, DOI) will be applied all data objects which will be produced by CLILNetLe including this DMP. Applying a DOI to all data objects when they are uploaded to Phaidra will allow future users to easily track and use these objects in their work.

Ensuring the application of a unique and persistent identifier, such as a Digital Object Identifier (DOI), to each data set is essential for effective data management, citation, and long-term accessibility. Here are the steps and considerations for ensuring this:

1. **Data Management Plan (DMP):** Our DMP outlines how data will be collected, organized, documented, and archived. It also specifies the importance of using DOIs for data sets.
2. **Select a DOI Provider:** This is not necessary because Phaidra can do this for the COST action.
3. **Data Set Metadata:** We will create detailed metadata for each data set. Metadata will include information about the data's title, creators, contributors, description, date of creation, format, and any relevant licensing or access restrictions.
4. **DOI Registration:** When the data set is ready for publication or sharing, we will register it with our chosen DOI provider.
5. **Ensure Data Accessibility:** We will make sure that the data set is accessible online or through a repository depending on the data each working group is generating. We will also ensure that the data set and its DOI remain accessible and functional over time.
6. **Long-Term Preservation:** Depending on the type of data we will consider data preservation strategies to ensure the longevity of the data and the associated DOI. This may involve data archiving, backups, and periodic reviews of DOI links to ensure they still resolve correctly.
7. **Citation Guidelines:** We intend to provide clear guidelines on how to cite the data set using the DOI. This includes specifying the format for citing the DOI in publications and ensuring that data citations are included in research papers and other relevant works.
8. **Monitoring and Maintenance:** We intend to continuously monitor the data sets and their DOIs to address any issues that may arise. We will also consider updating metadata if needed, and renewing DOI registration if required.
9. **Collaborate with Data Providers:** If working with multiple organizations or collaborators, we intend to establish clear communication and agreements regarding DOI assignment and data management responsibilities.

10. **Compliance with Standards:** We will ensure that our data set and metadata comply with relevant data standards and guidelines, such as the FAIR principles (Findable, Accessible, Interoperable, Reusable).

By following these steps and best practices, we can ensure the effective application of a unique and persistent identifier like a DOI to each data set, promoting data discoverability, accessibility, and long-term usability in research and scholarly communication.

8. Data management responsibilities and resources

This section refers to the duties and resources allocated for data management within the action. Given that COST Actions are collaborative in nature, responsibility for the collection and storage of different data elements is delegated across the action. The roles required for this DMP include:

Data Capture: The member who is responsible for data capture coordinates identification of appropriate data type for research purposes, identification of who/where/when/from whom the data will/ will be collected, ensuring the reliability and validity of collected data, and ensuring the ethical precautions related to data collection process.

Metadata production**:** The member who is responsible for data capture coordinates definition of metadata standards, implementation of metadata capture process, development of metadata tools, and ensuring metadata quality (validity and reliability).

Data Quality**:** The member who is responsible for data capture coordinates definition of data quality standards, identification of data quality issues, improvement of data quality, and evaluation of data quality.

Storage and Backup**:** The members who are responsible for data capture coordinates determination of how the data will be stored, determination of data storage policies, and providing data security.

Data Archiving**:** The members who are responsible for data capture coordinates definition of data archiving policies and procedures, identification and monitor of data archiving methods and procedures, implementation of data archiving processes, including data migration and preservation, and ensuring long –term usability and accessibility of data

Data Sharing: The members who are responsible for data capture coordinates identification of data sharing policies and guidelines including permissions, establishment of data sharing collaborations, monitor of data sharing activities, and development or finding secure data sharing platforms

The delegation of these responsibilities is outlined in the table below.

	WG1	WG2	WG3	WG4	WG5
Data capture	WG	WG	**WG**	**WG**	**WG**
Metadata production	[to be decided in WG]	[to be decided in WG]	**[to be decided in WG]**	**[to be decided in WG]**	**[to be decided in WG]**

Data quality	[to be decided in WG]	[to be decided in WG]	**[to be decided in WG]**	**[to be decided in WG]**	**[to be decided in WG]**
Storage and backup	Data Custodian, University of Vienna				
Data archiving	Data Custodian, University of Vienna				
Data sharing	COST Action Leadership Team				

Responsibility of implementation of DMP

The implementation of the DMP is the responsibility of all members of the COST Action involved in any type of data collection. However, if further clarity is required for any aspect of the DMP, any of the authors listed at the outset can be consulted each of whom belong to one of the individual working groups.

Updates to DMP

This DMP will be updated periodically as and when new decisions need to be made with regards to the collection and storage of data. This will be a minimum of every 6 months during the lifetime of the COST Action.

Financial Resources for DMP implementation

No specific resources are provided for the implementation of the DMP; however, grants (VMG, STSMs) will allow researchers to implement parts of the DMP but should be applied for separately during the designated grant periods. Also, the Phaidra repository is free to use and will not incur any fees.