MeMAD Deliverable

D1.1 Data Management Plan

Grant Agreement number	780069
Action Acronym	MeMAD
Action Title	Methods for Managing Audiovisual Data: Combining Automatic Efficiency with Human Accuracy
Funding Scheme	H2020-ICT-2016-2017/H2020-ICT-2017-1
Version date of the Annex I against which the assessment will be made	3.10.2017
Start date of the project	1.1.2018
Due date of the deliverable	30.6.2018
Actual date of submission	29.6.2018
Lead beneficiary for the deliverable	Yle
Dissemination level of the deliverable	Public

Action coordinator's scientific representative

Prof. Mikko Kurimo
AALTO –KORKEAKOULUSÄÄTIÖ, Aalto University School of Electrical Engineering, Department of Signal Processing and Acoustics
mikko.kurimo@aalto.fi

MeMAD project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 780069. This document has been produced by the MeMAD project. The content in this document represents the views of the authors, and the European Commission has no liability in respect of the content.

MeMAD

Methods for Managing
Audiovisual Data

memad.eu

info@memad.eu

Twitter - @memadproject

Linkedin - MeMAD Project

Authors in alphabetical order		
Name	Beneficiary	e-mail
Sebastian Andersson	Lingsoft	sebastian.andersson@lingsoft.fi
Sabine Braun	SURREY	s.braun@surrey.ac.uk
Jean Carrive	INA	jcarrive@ina.fr
Maija Hirvonen	UH	maija.hirvonen@helsinki.fi
Harri Kiiskinen	YLE	harri.kiiskinen@yle.fi
Tiina Lindh-Knuutila	Lingsoft	tiina.lindh-knuutila@lingsoft.fi
Mikko Kurimo	AALTO	mikko.kurimo@aalto.fi
Jorma Laaksonen	AALTO	jorma.laaksonen@aalto.fi
Lauri Saarikoski	YLE	lauri.saarikoski@yle.fi
Raphaël Troncy	EURECOM	raphael.troncy@eurecom.fr
Dieter Van Rijsselbergen	Limecraft	dieter.vanrijsselbergen@limecraft.com

MeMAD
Methods for Managing
Audiovisual Data

memad.eu

info@memad.eu

Twitter - @memadproject

Linkedin - MeMAD Project

Abstract

This document describes the data management processes in the project Methods for Managing Audiovisual Data: Combining Automatic Efficiency with Human Accuracy (MeMAD)

Contents

MeMAD

Methods for Managing
Audiovisual Data

memad.eu

info@memad.eu

Twitter - @memadproject

Linkedin - MeMAD Project

Data Management Plan

v.1.0 M6/36

This document describes the current status of project's data management, and will provide the basis of further work on developing common data management practices during the project.

Development of this plan is an iterative process and will be continued. Next version of the Data Management Plan is due M18 of the project as Deliverable 1.4.

1. Data Summary

The purpose of data collection and generation within the project is to facilitate the development and evaluation of methods for multimodal analysis of audiovisual content. Very large part of the data that is gathered and used by the project is either already publicly available research data or proprietary, strictly licensed audiovisual data from industrial partners. The main data produced by the project is in the form of computer program code and algorithms, trained Machine Learning models, metadata for media produced by the ML systems, and processed AV content. In addition to that interview, observation and test data will be collected in user experiments.

The research and evaluation data the project will use is in three main formats:

a) audiovisual digital data

b) general metadata, subtitles and captioning aligned to audiovisual content

c) specific metadata describing the content of the audiovisual material

The project will generate data of following type:

a) annotated datasets of audiovisual data

b) program code and algorithms

c) trained models using neural networks and supervised machine learning algorithms

d) interview, observation and test data

In addition, there are intermediate data types used within the project that are not necessarily preserved:

a) AV content processed to a format more suitable for further analysis (resampling, transcoding, etc.)

b) Intermediate data types for metadata and AV aligned data (subtitles, content descriptions, etc.)

c) Datasets resulting from program code development.

d) User experience data relevant only for intermediate purposes.

MeMAD uses mostly previously created audiovisual content for research purposes; for testing, raw footage by project partners and external partners can be made available. Several freely available research licensed datasets are used by the different work packages for their own specific needs. Industrial partners within the project will provide datasets consisting of their own media for the project. External partners are invited to provide datasets for use in training and testing the systems and methods developed by the MeMAD consortium. A detailed description of the research datasets is provided in the Deliverable D1.2

The research and evaluation data is obtained from two major sources:

a) state of the art research data corpora that have been collected

b) publication-quality and published media from industrial and external partners

State of the art research corpora are obtained by each partner and work packages individually according to their own research needs. A list of the used datasets is kept centrally. Publication-quality datasets are obtained from industrial partners. Additional datasets are made available by some external partners. The deliverable 1.2 Collection of Annotated Video Data reports these datasets in detail. Within the project, a summary of the datasets is kept centrally, and the partners/ work packages (WP) are invited to mark the datasets they are using.

The size of the research and evaluation data sets is large. Current estimates are, based on the research and evaluation data sets defined during the first quarter of the project, that the largest research oriented datasets are tens of terabytes in size.

The data set would be of immense use for any other actors working with automatic analyses of audiovisual data, including general AI research, media studies, translation studies etc, as well as industrial actors developing methods of media content management.

2. FAIR data

Data management in MeMAD is guided by the set of guiding principles labelled FAIR. The purpose of these principles is to make data Findable, Accessible, Interoperable and Reusable.

In order to be findable according to these principles, the research data has to be described using a rich set of metadata. This metadata must then be assigned a unique, specified identifier, which will be registered in an indexed or searchable resource.

According to the accessibility principle, this set of metadata has to be accessible using standardized communications protocols that is free and universally implementable, and that allows for the authentication and authorization procedures when needed. The principle also dictates, that this metadata has to remain accessible through these means even though the dataset itself is not or no longer available.

The interoperability principles dictates that the metadata must use a formal and accessible language for knowledge representation that is at the same time also shared and broadly applicable. The vocabularies used in describing the data should also follow FAIR principles, and include qualified references to other metadata.

In order to further the re-usability of the data, the FAIR principles dictate that the metadata should be composed of a plurality of accurate and relevant attributes that are associated with their provenance and follow domain-specific standards. The metadata must be accompanied with a clear and accessible license for the use of the data.

It is understood, that data management as practiced currently in MeMAD does not fully conform to the FAIR principles. This document describes the current adopted practices within the project, in order to facilitate the integration of practices during the successive iteration of data management practices. The aim of MeMAD is to create an integrated set of data management practices during the project, and the FAIR principles will be used to guide the process of data management practice development.

2.1. Making data findable, including provisions for metadata

In the first phases of the project, no overarching naming scheme is used. For the legacy datasets produced by the project as deliverable D1.7 and aimed for wider dissemination after the end of the project, a naming scheme will be devised.

The specific naming schemes to be used in the preparation of the legacy datasets will be decided after M24 of the project, when preparations for the collection for legacy data is scheduled to begin.

Currently each WP and partner uses its own naming schemes:

AALTO

Aalto University provides data of different types, including automatically generated annotations for audiovisual data, and trained machine learning models. In the initial phases of the project the data will be named according to the internal conventions of the research groups. During the project we aim to adopt the best practices decided within the consortium.

UH

Metadata of the user data (interview, observation and test data) produced during the project will be made findable through FIN-CLARIN, a Finnish data repository which part of the international CLARIN consortium. Describing and naming the data will occur in compliance with the FIN-CLARIN guidelines.

EURECOM

EURECOM provides also data of different types including: ontologies and vocabularies that normalize the meaning of terms useful for describing audiovisual content; annotations resulting from an automatic transformation of legacy metadata or from an information extraction process run on the various modalities of the audiovisual content; trained machine learning models.

Most of the annotations data will be represented in RDF, a graph-based model standardized by the W3C, and will follow the linked data principles. This means that each node and vertex in the graph is represented by a dereferencable URI. We plan to adopt the base URI <http://data.memad.eu/> when defining the scheme for naming those objects.

SURREY

The University of Surrey advises researchers on how to make their data findable. This includes, for example, advice on

- Creating data statements ensuring that data is clearly labelled and described with regard to the terms on which the data may be accessed, any access or licensing conditions/constraints, and legal or ethical reasons why data cannot be made available;
- Applying a licence to research data;
- Depositing research data into publicly accessible data repositories to enable researchers to make their data;
- Documenting data to ensure other researchers can access, understand and reuse the data. Including embedding of metadata.

Surrey dataset naming convention for the audiovisual data used in the MeMAD project will be along the following lines:

> [Surrey]_[MeMAD]_[Datasource]_[Genre]_[Dataset]_[version]

An example being:

> Surrey_MeMAD_ BBC_Drama_EastEnders_v.1

The University also creates an official publicly discoverable metadata record of where our data is held, such as in an external repository. Information regarding suitable places of deposit are kept up to date at an Intranet site of the University of Surrey.

YLE

Yle provided datasets are named as follows:

> [Yle]_[project]_[DatasetID]_[DatasetName]_[dataset_version]
> Yle = "Yle"; name of the company
> project = "MeMAD"; the name of the research project / context
> DatasetID = running number identifying the dataset (three digits, starting from 001).
>
> DatasetName = Human readable name to help identification. No spaces, no special characters, no underscore
> DatasetVersion = number describing the changes in the content.

Each dataset produced by Yle will include a set of metadata describing the dataset in RDF/XML format using DCMI Terms.

Yle datasets contain AV media and metadata describing it. Metadata is provided as XML files and mapping information between medias and metadata are included in the metadata.

Limecraft

Limecraft does not generate or bring into the project original audiovisual media files or prior available metadata, those are delivered by other partners in the consortium who will act as users of the Limecraft platform. In that case, we reuse names originally employed to identify the original media. Upon exporting this media from the platform, the same naming is reused.

Any metadata generated by the project's deliverables or by internal components of the Limecraft platform will be stored in an optimized database format not suitable for direct external use. However, all of this metadata will be accessible through the platform's API using the original media's naming identification of this information. When using scripting tools to perform these exports, Limecraft will ensure that the naming conventions used for the naming of the original media will also form a part of the naming for the derived metadata, e.g., "<original_clipname>_transcript_<language>.json" or "<original_clipname>_ner.xml".

The metadata generated and stored by Limecraft systems will be made available according to the exchange formats defined by WP6 (cf., D6.1, D6.4 and D6.7).

Lingsoft

Lingsoft will follow the format and naming conventions in media production industry along with best practices decided within the consortium.

INA

INA provides media and metadata following its internal conventions.

Media files encode one hour of tV or radio stream and follow the naming scheme <channel-id>-<yyyymmdd>-<hhmmhhmm>, yyyymmdd giving broadcast day and hhmmhhmm giving start and end hour. For instance, FCR_20140519_15001600.mp3.mp3 encode the radio stream of “France Culture” radio station on 19th may 2014, from 15:00 to 16:00.

Metadata are provided as CSV files, with various fields as : identifier of program, channel, start and end times, program title, summary, descriptors, credits, themes.

Mapping information between medias and metadata are provided separately as CSV files.

***

The practices relating to research metadata creation will be discussed in the further stages of the project. No obvious choices for a common metadata standard that could be adopted exist, and in the first stages of the project, the data management practices are very much related to individual Work Packages and their work. In the later stages, the requirements of the common prototypes will provide the framework within which the data must be managed, and the deliverables D6.1, D6.4 and D6.7 (Specifications Data Interchange Format, v. 1 to 3), will provide the guidelines on how to document the data.

2.2. Making data openly accessible

The data used and produced within the project can be divided in five groups according to differences in licensing and re-useability.

1. Research-oriented data obtained from public repositories
2. Research and evaluation data obtained from industrial partners
3. Annotated media data produced from groups 1 and 2 during the project
4. Algorithms and program code produced by academic research groups
5. Proprietary technologies developed by industrial partners.

Of these data types, the data in groups 3 “Annotated media” and 4 “Algorithms and program code produced by academic research groups” is the easiest to open for public re-use and will be made available as widely as possible.

Data in group 1 “Research-oriented data obtained from public repositories” often comes with a licence that does not allow re-distribution even though use for research purposes is free; this data is already available for research purposes, and therefore, a re-distribution within this project is not even desirable.

Data in group 2 “Research and evaluation data obtained from industrial partners” is typically published media data, which has strict licences concerning re-use and distribution, for example, tv-shows produced by broadcasting companies. This group also includes the user data collected during prototype testing. An open access publication of this kind of media is at least prohibitively expensive, at worst legally impossible, and in the context of MeMAD, the aim is not to make this data set publicly available to parties outside the project. Possibilities to re-license these datasets on terms equal to the ones used by MeMAD are pursued as default.

Data in group 5 “proprietary technologies developed by industrial partners” concerns tools and methods that the industrial partners contribute to the research project in order to facilitate and evaluate certain phases of the research. They reflect a considerable economic investment on the part of the industrial partners, and are aimed for developing further technologies and solutions with commercial purposes, thus not suitable for open distribution.

Project partners have currently different settings regarding the research data, and these will be described here.

AALTO and EURECOM

Both Aalto University and EURECOM have an Open Access Policy and strive to publish everything in as open way as possible. The same principle also applies to data and source code produced by these parties.

UH

University of Helsinki / WP4 will deliver reports on multimodal, multilingual, and discourse-aware machine translation. Any computational models or software developed in this process will be made available through freely accessible platforms like Github, keeping everything as open as possible as often as possible.

SURREY

The University of Surrey has a general Open Access Policy and aims to publish all research outputs as openly as possible. For example, Surrey publications and conference presentations will be deposited in the University’s repository and other suitable repositories. However, any audiovisual datasets used in MeMAD are unlikely to be open access due to licensing restrictions, although the annotations minus video data can be open access.

YLE

Yle provides the project with a selection of AV material and related metadata from it's broadcasting media archives. The AV material, altogether ca. 500 hours, consists of in-house produced TV programs. However, the rights of Yle are limited to typical business use such as broadcasting, and specifically do not include open distribution. A license agreement with the national copyright holder's organization is developed, which allows Yle archive material to be used freely within the MeMAD project and distribution of the material to researchers for project purposes. Based on this license, open access distribution of this media dataset is not possible, but the license agreement takes into account the need to make the project data FAIR.

The selection of programming metadata consists of a single month's TV programme metadata. This includes the times of transmission, content descriptions, classifications and the producing personnel. This data is not limited by copyright, but as the data has originated from in-house production processes for a specific use, it's opening may be limited by issues related to e.g. personal or journalistic data. Yle metadata set will be included in the project legacy open access, if no limitations to do this are identified during the project.

Limecraft

Concerning data of group 3, Limecraft will share the output from automatic analysis generated during the MeMAD project if sharing this is not prohibited due to business needs or the copyright restrictions of the original media they are based on.

Concerning data of group 5, most of the technologies Limecraft develops as part of MeMAD will not be made openly available by default. On the other hand, Limecraft will evaluate the open distribution of components developed during the project if those are parts that form an extension to a sizable existing open source component, or in case that the open distribution of a component makes sense economically, e.g., to enforce the commercial ecosystem that Limecraft intends to build around MeMAD technologies.

Lingsoft

Lingsoft will share the output from automatic analysis generated during the MeMAD project if sharing this is not prohibited due to business needs or the copyright restrictions of the original media they are based on.

INA

Since 1995, INA has been the legal depository of French television and radio. Legal deposit is an exception to copyright and INA has no intellectual property rights over the content deposited. The cataloging data (title, broadcast date, producer, header, etc.) are accessible for free, in accordance with the rules in force, by a search engine located on the site http://inatheque.ina.fr. INA also markets a collection mainly made of content produced by public television and radio stations, for which INA holds the production rights. INA thus offers broadcasters and producers excerpts and full programs, and pays back a contribution to the rights holders.

To promote research, INA provides for strictly research purposes (academic or industrial), various collections available on accreditation through the Ina Dataset web site (http://dataset.ina.fr). INA proposes to MeMad's partners, on the conditions of use described on Ina Dataset web site, a specific corpus of television and radio programs related to European elections in 2014.

INA also offers an open data collection of metadata on the thematic classification of the reports broadcast on the evening news of six channels (TF1, France 2, France 3, Canal +, Arte, M6) for the period January 2005 -June 2015), available at https://www.data.gouv.fr/fr/organizations/institut-national-de-laudiovisuel/.

****

During the early phases of the project, each project WP is responsible for its own data collection and storage. Partners providing research datasets will distribute the data using their own services. A central repository for all created research data is planned for the legacy dataset. It has not yet been decided whether this repository will be based at one of the research partner's own repository service, or whether some kind of public repository service is to be used. The final depository for research data remains to be discussed in the later stage of the project, mainly during the Project Task T1.3 during M30—36 of the Project.

The program data ("code") will be stored as a git repository, and can be accessed thus by both via the www-interface to the repository as well as with git directly. Documentation for the Git system is available on the internet for free, and the use of the program is discussed on several open forums worldwide. Program code used to analyse and process the datasets that is based on algorithms and techniques discussed and presented in scientific publications, is open source by default, and the released data sets will contain information on the relevant program code for their use. However, in the case of products intended for commercialization by the industrial partners, the release of the program code is not possible by default.

Research and evaluation data is distributed via suitable tools. As most of the previously prepared research datasets are available either as open access or via specific agreements, the partners using them acquire the data directly from the providers. Regarding research data from MeMAD industrial partners (Yle, INA), the partners have their own systems for distributing large datasets. INA data is available on the INA ftp server, and the Yle data will be distributed via a high speed file transfer service suitable for distributing large datasets.

The prototype applications developed during the project's first year will have specific needs for data transfer and distribution; these will be addressed and discussed during the phase of first installments during the period M6-M12. Technical solutions for distributing the legacy datasets will depend on the repository chosen for the legacy dataset deposition, and will not be the main concern of this project; these matters will be discussed during the relevant project task in M30-M36.

Such project legacy datasets that do not contain licenced nor sensitive information will by default be open for access to all interested parties, and therefore no restrictions will be imposed on their use. This does not apply to the proprietary media data provided by industrial partners. Whether it will be possible to have these as part of any kind of accessible legacy dataset is still an issue that needs to be discussed within the partner's own organization as well as with the relevant copyright representatives. In the case some kind of restricted distribution is deemed possible, the access will most likely be granted only by separate request to the parties holding the rights to the data, and will include the requirement of agreeing to terms of use for the data.

No need for a specific data access committee within the project is envisaged. The research data provided, while under a restrictive research licence, does not contain sensitive information on persons nor institutions.

Specific licensing issues will be addressed in combination of the legacy dataset creation in Task T1.3.

2.3. Making data interoperable

One of the main goals of the project is to create a set of interoperable research and evaluation data. The first six months of the project have as the main goal the creation of common interfaces and services for allowing the interoperation of data and tools among the research teams and data providers working in different countries. In practice this is rather straightforward, for the data is available in well-known and accessible formats.

In general, known best practices will be followed. As much as possible of the produced and used data is to be stored in formats that are well known and preferably open; structures text formats are preferred when suitable.

The standard definition is an important part of the first months of the project. As the first project deliverable (D6.1), a set of standards describing the formats for exchanging data is presented. This work is to be continued with further versions of the specification (D6.4, D6.7) These deliverables concern mostly the interoperability of the prototypes and frameworks within the project, which are proprietary technologies developed by the project partners.

Research data used within the project is easily usable, for AV material is delivered using well-known video formats, like MP4 and WMV, and metadata is distributed in structured text formats, like XML or JSON, which do not require proprietary technologies.

The legacy datasets will be described using a relevant metadata scheme, like DCMI. In the case it is necessary to create project- specific vocabularies/ontologies, mappings to commonly used ontologies can be provided.

2.4. Increase data re-use (through clarifying licences)

Data collected specifically for the project by its industrial partners as proprietary datasets, is strictly licenced, and in many cases the partners don't hold all the rights for the data or media. Therefore, it is highly difficult to license these datasets for open ended further use, especially under any kind of an open access license. Copyright societies granting licenses typically wish to limit the duration and scope of licenses in unambiguous terms, which doesn't favour open ended licenses that would be optimal for data re-use. Current approach is to acquire as open licenses as possible, and include in the agreement talks the idea and mechanisms for other parties to license the same dataset for similar purposes in the future.

In the cases where parts of proprietary datasets can be given further access to as parts of the project legacy dataset, their further use will most likely be limited to to research purposes because of business interests and IPR touching this data and media.

Interviews and user experience studies done in connection to the MeMAD prototypes may contain aspects which describe internal processes at the industrial partners. Opening this data for wider dissemination may result in disclosing information that has commercial interest, and may preclude these datasets from open distribution.

Data produced by the project itself can and will be open for re-use in accordance with the commercialization interests of the project industrial partners; this will take place either during or after the end of the project. Specific arrangements for peer review processes can and will be arranged when necessary.

3. Allocation of resources

As research data will be made FAIR partially as part of other project work, exact total costs are hard to calculate. Many of the datasets used already carry rich metadata, are already searchable and indexed, are accessible and presented in broadly applicable means and forms, are associated with their provenance and meet domain-relevant community standards.

Explicit costs for increasing FAIRness of the data are related at least to acquiring licenses for proprietary datasets in the form of license fees, but also in these cases part of the costs come from work associated with drafting license agreements and promoting FAIR principles among data and media rights holders and their representatives.

Direct license fee costs will be covered from Work Package 1 budget. Work hours dedicated to license negotiations and data preparation are covered from each partner’s personnel budget respectively, as they have allocated work months to Work Package 1 each. It is yet to be decided, how costs will be covered in cases where they benefit only parts of the consortium.

Each consortium partner has appointed a data contact person, and the overall responsibilities concerning data management are organized through work done in Work Package 1 dedicated to data topics.

Regarding the potential costs related to the long-term preservation of research data, these will be discussed in relation to the legacy dataset formation during the last year of the project.

4. Data security

In the first stages of the project, each WP or partner storing data has their own secure methods for storing data. Data is transferred either using secure cloud solutions, secure transfers over internet, or in the case of large datasets, specific secure download services or even physical transportation of the data on external media.

AALTO

All data collected and processed by Aalto University will be stored on internal network storage managed by Aalto University, or by CSC which is a non-profit state organization co-owned by the Finnish state and the Finnish universities. All data transfers are done using encrypted secure connections, and access to the files is restricted to project personnel.

UH

University of Helsinki stores the sensitive data on users it collects during the project on an internal/local network storage owned and managed by the University. This storage is secured and protected and access to it is restricted. If required, data sharing will take place using a sharing and downloading service specifically designed for transferring protected and non-public datasets securely over internet.

EURECOM

All data collected and processed by EURECOM is stored on internal network storage managed by EURECOM IT department. All data transfers are done using encrypted secure connections, and access to the files is restricted to project personnel. EURECOM servers are locked in dedicated room with a restricted badge access. EURECOM building is itself 24h secured within the campus.

SURREY

The University of Surrey provides mechanisms and services for storage, backup, registration and retention of research data during a research project and after its completion as part of the University's research data management policy. Data collected from users are anonymised and named under specific codes, which are also used for any annotations and files storing the data coding for analysis. These are stored separately from other datasets. All non-electronic data are kept in locked cabinets or drawers when not in use. Electronic data are stored on an internal network that is managed at Faculty level. Network access is secured using IT through password systems, file system access control, system monitoring and auditing, firewall, intrusion detection, centrally managed anti-virus and anti-spyware software, regular software patching and a dedicated IT support team overseeing all IT issues including data security and network security. All full-time and associate university staff are advised of data protection policies when they start working at the university. Research staff will normally have undergone research training (e.g. at PhD stage), which includes familiarisation with the UK research council code of conduction and the major principles of data protection.

YLE

Yle data is stored on an internal network share which is the same service as used for other company data and managed by Yle IT department. This storage is secured and protected and access to it is restricted. Data delivery will take place using specific sharing and downloading service specifically designed for transferring large datasets securely over internet. The data is delivered via personal download links, which can be requested, from Yle when needed.

Limecraft

Limecraft stores the project data either on storage in its internal network, or as part of the Limecraft Flow online platform infrastructure. For both environments, Limecraft follows the guidelines from ISO/IEC 27001 for best practices in securing data. Limecraft is also participant in the UK Digital Production Partnership and its “Committed to Security Programme”[1].

- Data stored in the internal Limecraft network is not accessible from the internet, except through secured and encrypted VPN connections. Access to this network is strictly controlled to only employees and storage systems require user authentication for access to data.
- Data stored as part of the Limecraft Flow infrastructure is hosted in data centers within the EU, and all conform to the ISO/IEC 27001 standard for data security. In addition to infrastructure security provided by Limecraft’s data center partners

[1] https://www.digitalproductionpartnership.co.uk/what-we-do/committed-to-security-programme/

(physical access controls, network access limitations), Limecraft's application platform also enforces internal firewalling and is only accessible for administration using dedicated per-environment SSH keys.
Any exchange of data is subject to user authentication and subsequent authorization (either from Limecraft employees which requires special access rights), or from clients who's access is strictly confined to the data from their own organisations. Additionally, any exchanges occur exclusively over encrypted data connections.

Lingsoft

Lingsoft stores the data it collects during the project on an internal network storage owned and managed by Lingsoft or third party data management providers within European Union. All storage is secured and protected and access to it is restricted. If required, data storage and management can be also restricted only to servers owned and managed by Lingsoft. If required, data sharing will take place using a sharing and downloading service specifically designed for transferring protected and non-public datasets securely over internet

INA

The INA corpus is made available to the MeMad project partners via a secure FTP server hosted at INA (specific port, implicit encryption over SSL). Each partner has been provided with a specific login.

***

The long term preservation of the data that is opened for further use is still an open issue. Project deliverable D1.7 is the legacy dataset resulting from the project, and our current aim is to store this in a repository that will be responsible for the long term storage of the data. Deliverable D1.7 is due in month 36 of the project, and plans regarding it will be specified in next versions of DMP. Media datasets provided by Yle and INA are parts of their archive collections, and will be preserved and curated through their core business of media archiving.

5. Ethical aspects

Part of the research data may contain personal information and it will be handled following guidelines and regulation such as GDPR. A Data Contact will be nominated and a contact point on personal data related issues will be set up to answer queries and requests for personal data related issues.

Metadata provided by industry partners may have issues related to the journalistic nature of the original datasets. Some of these datasets, such as the metadata provided by Yle, have been designed and intended for in-house production use of a broadcaster, and opening this data to outside users may result in needs to protect sensitive or confidential information stored within the data. These issues are resolved by removing and/or overwriting sensitive and confidential information in the research data set before delivering it to the project.

The user data (interview, observation and test data) collected during the project from experiments and authentic workplace interactions between human beings are sensitive data and will be protected and handled with proper care and measures (see MeMAD DoA, Chapter 5).

6. Other issues

As all MeMAD partners are established institutions, often with several decades of practices in data management, there are procedures in place, which play an important role in the data management practices, especially in the first stages of the project. These partner-specific issues have been described above in relevant sections of this document.