MeMAD

Methods for Managing
Audiovisual Data

memad.eu
info@memad.eu

Twitter - @memadproject
Linkedin - MeMAD Project

MeMAD Deliverable

D1.4 Data Management Plan, update 1

Version 1.0

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	8.5.2019
Start date of the project	1.1.2018
Due date of the deliverable	30.6.2019
Actual date of submission	1.7.2019
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 theMeMAD project. The content in this document represents the views of the authors, and the European Commission has no liability in respect of the content.

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	LLS	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
Liisa Tiittula	UH	liisa.tiittula@helsinki.fi
Raphaël Troncy	EURECOM	raphael.troncy@eurecom.fr
Dieter Van Rijsselbergen	Limecraft	dieter.vanrijsselbergen@limecraft.com

Document reviewers		
Name	Beneficiary	e-mail
Jorma Laaksonen	AALTO	jorma.laaksonen@aalto.fi
Raphaël Troncy	EURECOM	raphael.troncy@eurecom.fr

Document revisions			
Version	Date	Authors	Changes
0.5	20.6.2019	Lauri Saarikoski	Small corrections proposed by project internal reviewers.

0.6	28.6.2019	Lauri Saarikoski, Raphaël Troncy	Additional minor corrections, added URI policy

Abstract

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

This is the second, updated version of the data management plan. Overall focus of the document has been moved from individual partners to consolidated project-based descriptions. Practices defined during the first half of the project have been described, references to metadata and data repositories have been added.

Contents

Introduction - Data Management Plan Version 2.0 M18 6

Data Summary 6

FAIR Data 8

Making Data Findable, Including Provisions for Metadata 8

Making Data Openly Accessible 9

Research Partners and Their Data and Source Code 10

Industry and Commercial Partners and Their Data 10

Yle Dataset 11

INA Dataset 11

Making Data Interoperable 13

Increase Data Re-use (Through Clarifying Licences) 13

Allocation of Resources 14

Data Security 15

Ethical Aspects 16

1 Introduction

This document describes the current status of the MeMAD project's data management, and will provide the basis of further work on developing common data management practices during and after the project. Development of this plan is an iterative process and will be continued throughout the project. The last version of the Data Management Plan is due in M36 of the project, as deliverable 1.6.

The main changes compared to the previous version of the Data Management Plan are the following:

- Chapter 2: FAIR Data
    - Individual project partner descriptions have been replaced with consolidated project-based descriptions.
    - Practices defined during the first half of the project have been described, references to metadata and data repositories have been added.
- Chapter 4: Data security
    - Focus has been moved from individual partners to the project platforms.

Chapters 1, 3 and 5 contain no major changes.

These changes aim to incorporate the better understanding about the project's needs for the data management and to address the feedback from the project reviewers given in February 2019.

2 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. A very large part of the data that is gathered and used by the project is either already publicly available research data or strictly licensed audiovisual data from industry 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, interview, observation and test data will be collected in user studies and 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 the following type:

a) annotated datasets of audiovisual data

b) program code and algorithms

c) trained models using neural networks and machine learning algorithms

d) survey, 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 and development purposes, project partners and external partners provide additional audiovisual content. Several freely available research licensed datasets are used by the different work packages for their own specific needs. Industry 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) published or archived media from industry partners and external partners

State of the art research corpora are obtained by each partner and work package individually, according to their own research needs. Additional datasets are provided by project partners, mainly Yle and INA. The deliverable D1.2 Collection of Annotated Video Data describes 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 total size of the research and evaluation data sets is large. Current estimates, based on the research and evaluation data sets defined during the first half of the project, are that the largest research oriented datasets are tens of terabytes in size.

These data sets would be of immense use for any other parties working with automatic analyses of audiovisual data, including general AI research, media studies, translation studies etc, as well as commercial parties developing methods for media content management.

3 FAIR Data

Data management in MeMAD is guided by the set of principles labelled FAIR[1]. The purpose of these principles is to make data Findable, Accessible, Interoperable and Reusable.

It is understood that data management as practiced in the early stages of MeMAD does not fully conform to the FAIR principles. This document describes the current practices within the project and facilitates the integration of practices during the project. 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.

3.1 Making Data Findable, Including Provisions for Metadata

Requirements of the common prototypes will provide the framework within which the project data must be managed, and the deliverables D6.1, D6.4 and D6.7 (Specifications Data Interchange Format, v. 1 to 3) provide guidelines on how to document the data. Respectively, project collaboration and data exchange provide the guidelines for internal project data management, described in more detail in deliverable D1.3 Data Collection and Distribution Platform.

Currently, project data stored to the project file server follows a systematic folder structure where folder naming states whether the data is primary data or annotations, which collection or software component is originates from, version numbers, run timestamps etc. Each folder contains a machine and human readable file that follows the LDAP Data Interchange Format LDIF[2] and contains only elements following Dublin Core DCMI Metadata Terms[3]. A minimum set of metadata elements and folder naming conventions for the project are defined in deliverable D1.3 in detail. This aims to describe the project data semantically, interlink project data and annotations across work packages, and provide sufficient additional search handles for the project participants. This will also make the project result dataset in deliverable D1.7 easier to select and collect as dependencies between project sub-datasets, annotations and software versions have been recorded along the data.

One of the research data produced by the project is a so-called knowledge graph representing in RDF the legacy metadata associated to the audiovisual programs as well as some of the automatic analysis results. This knowledge graph follows the Linked Data principles, which means that every object is identified by a dereferencable URI. The project has established a policy to mint those URIs following some existing best practices from the (Semantic Web) community. First, the MeMAD ontology has for namespace URI <http://data.memad.eu/ontology#> with the recommended prefix to be “memad”.

[1] https://www.go-fair.org/fair-principles/
[2] https://tools.ietf.org/html/rfc2849
[3] http://www.dublincore.org/specifications/dublin-core/dcmi-terms/

Second, the general pattern for identifying metadata object is
http://data.memad.eu/[source|channel]/[collection|timeslot|series]/[UUID] where:

- source | channel (in lower case)
  - channel codes for INA: ['fcr', 'fif', 'fit', 'f24', 'fr2', 'fr5']
  - channel codes for Yle: ['tvfinland', 'yle24', 'yleareena', 'yletv1', 'yletv2', 'yleteema', 'ylefem', 'yleteemafem']
  - 'surrey' for the material used by the University of Surrey
- collection | timeslot | series (in lower case and in ASCII and slugified)
  - we replace: white space (' '), semicolon (':'), comma (','), slash ('/'), quote ('''), brackets ('(' or ')' or '[' or ']'), exclamation marks ('!'), interrogative marks ('?'), hash sign ('#') by a hyphen '-'.
  - we delete the consecutive hyphens to only have one, at most; we do not end by an hyphen; we do not start by a hyphen.
- UUID = MeMAD custom hashing function using a seed where:
  - seed for INA is "record ID" (of a program OR a subject)
  - seed for Yle is "guid" OR "contentID"

Finally, media objects are identified using the pattern
http://data.memad.eu/media/[UUID]

For the result datasets produced by the project as deliverable D1.7 and aimed for wider dissemination after the end of the project, a naming scheme for individual files will be devised to improve their findability. The specific naming schemes to be used in the preparation of the resulting datasets will be decided after M24 of the project, when preparations for the collection of resulting data is scheduled to begin.

Currently each WP uses its own naming schemes according to internal conventions of the research groups, typically following systematic structure that states e.g. data origin, version numbers etc. The aim is to make individual files findable and identifiable even when no additional metadata is provided.

The next section describes how the project data is meant to be distributed. Parts of the project data will be stored into open repositories and for the license-restricted datasets, metadata entries will be created into relevant data catalogues, currently CLARIN[4] and META-SHARE[5], which improves their findability. Once the repositories to be used have been chosen, the project will adjust its metadata guidelines to ensure compatibility with the target repositories.

3.2 Making Data Openly Accessible

The data used and produced within the project can be divided into five groups according to differences in licensing and reusability:

1. research-oriented data obtained from public repositories

2. research and evaluation data obtained from project industry partners

[4] https://www.clarin.eu/
[5] http://www.meta-share.org/

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 project industry 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 project industry 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. 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 project industry partners” concerns tools and methods that the industry 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 industry partners, and are aimed at developing further technologies and solutions with commercial purposes, thus not suitable for open distribution.

Research Partners and Their Data and Source Code

The MeMAD project strives to publish all its research in as open a way as possible. This principle applies to data and source code produced by the research partners within the project.

Industry and Commercial Partners and Their Data

Commercial partners in the MeMAD project will share the output from automatic analyses generated during the MeMAD project if sharing them is not prohibited due to business needs or the copyright restrictions of the original media they are based on.

Concerning the data of group 5, most of the technologies Limecraft, Lingsoft and LLS develop as part of MeMAD will not be made openly available by default. On the other hand, Limecraft, Lingsoft and LLS 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 cases where the open distribution of a component makes sense economically, e.g., to enforce the commercial ecosystem that Limecraft, Lingsoft and LLS intend to build around MeMAD technologies.

Yle Dataset

Yle provides the project with a selection of AV material and related metadata from its broadcasting media archives. The AV material, altogether ca. 500 hours, consists of in-house produced TV programs. 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 societies has been established, which allows Yle archive material to be used freely within the MeMAD project and also the distribution of the material to researchers for project purposes. Based on this licence, open access distribution of this media dataset is not possible, but the licence agreement takes into account the need to make the project data FAIR.

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

INA Dataset

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 commercial), various collections available on accreditation through the INA Dataset web site (http://dataset.ina.fr). INA proposes to MeMAD's partners, in relation to the conditions of use described on the INA Dataset web site, a specific corpus of television and radio programs related to the 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/.

While the primary data from the AV sets will not be openly accessible, the project will create metadata entries of these datasets into CLARIN and META-SHARE, accompanied with contact information needed for licensing and accessing these datasets.

During the project, created research data is first stored to the project's internal file sharing platform and selection of this data will be included in the project resulting dataset as deliverable D1.7. The final depository for research data remains to be discussed in the later stages of the project and the final decisions will be made in task T1.3 during M31-36 of the project. This repository will be taken into active use by the project as soon as it is available.

The program data ("code") will be stored as a Git repository[6], 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 project industry 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 project industry 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.

Technical solutions for distributing the project result 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 M31-36.

Such project result datasets that contain neither 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 or data provided by project industry partners. Whether it will be possible to have these as a part of any kind of accessible result dataset is still an issue that needs to be discussed within the partners' own organizations as well as with the relevant copyright representatives. In circumstances where 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 the 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, contains neither sensitive information on persons, nor institutions. The user data collected during the project is sensitive by nature, and person-related details will not be quoted or published. The data are used only for research purposes, and recordings in which persons may be identified will not be shown in public.

[6] https://github.com/MeMAD-project

Specific licensing issues will be addressed in combination with the project result dataset creation in task T1.3.

3.3 Making Data Interoperable

One of the main goals of the project is to create a set of interoperable research and evaluation data. The following have been selected as the interoperable data formats:

Data Type	Data Format	Explanation
video	video/mp4	video data
subtitles	Advanced SubStation Alpha	subtitles/captions for videos
ontology	text/turtle	ontology encoded in OWL/RDF
knowledge graph	text/turtle	RDF triples and named graphs, following a number of well-known ontologies such as EBU Core, NIF, Web Annotations, etc.
raw media analysis results	csv or json	media content annotations
structured data	application/xml	multiple uses
structured data	application/xml	multiple uses

Table 1. Interoperable data 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; structured text formats are preferred when suitable.

A set of standards describing the formats for exchanging data is presented as part of the project prototype work and it is reported in more detail in project deliverables (D6.1, 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.

The project result datasets will be described using well known ontologies including EBU Core, DCMI, and Web Annotations. In the case it is necessary to create project- specific vocabularies/ontologies, mappings to commonly used ontologies is provided.

3.4 Increase Data Re-use (Through Clarifying Licences)

Data collected specifically for the project by its industry partners as proprietary datasets is strictly licenced, and in many cases the partners do not 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 licence. Copyright societies granting licences typically wish to limit the duration and scope of licences in unambiguous terms, which does not favour open ended licences that would be optimal for data re-use. The current approach is to acquire licences which are as open as possible, and include in the agreement negotiations mechanisms for other parties to licence the same dataset for similar purposes in the future.

In cases where it is possible to extend access to a part of licensed datasets as an element of the project resulting dataset, its further use will most likely be limited to research purposes, owing to business interests and IPR impacting both data and media.

Licencing challenges affect mainly the primary data - video, audio and most ancillary data such as subtitles - but parts of this data, e.g. neutral metadata elements could and should be shared.

Secondary data such as annotations created during the project should be more straightforward to share for re-using. The project aims to share these, but it is yet to be decided whether they will be shared separately or as a part of deliverable D1.7 which is the collection of data resulting from the project. Also here some layers of licensing may be needed, as some types of annotations are closer to the original copyrighted data (e.g. ASR results) than other ones (e.g. extracted keywords).

Interviews and user experience studies conducted in connection with the MeMAD prototypes may contain aspects which describe internal processes at the industry partners’ organisations or sensitive personal information about the interviewees. Disclosing information that has commercial interest or sensitive personal information 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 industry 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.

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

Direct licence fee costs will be covered from Work Package 1 budget. Work hours dedicated to licence negotiations and data preparation are covered from each partner's personnel budget respectively, as they have allocated work months to Work Package 1 each.

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 project resulting dataset formation during the last year of the project (deliverable D1.7).

5 Data Security

Each of the project partners have their policies and means to keep the data safe on their sides with secure methods of storing and transferring the data and access control on shared data.

Project internal data platform is provided by INA and follows their security policies. This is described in more detail in the project deliverable D1.3.

The project prototype uses Limecraft Flow[7] as platform. Limecraft follows the guidelines from ISO/IEC 27001 for best practices in securing data. Limecraft is also a participant in the UK Digital Production Partnership and its "Committed to Security Programme"[8].

- 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 (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 whose access is strictly confined to the data from their own organisations). Additionally, any exchanges occur exclusively over encrypted data connections.

The long term preservation of the data that is opened for further use is still an open issue. Project deliverable D1.7 is the 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 during the second half of the project. 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.

[7] https://www.limecraft.com/

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

6 Ethical Aspects

The Project follows the guidelines for responsible conduct of research[9].

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, user studies 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).

[9] See for example the Finnish National Board on Research Integrity https://www.tenk.fi/en