MeMAD
Methods for Managing
Audiovisual Data

memad.eu
info@memad.eu

Twitter - @memadproject
Linkedin - MeMAD Project

MeMAD Deliverable

D1.7 – Collection of annotations and / or video data resulting from the project

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	23.6.2020
Start date of the project	1.1.2018
Due date of the deliverable	31.03.2021
Actual date of submission	31.03.2021
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
Jorma Laaksonen	AALTO	jorma.laaksonen@aalto.fi
Mikko Kurimo	AALTO	mikko.kurimo@aalto.fi
Lauri Saarikoski	YLE	lauri.saarikoski@yle.fi
Anja Virkkunen	AALTO	anja.virkkunen@aalto.fi

Document reviewers		
Name	Beneficiary	e-mail
Umut Sulubacak	UH	umut.sulubacak@helsinki.fi
Raphaël Troncy	EURECOM	raphael.troncy@eurecom.fr

Document revisions			
Version	Date	Authors	Changes

Abstract

This deliverable describes the process and output of curating and storing data resulting from the MeMAD project for post-project use. Different data types such as software source code, publications, annotations and machine learning models are described separately, each with their respective policies for curation and storage solutions. Main platforms selected for the results storage are GitHub for source code and Zenodo for other data items.

MeMAD project has also designed three datasets and licensed them for post-project use, based on the experience gathered during the project work with the proprietary data from Yle and INA media archives. These datasets combine the original media and metadata with annotations and alignments created by the MeMAD project. Datasets and their licensing mechanisms are described in the last part of this deliverable.

Contents

1 Introduction

This deliverable reports the work performed in Task 1.3 of the MeMAD project: the process of collecting, curating and storing project results and working material into selected repositories. Different types of data involved in the project are described in this document, as well as selected guidelines and storage solutions for each data type. The project data management plan (DMP, project deliverable D1.6) provides a good starting point for this work, as it covers the general principles and guidelines of MeMAD project data management. This document focuses on the practical applications of those principles and guidelines.

By the end of the project, an impressive collection of automatic media analysis and translation technologies have been collected to the project's main GitHub repository[1], providing an overview of the technical work done by different work packages. This also facilitates building on this work since the items are free to use. Other main project outputs have been gathered to the MeMAD project page on Zenodo[2], where the project public deliverables and main dissemination items will be openly available even after the activity on the project website eventually quiets down.

The third major element of the project results is the selection of datasets[3] that have been designed and licensed by the MeMAD project for future use, including use by 3rd parties. These datasets have been created based on the experiences gathered during the project from working with the Yle and INA licensed datasets, described already earlier in detail in deliverable D1.2. Since INA already has an existing framework for providing data to researchers[4], MeMAD efforts have focused on Yle data, extending the supply and coverage of in-domain datasets for media industry related research.

Chapter 2 of this document focuses on the project outputs and the data that has been created and gathered during the project. Chapter 3 describes the follow-up datasets that have been produced and licensed by the project for post-project use.

[1] https://github.com/MeMAD-project/mmca
[2] https://zenodo.org/communities/memad/
[3] https://developer.yle.fi/en/data/avdata/index.html
[4] http://dataset.ina.fr/

2 Resulting data from the project

This chapter describes which parts of the project data have been selected and curated for long-term storage after the project, and the platforms selected for this purpose. Large portion of the project data consists of source code and pre-trained machine learning models, but also project publications and different types of dissemination material have been judged to have potential value after the project has ended.

Using the project data management plan as a basis for this work, project teams reviewed outputs of their project work and decided which items need to be archived for post-project use and which can be discarded as e.g. intermediary development versions or drafts. These are the main guidelines the project followed in this work:

- The project should archive final or stable versions of the data.
- For items under active development, e.g. code on GitHub, forks or releases will be made to record the status at the end of the project and to provide a reference point for future use.
- The project should archive everything that has been defined as public in the project plan.
- The project should archive everything that is directly referenced in the project deliverables.
- Other items may be archived if judged relevant or potentially valuable for future use.
- Data from user evaluations is handled in line with the Research Information Sheets of each evaluation run.
- The project website will be made available until November 2024, but it will not be archived by the project as such. However, generic archiving services such as the Internet archive[5] may store the web site content permanently. Original pieces of content such as blog posts and demonstration videos may be archived separately, as well as some key dissemination items such as the list of publications and the project's final webinars.
- Cross-references between items on different platforms should be created, e.g. in cases where code is stored on GitHub and pre-trained machine learning models are made available on Zenodo.

Each project partner has been responsible for archiving the selected project outputs to the selected storage services. This work has been coordinated by Yle and it has taken place during the last 6 months of the project.

Since different types of project outputs have different target audiences, the platforms selected for storage and data selection differ slightly from one data type to another. The following sections describe these aspects by data type.

[5] https://archive.org/

2.1 Types of data

As listed in the data management plan, the project has produced data in the following categories:

a) Annotated datasets of audiovisual data

For project work packages, project partners have annotated small media datasets for testing and evaluation purposes both manually and automatically. Mostly these annotations are useful only in connection with the original media that has been annotated.

Annotations that do not include copyrighted material and refer to publicly available content are stored on the project's Zenodo page[6]. To some extent, these annotations are publicly available also through the MeMAD knowledge graph[7].

The Finnish Parliament has provided video recordings[8] and transcripts[9] of plenary sessions from 2008 onwards as open data. Aalto University has built a pipeline to align[10] the transcripts to the videos to create, at the time of writing, over 3000 hours worth of speech recognition training data[11]. In addition to training better Finnish and Finland-Swedish automatic speech recognition systems than before, the videos allow multimodal studies that take into account the visual modality. The same applies to speaker recognition and diarisation studies as the data also contains speaker annotations of more than 450 unique speakers.

Chapter 3 of this deliverable discusses in detail the cases where annotations refer to copyrighted media or include copyrighted elements such as transcripts or translations of the original content. For these cases, the data has been collected into three datasets, which have been licensed for further use.

b) Program code and algorithms

Code produced by the project has been gathered to the project GitHub site[12] in the status it has been in when the project work on the code has been completed.

c) Trained models using neural networks and machine learning algorithms

Machine learning models produced by the project have been stored on and shared through Zenodo, since many of these models exceed the single file size limit on GitHub and with less need for active version control features compared to software source code.

[6] https://zenodo.org/communities/memad
[7] https://data.memad.eu/
[8] https://verkkolahetys.eduskunta.fi
[9] https://avoindata.eduskunta.fi
[10] https://github.com/aalto-speech/fi-parliament-tools
[11] https://doi.org/10.5281/zenodo.4581941
[12] https://github.com/MeMAD-project

d) Survey, interview, observation and test data

The project has gathered different types of research data as part of the evaluations and user tests performed mainly by work packages WP5, WP6 and WP7. By default these data, such as interview recordings and transcripts, are seen as sensitive and will be destroyed after the project has ended, unless the research information sheet for the part of the project in question has stated otherwise.

Questionnaires used and interview scripts have been included in the project deliverables where those parts of the project have been reported, and they will not be archived separately.

e) Dissemination data such as publications, demonstration videos and web content

Public project deliverables have been stored to the MeMAD Zenodo project. Other publications are typically already stored at e.g. university or publisher repositories. A list collecting the project publications[13] has been stored in Zenodo for a comprehensive overview of the project, but individual publications have not been duplicated there. The list of publications includes links to the original repositories.

Selected key dissemination materials have been stored in Zenodo. The video recordings from the three project closing webinars organised in early 2021 are available in YouTube, and the project dissemination report stored in Zenodo will contain links to these videos.

3 Post-MeMAD licensed datasets

Part of the MeMAD project work has been providing the research teams access to proprietary media and data collections from project partners Yle and INA. One of the main benefits in this has been the possibility to develop, test and evaluate research and prototypes with in-domain data, bringing the project work closer to the realistic use environment and use cases than would have been possible with generic research oriented datasets.

Access to this data has been for the duration of MeMAD project and for the project purposes only. By using different subsets of data during the project for various purposes, both the data providers and data users have learned about the suitability of this data for different branches of research. As a result, project partners have been able to identify a number of datasets that could have potential use and value after the project. These datasets and the process of designing and licensing them are described in this chapter.

In general, easy access datasets for in-domain media industry data, especially the medias (videos) themselves, are scarce and biased towards major languages. Acknowledging this, and the fact that INA already has mechanisms in place to provide datasets to the R&D

[13] https://doi.org/10.5281/zenodo.4541252

community[14], the project effort was focused on data originating from Yle collections to expand the supply of media and data.

The market for media-related datasets in Finland is still evolving and the MeMAD project can act as an example of successful data licensing practices on data related to media and creative industries. The same effect can also be extended to other countries in the same stage of data market development.

During late 2020 the project partners gathered ideas for post-project datasets based on their experiences during the project. These ideas were refined into concrete propositions for datasets, out of which three most potential ones were selected for licensing talks with the rights holders. The three datasets are described in more detail in the following sections.

Criteria for selecting these datasets included:

- Size of the potential audience for the dataset: Who would be interested in using this data?
- Uniqueness of the dataset, e.g. in coverage of domain or languages: Is something similar already available from other sources?
- Amount of project effort that had already been invested in refining or enriching the data: some benchmark and evaluation datasets were engineered already during the MeMAD project and should be made available for wider use.
- How close was the relationship between the dataset and other MeMAD work? Was the data e.g. referred to in project deliverables or dissemination items?
- Availability of data at the source for the intended purpose: Is it possible to collect enough data for e.g. training ML-models, or is the dataset intended for test, evaluation and benchmarking purposes only.

In addition to preparing these datasets, the project identified that some parts of the project data should be publicly available, but rather in the form and function of a demonstration and dissemination item instead of a dataset. For example videos demonstrating the automatically translated subtitling developed in the project were embedded to a blog post by Yle[15]. This way they are publicly accessible and can be referred to in e.g. studies and teaching, but no additional licensing is needed since no copies of the videos have been distributed.

Based on the three project dataset propositions Yle, representing the MeMAD consortium, went through a number of negotiation rounds with the Finnish copyright society Kopiosto[16], representing collectively different rights holder groups of audiovisual media professionals. These negotiations took place during November 2020 - February 2021. While the basic terms of use for the data were relatively easily agreed upon, the amount of data and the duration of the licensing agreement needed to be revisited multiple times before the output was satisfactory for both parties.

[14] https://dataset.ina.fr/

[15] https://yle.fi/aihe/artikkeli/2021/01/14/korean-kriisi-ja-zombeja-tornadossa-ylen-kokemuksia-automaattisista

[16] https://www.kopiosto.fi/en/

Summarising the main points from these negotiations, both parties recognised the need to support research activities by making media and data accessible. However, it was partly unclear how large the demand for this type of data actually is, which in part led to this agreement being aimed to test the market. The resulting agreement can be described as a compromise between an acceptably low price, reasonable licence duration and dataset sizes, and well-enough contained risks from the rights holders' point of view.

One motivation in the negotiations was also to minimise the need for additional talks or management during the licence period caused by e.g. the quota of licensed projects running out or licences expiring before research projects using the data could be finished. Assuming that e.g. a doctoral thesis or a research project could be finished in five years time, the licence allows projects to keep using the data even after the initial licence contract period of 21 months has ended. The initial licence grants a maximum of 50 different research projects access to the data for free, which was judged to be a large enough number to simulate an unlimited access approach while still managing the risk of unexpectedly large demand for the data. If the need for data or number of projects requesting data exceed expectations, there is always the possibility to expand this initial licence agreement or negotiate additional licences.

As a general remark it must be noted that the licence agreement has some conditions and features that stem from Finnish legislation and the mandate of Kopiosto from different rights holder groups. For example, the content genres in the datasets were restricted to journalistic factual content to reduce the number of different rights holder groups included (leaving out e.g. actors). Also, the agreement is based on the legal framework of extended collective licensing[17], which is why the data access can only be initially granted to licensees located in Finland. However, they can share the data with international research project partners as long as all parties using the data comply with the original terms of use.

Compared to the original proposition by MeMAD, some modifications to both the licence terms and dataset contents were made during the negotiations. Agreement duration and dataset sizes were reduced to lower the overall price. The size of the dataset 3 was intended to be large enough for also training machine translation models instead of just evaluation and test purposes, but the licence costs for this was judged too high for the MeMAD project. Also the size of dataset 2 was reduced from 200 program hours into 60 hours to lower the licence price.

Final dataset versions are described in the following section, followed by a description of how to gain access to the data.

[17] https://minedu.fi/en/extended-collective-licensing

3.1 Dataset contents

Dataset 1: Yle media evaluation dataset

This dataset is a compact, multi-purpose dataset focusing on in-domain video and audio data. It includes the demonstration, evaluation and test content used in and enriched during the MeMAD project. It acts as the first public benchmark dataset for ASR, speaker diarisation and NER in the media-industry context for two low-resource languages: Finnish and Finnish-Swedish. It also provides video examples of different visual content types, supported by rich metadata. This data has been successfully used for these purposes already in the MeMAD work packages WP2, WP3 and WP5.

Contents

Part A: Audio, subtitles, transcripts and metadata of 14 test programs: 4 episodes of Pressiklubi and Obs Debatt each, 3 episodes of Strömsö and 3 EU-election studio debates, with a total duration of 549 minutes and 51 seconds.

Part B: Video, audio, subtitles, transcripts and metadata of 7 demonstration programs that have been annotated in the MeMAD project: Uutiset Lounais-Suomi, Spotlight, Sohvasurffaajat, Vallankumouksen lapset, Kuningaskuluttaja, Strömsö, Egenland, with a total duration of 3,41 hours. Different annotations created by the MeMAD project are included in this dataset, including automatically generated raw captions and captions after the caption post-processing described in deliverable D2.3, face detections and recognitions, speaker diarisations, language identifications, speech recognitions, visual location recognitions and audio background recognitions.

Dataset 2: Yle multimodal media and machine translations dataset

This dataset is designed to support the currently scarce market of professional visual media and multimodal translation datasets. It focuses on programs with subtitles available in multiple languages, with the added value of providing also the video and audio signals in addition to the textual content. The main languages in the dataset are English, Swedish and Finnish.

Contents

Video, subtitles and program metadata of 112 programs from Yle archive collection, with a total duration of 60 hours. The amount of content for each language pair are: FIN-ENG 16,58 hours, FIN-SWE 39,74 hours, SWE-ENG 3,46 hours (same items may be included in more than one language pair because of multiple parallel subtitle languages).

To complement these language pairs, the dataset includes also 5,98 hours of mixed media content to provide test content from few typical professional media types such as news broadcasts for the purposes of visual and audio analysis.

Dataset 3: Yle machine translated subtitles evaluation dataset

This dataset serves the needs of text-based machine translation research, with the focus on subtitles translation in the context of the professional media industry. It contains the benchmark dataset for subtitles machine translation engineered by the University of Helsinki during the MeMAD project, based on Yle subtitles. Included languages are Finnish, Swedish and English. This data has been successfully used in WP4 and WP6 of the MeMAD project for machine translation evaluation purposes[18].

Contents

Multilingual parallel subtitles from 44 programs, cleaned into 10,3k aligned sentence pairs. This dataset does not contain audio or video, but the total duration of the original media is 22,46 hours.

Table 1 summarises the dataset contents.

Dataset	Content	Modalities	Amount of data	Intended use
1 part A	ASR and NER benchmark programs	Audio, subtitles, transcripts, metadata	14 programs (+2 of the demonstration programs below), 9,16 hrs	Gold standard evaluation data for ASR, speaker diarisation and NER
1 part B	MeMAD demonstration programs	Video, audio, subtitles, transcripts, metadata, MeMAD annotations	7 programs, 3,41 hrs of content	Demonstration of MeMAD technologies with a variety of genres and languages.
2	Multimodal media and machine translations dataset	Video, audio, multilingual subtitles	59,95 hours of content	Testing and developing multimodal and multilingual analysis and translation technologies
3	Professional multilingual subtitles	multilingual subtitles, metadata	44 programs, 10.3k aligned sentence pairs (equal to 22,46 hours of content)	Text based machine translation testing and evaluation

Table 1. Contents of post-MeMAD licensed datasets

[18] See e.g. https://www.aclweb.org/anthology/2020.eamt-1.13/ and https://www.aclweb.org/anthology/2020.amta-pemdt.6/

3.2 Dataset technical details

All three datasets follow the same basic structure: media files and their original metadata form the core of these sets, and additional annotations and cleaned up versions of the data have been added to ease the use of these datasets. A readme-file documenting the dataset has been added to each dataset, explaining the data structure, semantics and background as well as providing contact information from the data provider. A copy of the terms of use document is included in each of the datasets.

For the first release of these datasets, the original metadata and media files are based on the Yle production formats to reduce the amount of data engineering needed to release these datasets. Video files are Yle house format in browse-quality mp4 files with frame height varying between 540 and 600 pixels and frame width varying based on aspect ratio in relation to this. The original metadata files follow the XML structure of Avid AXF files with sensitive or confidential metadata elements having been filtered out. Since the MeMAD project WP3 has produced a mapping between the Yle XML structure and the EBUCore standard[19], data users interested in the metadata elements should have sufficient support readily available to restructure the data into another format if needed.

For the convenience of use, alternative formats for some of the data have been included in the datasets, e.g. different subtitle tracks are available as standard SRT files even though all the subtitles are included in the original XML files.

Datasets 1 and 3 also include annotations and refinements to the original data produced by the MeMAD project, aimed to both disseminate the project results and to ease the productive use of these datasets. Dataset 1 includes gold standard annotations for ASR, speaker diarisation and NER, as well as demonstration annotations by the MeMAD project. These results are available as time coded text files and Advanced SubStation Alpha[20] (ASS) subtitle files. Dataset 3 includes a copy of the original subtitle data that has been cleaned up and aligned by the MeMAD project WP4. Formats of these annotations vary in the first release, depending on the data source, supporting the direct association of these annotations and the MeMAD technologies found on GitHub. All formats used have been described in the readme-files of each dataset. Based on the feedback from data users and in collaboration with them, the data can be further harmonised and standardised.

As an example, dataset 3 includes

- on the root level, a readme-file and a copy of the licence / terms of use document
- a subdirectory with the original XML files from Yle, named as

  <Yle unique media id>.xml

  e.g. “MEDIA_2013_00611491.xml”
- a subdirectory with copies of the original subtitles in SRT format, named with the same media id as above and with a postfix indicating the subtitle language

[19] https://github.com/MeMAD-project/rdf-converter

[20] http://moodub.free.fr/video/ass-specs.doc

- a subdirectory with a copy of the subtitle data in a cleaned up form by University of Helsinki, ready to be used for evaluation purposes. This data is summarised by a set of files named

  <src_lang>-<tgt_lang>.test.clean
  e.g. “FIN-ENG.test.clean”

  where the data is segmented into sentences and aligned across source and target languages.
- along the cleaned data, also additional raw data from the alignment steps of the cleaning procedure. These files allow users to implement their own alignment procedures on the same data, or custom filters on the provided alignment output, rather than using the summary files.

3.3 Framework for licensing and accessing the data

To provide access to these three datasets resulting from the MeMAD project, Yle has acquired a licence to distribute copies of the datasets. The licence has been granted by the Finnish copyright society Kopiosto, representing collectively the necessary rights holders.

This license grants Yle the right to distribute copies of the datasets for 1,75 years, until the end of 2022. To gain access to the data, users must register their projects and accept the terms of use at the Yle website[21]. Full terms and conditions of use can be found on the website[22]. To summarise the main points, the Terms of use

Allow

- The use of this data for research purposes within the registered project and by the registered project partners for the duration of the project (with maximum duration of 5 years). ‘Project’ can here refer to a number of research activities, including also e.g. research and benchmark challenges.
- Data mining for research purposes by the registered project.

Forbid

- The use of this data in projects other than the one registered
- Public presentation of this data. Citation in publications and presentations is of course allowed following the standard practices.
- Further distribution of this data to 3rd parties.

Obligate

- The data user to destroy all copies of this data when the registered project ends.
- The data user to ensure all possible project partners using the data comply with the Terms of use.

[21] https://developer.yle.fi/en/data/avdata/index.html
[22] https://drive.google.com/file/d/1m2xL4VQhdhEEyqyLp6Fu0TT0HO6h9t4P

Yle will host this service for the time being, but optionally this work could be taken over by 3rd party services specialised in this area such as FIN-CLARIN[23] or ELRA[24].

This framework and licence agreement will be in place until 31.12.2022. The rights holders see this as a pilot licence, and during the licence period all parties concerned will learn about the market for these types of datasets and other possible questions to consider in this line of work. While permanent access to the data would be the optimal solution for data users, this compromise was acceptable to all parties, keeping in mind that e.g. the European legislation on data mining and copyright is undergoing major changes at the time of writing this.

4 Summary

This document has summarised the work of making the MeMAD project outputs FAIR[25] from a data perspective. This serves also the purpose of project dissemination, as the project results and data have been made available on multiple public platforms.

By using well known standard platforms such as GitHub and Zenodo for project results storage, MeMAD has made its outputs findable, accessible and easily reusable. With the support of the project data management plan and the design of mentioned platforms, the project has sought to make the data also interoperable by e.g. using standard data formats wherever feasible.

Through clarifying licences for the Yle media archive data, MeMAD project has cleared the way for future projects, removing some of the typical obstacles projects encounter trying to find suitable in-domain data to work with.

[23] https://www.kielipankki.fi/language-bank/
[24] http://catalog.elra.info/en-us/
[25] https://www.go-fair.org/fair-principles/