MeMAD

Methods for Managing Audiovisual Data

memad.eu
info@memad.eu

Twitter - @memadproject
Linkedin - MeMAD Project

MeMAD Deliverable

D6.3 Evaluation report, initial version

Version 2.0

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	3.10.2017
Start date of the project	1.1.2018
Due date of the deliverable	31.12.2018
Actual date of submission	31.05.2019
Lead beneficiary for the deliverable	Limecraft
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
Inari Köngäs	YLE	inari.kongas@yle.fi
Tiina Lindh-Knuutila	LSS	tiina.lindh-knuutila @lingsoft.fi
Lauri Saarikoski	YLE	lauri.saarikoski@yle.fi
Dieter Van Rijsselbergen	Limecraft	dieter.vanrijsselbergen @limecraft.com
Kim Viljanen	YLE	kim.viljanen@yle.fi

Document reviewers		
Name	Beneficiary	e-mail
Mikko Kurimo	AALTO	mikko.kurimo@aalto.fi
Maarten Verwaest	Limecraft	maarten.verwaest @limecraft.com
Liisa Tiittula	Helsinki University	liisa.tiittula@helsinki.fi

Document revisions			
Version	Date	Authors	Changes
0.9	19/12/2018	Kim Viljanen, Lauri Saarikoski, Inari Köngäs	First version of the evaluation report.
1.0	28/12/2018	Dieter Van Rijsselbergen, Tiina Lindh-Knuutila	Finished missing sections on technical details, processed internal review.
1.1	24/05/2019	Dieter Van Rijsselbergen, Kim Viljanen	Major revision of the report, added clarification of methodology, aim of the evaluation and impact on the project plan.
2.0	29/05/2019	Dieter Van Rijsselbergen	Processed internal review of version 1.1.

Abstract

This is the initial evaluation report of MeMAD prototype, first of three, which reports end user feedback on the prototype and the underlying components that it makes use of. We interviewed eight media professionals by showing them processed results from the initial version of the MeMAD prototype and we collected their first impressions and feedback on how the MeMAD outcomes would best serve the various needs of the media industry.

The main outcome of the evaluation was that we able to confirm that the MeMAD use cases defined in the project plan align well with the daily needs of the industry. The feedback gathered did provide us with valuable input on how to improve the further execution of the project. In particular, we noted the impact on the potential use cases and user stories envisioned for implementation, as well as learned which considerations are important for the further definition of metadata exchange formats, and finally gained new insights into how the future evaluation can be aided by the input from our interviewees.

Contents

1 Introduction and aim of the first evaluation

This is the initial evaluation report of MeMAD prototype, the first of three, which will report end user feedback on the prototype and components it uses.

Because the project is still in early stages, the evaluation of the prototype focused on those components and specifications of the prototype that currently exist, including example data created with automatic content analysis modules and the representation of data as-is in the current implementation of the platform's user interface. More complicated features such as the workflows and orchestration of several analysis modules and application-specific user interfaces will be evaluated in the upcoming evaluation rounds.

The main goal of this first evaluation was to learn the following from media industry professionals through a series of interviews:

- Verifying basic assumptions made in the project, including its four main project use cases and the foundations on which they are constructed, in particular: the usefulness of multi-modal content analysis and (meta)data it produces, the potential to implement novel storytelling methods and the approaches proposed for applying multi-lingual machine translation.
- Learning about the validity of the sub-use cases and functional user stories devised in deliverable D6.1 with a limited but relevant user group. Any identified gaps in the considered use cases can then be evaluated by a larger group of stakeholders and can then be incorporated in the future work plan of the project.
- Learning requirements from end users about usable metadata in their existing production processes and the features those should support when choosing file formats or presentations of this data in GUIs used in the production process. Also, we wish to get a sense of how the quality of those metadata should be evaluated later on when more mature implementations of the MeMAD prototype are available for testing.
- Finally, we wanted to gather first impressions from the results produced by the initial iteration of the MeMAD prototype (as described in deliverable D6.2), both with respect to the metadata it produces at this stage, and the user interfaces it currently presents to end users. Also, we wanted to learn the usefulness and value of the proposed solutions in the interviewee's' daily work.
  This first round of evaluation will serve as a learning experience for designing the subsequent evaluations with users in subsequent evaluation rounds.

No exhaustive qualitative analysis of the MeMAD components for audiovisual content and metadata processing was made yet, as most versions we tested were available only in a first version and they have not been fully optimized for the use cases of the project yet. We will spend this effort later when we can compare different versions of the same components to gauge their individual and combined improvements.

We were also cautious not to make any final conclusions from the current state of the prototype, as no explicit new interfaces have been developed yet, rather we re-used existing functionality to fit in newly integrated components and the data they produce (e.g., video captions, NER results, audio classification). Regardless of this fact, we did want to show the current state of the MeMAD developments to a selection of media professionals already, aiming to initiate discussion about MeMAD's potential within their professional networks to hopefully lead to a closer co-operation in future years, both as part of the internal project evaluations but also as a way of boosting the project's later dissemination activities.

The content of this report is as follows. In Section 0, we describe the evaluation setup and method we used for gathering the feedback used in this report, including who was interviewed, how the interviews were conducted and how we selected the input data for the interviews. Secondly, in Section 0, we break down the results of the interviews in three sections: feedback concerning metadata, feedback concerning the project use cases and finally feedback concerning the use of the first prototype. We further analyze the results in Section 4, where we also translate our observations into the exact impact this evaluation has on the project work plan and its execution. We conclude with an assessment of this evaluation round and look towards the next actions in this work package.

2 Evaluation setup and method

The evaluation was done through eight interviews we conducted in December 2018 at the Finnish Broadcasting Company Yle. Interviewees were chosen from different areas of media production work in order to get feedback covering most of the typical workflows and user needs in a full service media company. We selected the interviewees so that their roles would match the MeMAD goal of increasing the usage or re-use of content, both as audiovisual media and its various metadata.

Interviewees work roles included:

1. Production co-ordinator (content description, reporting of content and music)
2. Archive editor / archivist (content description, archiving and information retrieval)
3. Translation co-ordinator (organises who translates and subtitles what media content)
4. Web analyst (analyses the statistics of content use and audience behaviour)
5. Online product owner (manages the development of an online service)
6. Video editor (editing of programs and inserts, archiving)
7. TV director (clipping and organizing footage, editing process)
8. Producer (TV-Producer has to know everything that goes on within a production)

Most of the interviewees were very experienced in their field of work. The average work experience in the media industry was 16 years, ranging between 1 and 28 years.

2.1 Interview organization and structure

The number of interviewees were selected to be quite small (eight), because the goal was to get a first impression on the usefulness of the results. According to standard practices in conducting user interviews[1], a relatively small number of interviews is enough to get a general overview of the strengths and weaknesses of a proposed solution. Already with this number of interviewees we started to see similar comments presented by several interviewees despite their varying professional backgrounds. Importantly, as the interviewees work as part of their professional community, and their answers also echo those of their colleagues and not just their personal opinions.

We conducted semi-structured interviews with each of these persons, lasting 30 to 60 minutes. Each conversation was guided by the online questionnaire which contained the interview questions (cf. Appendix 2) and provided the structure for the interviews. Interview notes were made in the questionnaire, and the interviews were also recorded for future reference and clarification in case information was left out of the notes. All interviews were done in Finnish and answers were initially noted down in Finnish.

[1] Mitchel Seaman: The Right Number of User Interviews, Medium 28.9.2015. cf. https://medium.com/@mitchelseaman/the-right-number-of-user-interviews-de11c7815d9, and also Jakob Nielsen: How Many Test Users in a Usability Study?, 4.6.2012, cf. https://www.nngroup.com/articles/how-many-test-users/.

While some numerical answers were also asked for, the interview contained mostly open-ended questions. The interviewees were encouraged to think out loud and present their ideas freely.

The interviews had a three-part structure:

- The first part concentrated briefly on the background of the interviewee including his or her role in the media production process and workspace at the company;
- In the second, main part of the interview the metadata examples and project use cases were discussed. The interviewees were asked a similar set of questions, focusing on the value and usefulness of presented data examples for the identified use cases and identifying potentially missing types of data. Based on the data examples, the interviewees were also asked what would be an optimal combination of metadata for their area of work, what this metadata should look like and if they can think of other uses for the data presented.
- Finally, an online demonstration of the current technical implementation of the prototype concluded the interview. In this part, users were asked for their first reactions on the live prototype and how well it would suit their specific work requirements.

After we had conducted all eight interviews, the notes of the interviews (and in some rare cases also the audio recordings) were analyzed in a qualitative manner with the goal to first, identify common themes between interviewees and second, find ideas for improvements and ideas for use cases. Section 4 of this report documents our findings of this analysis.

2.2 Selection of the evaluated metadata

During the interview, the interviewees were shown metadata and annotations produced by a variety of MeMAD components from the prototype platform - and asked to evaluate their value and usefulness for different aspects of their work. The components were selected based on readiness and availability. Each MeMAD partner had the opportunity to sign in their module for testing, but since the project is still in its early stages, not all planned components were yet available for testing. More details on which components were integrated into the first MeMAD platform iteration are described in deliverable D6.2.

The following groups of audiovisual content processing were selected for testing:

1. Automatic speech recognition (ASR) in Finnish, Swedish and English;
2. Named entity recognition (NER) from ASR transcript in Finnish and subtitles in Swedish;
3. Audio content analysis (AUDIO) containing speech vs music segmentation and sound classification;
4. Video content analysis (VIDEO) - automatically created video captions based on image analysis;

More details about the components can be found in Table 1 (cf. also D6.2 for more details).

Name of module	Short description of the module	Input	Output
ASR Finnish Lingsoft (1)	Recognizes Finnish speech	audio	timecoded text
ASR Swedish Lingsoft (1)	Recognizes Swedish speech	audio	timecoded text
ASR English (1)	Recognizes English speech (Not implemented by a partner in the MeMAD consoritium[2])	audio	timecoded text
NER Finnish Lingsoft (2)	Finds named entities in Finnish text	text	JSON highlighting the found term in the textual context
NER Swedish Lingsoft (2)	Finds named entities in Swedish text	text	JSON highlighting the found term in the textual context
Aalto DeepCaption (4)	Produces a short text description of image or video	video	text describing each video segment
Aalto Audio Tagger (3)	Produces the recognized sound events for each second audio/video	video or audio	text describing each audio segment's audio classification.
INA Speech Segmenter (3)	Splits audio streams into speech and music segments. Speech segments are labelled with gender information.	video or audio	structured text describing each audio segment's audio classification.

Table 1. Details on components used for creating the data examples for the interviews.

With the help of the above mentioned components, we analysed TV programs originating from Yle and INA archives, taken from the data sets discussed in deliverable D1.2.

The outcome of the analysis were metadata in various forms, which we chose eight examples to be shown to the interviewees for evaluation (cf. Appendix 1).

- Data example E1: ASR in Finnish
- Data example E2: ASR in Swedish
- Data example E3: ASR in English
- Data example E4: NER on Finnish ASR
- Data example E5: NER on Swedish subtitles
- Data example E6: Speech segmentation
- Data example E7: Audio tagging
- Data example E8: Deep captioning

[2] The ASR component for English was not developed by a consortium member, but rather the commercially available speech software from Speechmatics (cf. https://speechmatics.com/) was used for this purpose.

The specific examples were selected to reflect different strengths and weaknesses of the MeMAD modules, covering different languages covered by ASR components, rich variety of audio and visual environments and rich variety of topics covered in the programs. These include:

- Transcripts from unscripted politics-related interviews with mentions from places, nationalities (e.g., Romanian) and institutions (e.g., EU) which are hard to get right and are often mistaken for unrelated but similar sounding words.
- Current affairs programs with a great variety of depicted imagery which can make it challenging to offer accurate and relevant video content descriptions.
- Audio samples with mixes of music and voices which can complicate the audio classification as misclassifications of audio can easily happen.

2.3 Presentation of the evaluated metadata and use cases discussion

To properly focus the interviews on the potential of the metadata itself, we made the explicit decision to show each type of metadata to users in two formats.

The first form shown was as print-outs without being displayed as a part of a computer application. Considering the preliminary state of how certain metadata is presented in the first prototype of the MeMAD platform – metadata is shown using existing GUI elements, which are in many cases not optimized for showing the metadata in question in a complete or visibly pleasing form – we have opted to provide printouts with a more relevant markup, or a structured format which presents the available metadata in its complete form. This approach allowed us to avoid overly focusing the interview on technical implementation details of a certain user interface, when we really wanted to get feedback about the data and how it could provide benefits to different kinds of roles and processes in the media industry.
Later in the interview, we however also showed the metadata in question as part of the current implementation of the prototype, to also get feedback of the application.

As the output of the automatic content analysis components varied in formats and layouts, we considered this as an opportunity to learn about the users' preferences in different working contexts. Therefore, we did not unify the outputs of different components. Examples shown to interviewees were non-uniform and provided examples of alternate layouts and ways of representing data. For example, the three automatic speech recognition transcripts differed in the way text was split into paragraphs and whether speaker diarization or timecodes were visible in the printouts. This allowed us to benchmark interviewee preferences concerning the richness of the metadata presented and the way this would impact the usability of the data.

This part of the interview was structured to loosely follow the four project use cases (PUCs) of MeMAD project (as described also deliverable D6.1), but partially re-structured to better align with expected data use context and the work roles of the interviewees. In order to keep the length and scope of the interview manageable, the potential for each type of metadata (E1-E8) was cross-referenced with only a select number of logical groupings of use cases, instead of confronting our interviewees with the lengthy list of sub-use cases and user stories that we defined in D6.1.

The following groupings of use cases, we called them Work Scenarios, were used as the structure for this exercise:

- WS1: Searching and browsing media and data (both from archives or originally created content, cf. D6.1 sub-use cases 2.1, 2.2);
- WS2: Creating metadata and content descriptions (cf. D6.1, sub-use cases 2.3 and the creation stories from 2.2);
- WS3: Online service or application development;
- WS4: Subtitling, translations and accessibility (cf. D6.1, sub-use cases 4.1, 4.2, 4.3 and translation-related stories from 2.1).

From each use area perspective, the interviewees were asked a relatively similar set of questions, focusing on the value and usefulness of presented data examples for the use case in question and identifying potentially missing types of data. In particular, we were interested to learn:

- Which parts of the example metadata would be useful when performing tasks related to the working scenario?
- From this mentioned metadata, which would be the most useful?
- Looking at the metadata examples, what types of data are missing, if any? For which tasks would that data be useful?
- How would you grade the overall usefulness of shown sample data for the media production tasks in question? (1 to 5, 1 = useless, 5 = very useful)
- Provide a grade for the specific metadata example.

Finally, for each WS, we also asked the interviewees if they could think of other uses besides those tasks relevant to the WS for the metadata presented.

In the following section, we provide a break-down of the conducted interviews, organized first per section of the interview, and where applicable, per type of metadata that was evaluated.

3 The interviews

In the following we present the input from our interviewees, grouped into four viewpoints: 1) observations on the data examples, 2) data examples in the context of different work scenarios, 3) work scenarios compared with MeMAD use cases, and 4) first feedback on the current live version of the MeMAD prototype application.

3.1 Observations on the data examples

Based on the data examples, a number of improvement ideas were found concerning the data different MeMAD components currently produce. Some of these ideas might be expanded to guidelines what good quality data should look like when it is automatically created for different purposes.

Timecodes are essential for almost all purposes concerning video or audio. They should be present in the data every time, including also end times for events / annotations. Timecodes should be presented in a uniform format across technical components, e.g. following the SMPTE timecode format: HH:MM:SS:FF.

Names, identities and topics are needed to add value to the data. For example, an annotation saying “A woman is talking” should be expanded into a richer version that tells who the woman is, what is she talking about and in what tone of voice? Similarly, diarization into “Speaker1, Speaker2, Music” should be expanded to include speaker names and the musical piece or recording.

Some users and use cases would benefit from a data hierarchy or linked data structure where it would be possible to select the amount of detail based on the need. Some use cases require summaries on a very general level (e.g. the main topics of a whole TV program), other require as detailed data as possible, e.g. analysis on a frame by frame level or each single sound, word or pixel. The data should hence be a combination of all levels, combining, for example, transcripts with NER and linking these NER entities to global identifiers.

The desired representation of the data also varies between users and use cases. Some use cases would require the data to be shown in human readable form, other roles in development and analytics require a structured technical representation (e.g. JSON).

Many missing features were also identified in the dataset. For example, the mood of the content (such as feelings the content may represent or communicate), language, location, details on persons, colors, etc. These ideas are described in more detail in the following sections.

3.1.1 Observations on Automatic Speech Recognition data (data examples 1-3)

Use cases that could benefit from ASR were suggested in the interviews. For finding archive content or raw material, or for browsing consumer online services, transcripts could be used for navigating and finding right segments or quotes from the content. For content creation, organizing and clipping raw footage and creating subtitles could

benefit from ASR results. And for organizing content production, transcripts could be used for booking translators for the content, if the content languages were known.

In the context of end-user services such as over-the-top media services (OTT), transcripts could be used to power recommendation systems or search engine optimization. Transcripts could obviously be used also for accessibility purposes, to produce textual versions or subtitles of audio content.

Improvement ideas to the ASR data content and presentation:

1. Even though not all provided data examples included this, time codes were felt that they should be part of all transcripts. In addition to start times, also end times for each segment should be present.
2. Similarly, the transcripts should be divided to sections with section headlines.
3. The transcripts should show, when the speaker changes.
4. Transcripts without timecode or sections were considered to be slow to read.
5. When visualizing transcripts, it should be made clear which parts represent speech, and when the text is a description of what is happening in the image.
6. The optimal version would be a combination of many (meta)data sources. For example, transcripts combined with information about the visual image.
7. The location and context would be useful. For example, are the persons talking in a bathroom or in a war zone.
8. Are the persons talking alone or are there other people observing?
9. What kind of expressions and emotions are present in the speech? In which volume are the words uttered? These characteristics are typically not yet conveyed by the output from current ASR systems.
10. The identity of the persons talking should be identified. “Male3” or “female2” was not considered to be descriptive enough. The context of the identified person would be useful, e.g. is the person a celebrity or a politician in a specific party.
11. Different abstraction levels of the same transcript would make the data useful for more use cases. E.g. automatically identifying and highlighting the main keywords (topics, concepts) from the transcript. Keywords would optimally also be linked to entity registries, such as Wikidata. Or, for example, giving an automatic analysis on a more general level what happened in the interview, such as “the male participant was talking a lot, but saying little” or “the female participants described the truth, the male participants were making jokes”.
12. Transcripts from ASR represent the speech as spoken, so any colloquial language used is retained. One interviewee commented that to make the language publicly presentable, an automatic transformation from spoken language to written language could be useful, depending on the application. In fact, in some cases, when serving the deaf and hard-of-hearing, this would actually not be desirable.

3.1.2 Observations on Named Entity Recognition data (data examples 4-5)

Use cases for named entity recognition were already present above in the speech recognition examples, and in many cases it would make sense to combine ASR results with NER results, making this a candidate for further exploration when building new workflows in the MeMAD prototype. Focusing on identified entities, listing them could

give a viewer or a listener an overview on the topics and themes of the content. Entities could also be used in searching content, both in consumer and professional contexts.

Identified entities would be useful for analytical purposes. Such entities could be used to group and filter analytical data such as number of media play starts in an online service. Entities combined with time code would allow more detailed analysis of the internal structure of the media content, for example, while watching a program or for analytical purposes.

Improvement ideas for NER data:

1. Time codes should be added to the entities so that it becomes obvious where in the video or audio the entity is mentioned or detected.

3.1.3 Observations on audio analysis data (data examples 6-7)

Use cases suggested for audio analysis data involved searching and browsing media, and using audio analysis results as alternative search strategies such as finding sound effects or music based on style instead of music genres which are highly subjective. For example, video editors often need to find specific style of music or music with a specific instrument, or music that represents abstract concept such as “power” or “anger”.

Also finding music locations inside media content and identifying musical pieces for automated cue sheet creation were mentioned as potential use cases. And as audio events within content could be located and identified, close captions for hard of hearing audiences could be created based on audio analysis data, describing sounds and music currently playing in the media.

Improvement ideas for the audio analysis data:

1. Emotional information about the sounds. How does the sound feel? What feelings does it generate in the listener?
2. Information about the musical piece or the music recording (name of the recording, composer, lyricist, arranger, musician, year, location, Music Company...).

3.1.4 Observations on video analysis data (data example 8)

Use case suggestions for video analysis data included describing material, for example, in archive context. This would power searching for material and also finding the exact events inside a video, e.g. in a recording of a music performance. For accessibility purposes, video analysis data might serve in audio describing content for visually impaired people.

For analytics, the use of video descriptions could perhaps be used to analyze audience behavior: What content works, what does not? Why does some content get more views than other? Why do the audience leave the program after the first minute? It could answer whether correlations can be found between those statistics and what’s visible in the audiovisual content.

Improvement ideas for the video analysis data:

1. If the video analysis would be combined with transcripts of speech, the combination could help in understanding the underlying video content better than each separately. Use cases for this could include, for example, generating content descriptions, translating the content more correctly or searching the right moment in a long video recording.
2. The segments in the current version are too short for some use cases. There should be a way to choose the granularity of the segments - e.g. each frame vs. each shot vs. each thematic part of the program.
3. Dividing the data to segments with segment headings would make the data easier to read and more useful.
4. Characterizations such as adjectives or adverbs in descriptions would be useful.

3.2 Data examples in the context of different work scenarios

In the following, we present the interview findings from the context of the work scenarios (defined in Section 3.2).

We asked the interviewees to give a numerical score representing the value of the example data as a whole for the different work scenarios (cf. Table 2).

Role	WS1: Searching & Browsing	WS2: Metadata creation	WS3: Development	WS4: Subtitling, translations & accessibility	Average (if given)
Archive editor	2	3			2,5
Producer	4			5	4,5
Product owner	1		2	3	2,0
Production co-ordinator	2	3		4	3,0
Subtitle co-ordinator	4			3	3,5
TV Director	3				3,0
Video editor	2	2		3	2,3
Web analytics	2		2	3	2,3
Average (if given)	2,5	2,7	2,0	3,5	

Table 2: How valuable are the data examples as a whole for each specific use case? Scale: 1 (no value) to 5 (very valuable). Empty cells represent missing answers.

Based on the answers and the data examples shown in the interviews, the different data types would be most useful for subtitles, translations and other forms of accessibility such as audio description (WS4) but also for searching and browsing (WS1) and metadata creation (WS2).

Regarding the use in development, the interviewees pointed out that the value of the data itself is difficult to estimate. The data gets its value by using it in real use cases and

with real users. Therefore, the data as presented now is not that valuable for development, but it could have great potential when a real use case is identified, and the service and data are developed together.

The potential of the data is better visible in the Table 3, where the answers from all interviews are summarized according to technology and work scenario.

	WS1: Searching & Browsing	WS2: Metadata creation	WS3: Development	WS4: Subtitling, translations & accessibility
ASR	4	3	1	5
NER	2	1	2	1
AUDIO	3		1	1
VIDEO	3	2	1	2
Amount of interviewees answering this WS.	**8**	**3**	**2**	**5**

Table 3. Which of the example data would be useful to the specific use case? (combination of all answers). Table was created by analyzing the freeform answers of the users. Not all interviewees commented on all work scenarios.

To summarize the interviews and the Table 3, speech recognition was considered useful in all work scenarios. Video captioning data was considered especially useful in documenting video content, but with potential use in all other work scenarios.

People responsible for developing services or analyzing the usage of services would prefer data that is more succinct and compact - that is, for example, main keywords instead of full transcriptions of the whole program.

3.3 MeMAD use case validation

During the interviews, a number of ideas on practical specific use cases for the data was presented (cf. Appendix 3). Although some individual comments focus more on quality of data or individual feature wishes for the prototype, we considered the list of ideas as an opportunity to validate the MeMAD use cases (project plan) and user stories (of which a first list was defined in deliverable D6.1) defined earlier in the project.

We noted that the subject of each of the MeMAD project use cases was referred to multiple times, and each of the interviewees mentioned at least one of them:

- MeMAD use case 1: Consumer media services; 24 occurrences
- MeMAD use case 2: Digital media production; 37 occurrences
- MeMAD use case 3: Linking data to external resources; 11 occurrences
- MeMAD use case 4: Subtitling, translations and accessibility; 14 occurrences

This confirms that the MeMAD main use cases chosen for the project reflect user needs well, given that the users interviewed are media professionals, without a background in

research and development in general or the MeMAD project specifically. We will however need to gather a more tangible assessment of the applicability of each use case before making clear decisions on which ones to focus on in future developments, which is something we will do at a later stage, with representatives from more companies than just Yle.

Potential new use cases and user stories identified as the outcome of the interviews are, for example, related to online service development, content and service analytics and automation of certain production or publishing jobs.

Most of the sub-use cases described in MeMAD deliverable D6.1 are well aligned with the interviewee's' comments. Only the specific sub-use cases 3.3 "Validating content for truthfulness" and 3.4 "Linking relevant advertising to content" were not mentioned. For the latter one this is no surprise, as all interviewees work for a public service company that has no commercial advertisement on its platforms.

Examples connecting to the initial MeMAD use case 3 "Linking data between resources", were typically mentioned in the interviews as something enabling other further uses for the data. For example, automatic identification of a speaker was mentioned as a way to use the tools, but this appears to be more of a feature or a requirement to make the other use cases more valuable - not an independent use case of its own.

We further summarize the impact of this feedback in Section 4.

3.4 Prototype user interface

As the final part of the interview, the current version of the MeMAD prototype user interface was shown to the interviewees (cf. Figure 1, or in more detail in D6.2).

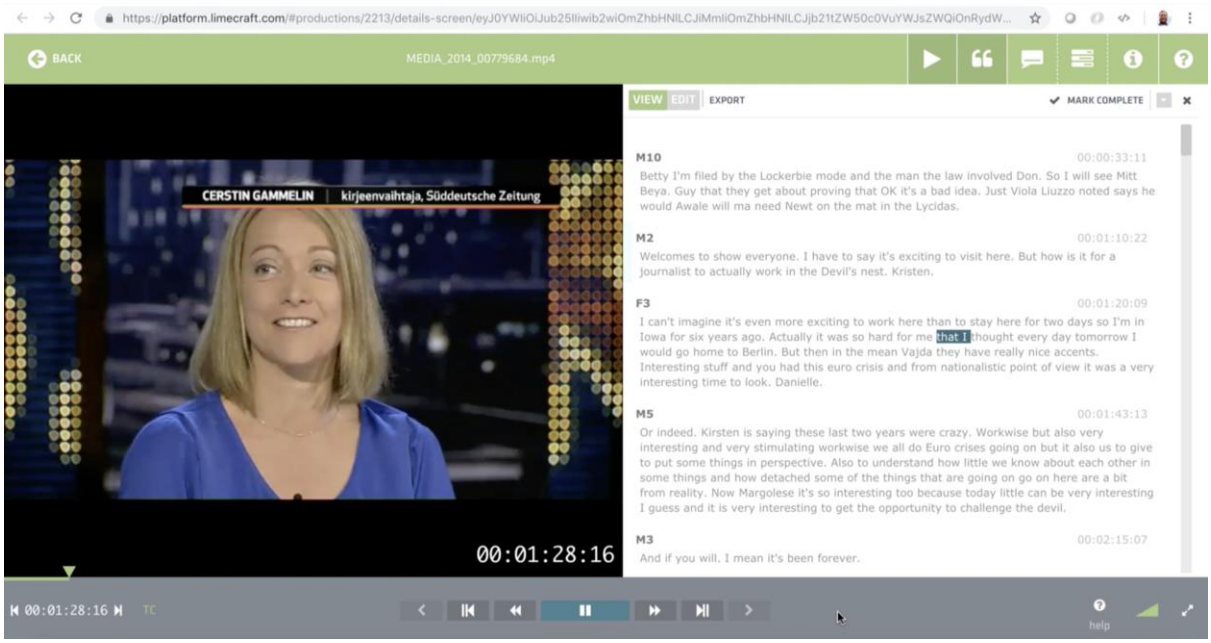

Figure 1: Screenshot of the MeMAD prototype.

The interviewees were asked for first reactions and also to give a numerical score on how useful the system shown would be for their work (cf. Table 4).

Role	Usefulness
Production co-ordinator	5
Archive editor	4
Subtitle co-ordinator	5
Web analytics	1
Producer	4
Product owner	1
Video editor	5
TV Director	5
Average	**3,75**

Table 4: How useful do you see the MeMAD prototype for your work? Scale: 1 (no value) to 5 (very valuable).

Based on the answers, the MeMAD prototype user interface would be useful as such for people working in different roles in media productions but for service development or analytics. The latter ones need the data as such, e.g. part of the analytic system or as an enabler for creating advanced new features for the audiences, but not the interfaces optimized for media production use.

Feedback from the users was generally positive and enthusiastic. The idea of showing the transcript next to the video was considered good, because it made following both the video and the transcript easy. The speech recognition quality was considered to be good, although some errors were also spotted.

The interviewees suggested that the user interface should contain more data, for example, a text description on what is happening in the image and what is the narrative structure of the video (e.g. “montage with music”).

One user would like to use the system for music so that the words sung by the singer would be shown next to the video (automatic speech recognition for sung music).

Another idea was to use the user interface for creating public content descriptions of TV programs in those cases when a script or (manually created) transcript is not available.

Other ideas were e.g. using the user interface for creating quotes from the program to the social media or using the user interface for directly doing video editing based on the text (as it is intended for even though we did not reveal that during the interview).

4 Analysis and impact on the project work plan

In the following we provide an analysis of the interviews, and we try in particular to summarize the impact that this first evaluation and set of interviews has on the further execution of the MeMAD project. Additionally, we also discuss our learnings and outline opportunities for future work concerning the evaluation process for MeMAD.

4.1 General discussion on the results

The key findings of the interviews are that there is a wide spread of media work contexts, each with individual needs and criteria concerning the data. In addition, current MeMAD technology components, use cases and prototype user interface are shown to be relevant to media professionals, but there is room for improvement in all areas. This is the first set of feedback that we will incorporate in the remainder of the project's execution in the following ways:

1. It will impact the use cases and user stories that are relevant to implement during the course of the project. Feedback from the interviewees provides us with additional use cases and applications to consider, as well as point out that some cases might not be as relevant as initially previsioned;
2. It also impacts the metadata use and formats specifications, which we need to take into account when defining final exchange format specifications, making sure that all features and requested data flexibility is correctly represented.
3. Finally, there's impact on the future evaluation of the prototype, which has become clear from gaining a better insight into the media production processes and the potential of the presented metadata produced by the MeMAD prototype.

Overall, for each data example given, at least one interviewee selected it to be useful. The media industry (at least based on these limited interviews) would welcome many kinds of new data about the media content, if it would be available, the quality matches the needs and if the data is presented in a correct way in the correct phase of the processes.

For some interviewees the data as such would be already useful, for others the data should be processed further to become valuable. This means that even though the first step of analyzing media content may be the same (e.g., speech recognition), for different use cases the data requires various types and amounts of processing to become useful for the end user working in a specific role in the media industry.

We did notice a clear discrepancy between the amount of suggestions made concerning the ASR data vs. the NER data in particular. Looking at Table 3, we also observe that the usefulness of NER data is rated much lower than the ASR data which clearly holds potential in all contexts of use. As far as we understand, this has two reasons. One is that the data itself is very simple, i.e., an identification of named entities in an existing text, so it begs less additional attributes to be available. One crucial attribute, however, has in fact been identified by the test panel, namely the need for timing information to be associated to detected entities. The other reason, we believe, is because the NER data acts as an intermediary to enable other kinds of applications, e.g., those linking between the source data (e.g., an ASR transcript) and Wikidata (as identified in Section 0) by means of

disambiguated and clearly identified entities. As such, this NER data by itself is less useful by itself, but it will be crucial to obtain in order to realize more advanced applications (e.g, those specified in Project Use Cases #1, #2 and #3).

Following the previous point, many interviewees have confirmed that they would like to have a combination of data in many cases. For example, speech recognition, video analysis, music recognition, person identification and location identification combined. Or combining this data with other data sources, such as a knowledge base (such as Wikidata in the particular case of Yle). Also, different users or use cases benefit from different amount of detail. For example, full transcript vs. a few keywords summarizing the main topic of the whole TV program.

This multitude of needs reflects the reality of especially audiovisual productions which is by nature a combination of many types of media (audio, video), many types of professionals (reporters, performers, directors, video editors, producers, web analysts, etc.), many types of technologies (light, audio, video, movements, ...), many types of genres and topics (sport, news, documentaries, drama...) and many types of audiences (children, adults, seniors, special audiences...). This all means that also the data should reflect a multitude of different viewpoints to serve the needs for different persons in different roles during different parts of the media creation, publishing and experience process.

4.2 Impact on considered use cases and user stories

In addition to the general observations above, we now summarize the specific impact the first evaluation has on various aspects of the MeMAD project, the first of which is the use cases and user stories defined in deliverable D6.1. The table below summarizes our findings and subsequent impact.

Observation	Impact
A potential usage context was identified concerning content analytics related to consumer behavior and analytics used for structuring of online services portfolios. In particular, applications could include grouping and summarization of associated metadata for use in analytics about media consumption: e.g., to analyze content play starts per character or named topics identified in the content.	We will take additional user stories under consideration focusing on service development and analytics, based on both content metadata and consumer behaviors. These will be added to the user stories defined for Project Use Case #1.
Use cases were suggested by the interviewees to use audio analysis data for searching and browsing media, e.g., by finding sound effects or music based on style instead of music genres which are highly subjective. For example, video editors often need to find specific style of music or music with a specific instrument, or music that represents abstract concept such as “power” or “anger”.	We will consider extending the existing user stories that cover music, presence of music, identification of music (which is already somewhat described in D6.1 – user story 2.3.1) with functionality to search for content by using audio attributes such as instruments, or style (to the extent these concepts are quantifiable and can be implemented in practice).
Many missing features were identified in the current dataset. For example, the mood of the content (such as feelings the content may represent or communicate), language, location, details on persons, colors, etc.	An additional user story will be taken under consideration which deals with looking up content by presence of certain detected emotions, both in the audio signal, and in the transcribed text.
Many interviewees would like to be provided with a combination of data. Additionally, this data can be combined with other data sources, such as a knowledge base (such as Wikidata in the case of Yle).	While we had identified the linking with external data mostly as useful resources for consumers, the use cases should be updated to reflect an interest in this topic for the media production use cases also, i.e., those represented in Project Use Case #2.
Different users and use cases benefit from different amount of detail in the data they use. For example, a full transcript vs. a few keywords summarising the main topic of an entire interview.	This requirement needs to be translated into one or more matching user stories which explicitly demand this layering of available metadata. While sub-use cases 1.1, 1.2 and 2.1 and 2.2 already hint somewhat in this direction, it needs to be made more explicit such that the consortium is better informed of the need for this functionality and such that it could become a distinguishing vital feature of the MeMAD prototype.

Table 5: Impact of evalution on the project's proposed use cases.

Those newly suggested use cases that featured most prominently in the interviews have been described in the table above, an exhaustive list of use case ideas has been attached in Appendix 3.

4.3 Impact on metadata usage requirements and formats specifications

The first evaluation also has an impact on the metadata exchange formats specifications and conditions surrounding the use of the metadata defined in deliverable D6.1. The table below summarizes our findings and subsequent impact.

Observation	Impact
Data should reflect a multitude of different viewpoints to serve the needs for different persons in different roles during different parts of the media creation, publishing and experience process. Different roles will require varying representations of the same data depending on the task being executed, which was already clear from largely different contexts of use (consumers vs. production, cf. D6.1), but it will also play an important part within the various tasks involved in media production itself.	This observation will have an impact on the metadata formats we define in subsequent versions of D6.1 (i.e., D6.4 and D6.7); if data is to be re-used between consumer processes or processes within the production chain, we need to ensure we can unify data as much as possible using a single data format and avoid using a variety of formats that require translations or conversions.
Many interviewees would like to be presented with a combination of data. For example, speech recognition, video analysis, music recognition, person identification and location identification combined.	We have observed a strong interest in multi-modality and the combination of various types of metadata. This is not only a fact for as far as actual processing and analysis of audiovisual content is concerned (which is one of the main objectives of the MeMAD project), but also holds for the presentation of metadata to users. This presents a challenge, namely to provide application GUIs that remain insightful and that can be used efficiently even with large amounts of multi-modal metadata available at every given point in time in a set of audiovisual content. This will form a key point to keep in mind while executing the User Centered Design (UCD) process for the further development of the MeMAD prototype.
Different users and use cases benefit from different amount of detail in the data they use. For example, a full transcript vs. a few keywords summarizing the main topic of an entire interview.	While some of the tools delivered in the MeMAD project provide very granular analysis results, for many tasks and users, this level of details is not relevant. It will be vital to define additional user stories which focus on this aspect and ensure the consortium can provide software components that are able to 1) combine granular metadata into summarized sections, thereby isolating those parts that 'belong' together to obtain useful segmentations, and 2) that can find

	those terms and named entities that can correctly summarize program sections while discarding less relevant or completely irrelevant descriptions.
Timecodes are essential for almost all purposes concerning video or audio. They should be present in the data every time, including also end times for events / annotations. Timecodes should be presented in a uniform format across technical components, e.g. following the SMPTE timecode format: HH:MM:SS:FF.	Deliverable D6.1 already identified the need for exchange formats that include the notion of timing information. However, it is likely even opportune to discard those formats for which no timing data is defined and push for formats that have the capability of incorporating timing information in any case. This will also force all partners in the consortium to incorporate the concept of time into their processing components and as such ensure that time-varying metadata becomes a first-class citizen in the MeMAD ecosystem.
Many missing features were also identified in the dataset. For example, the mood of the content (such as feelings the content may represent or communicate), language, location, details on persons, colors, etc.	In further defining the exchange formats for deliverables D6.4 and D6.7, we will ensure that these also are taken into account if the corresponding use cases are retained for development.

Table 6: Impact of the evalution on the project's metadata usage requirements and formats specifications.

4.4 Impact on the future evaluation of the prototype

Finally, we have also considered in which way the first evaluation will have an impact on future and more extensive evaluations of the prototype, specifically if we want to define proper evaluation frameworks for the various technology components in the MeMAD prototype which we can then use to organize repeatable and reliably quantifiable tests that reflect the expectations of end users for each use case.

Observation	**Impact**
The media industry would welcome many kinds of new metadata about the media content, if it would be available, the quality matches the needs and if the data is presented in a correct way in the correct phase of the processes.	This observation has important implications on the evaluation of the MeMAD prototype and performance of the underlying components. In particular, we must evaluate the accuracy of the machine-learning analytics results per use case or production process, as the requirements might be different for each. E.g., ASR results used by subtitlers will be subject to different (higher) demands for accuracy than for editors or journalists who will not feature the ASR output directly in the program. We need to define individual evaluation criteria for each use case and production context.
For some interviewees the data as such would be already useful, for	The impact of this observation is two-fold:

others the data should be processed further to become valuable. This means that even though the first step of analysing media content may be the same (e.g., speech recognition), for different use cases the data requires various types and amounts of processing to become useful for the end user working in a specific role in the media industry.	• First, it has been identified in D6.1 which components are dependent on which other components. An addition needs to made to ensure that we get a clear insight into the which data is to be re-used as input for processes that we might not have considered yet. • Secondly, as generated data will often serve as the source of a subsequent (automated) processing step, we need to measure the impact of error propagation in the quality of the end result of a processing chain, and the effect this has on usefulness of the results. E.g., if ASR is followed by a NER process, errors in the transcript could lead to misidentification of entities, or it could associate completely irrelevant named entity tags with a piece of content. Or even more in the case of translation, mis-detected ASR results could deliver blatantly incorrect translations. In practice, this means that measurements will need to be made at every stage of the executing process chain, as well as different source data needs to be presented to each stage (i.e., both un-modified and manually corrected data) to properly gauge the effect of propagated errors, and to determine under which circumstances the quality of finally produced data satisfies the end users.

Table 7: Impact of the evalution on the project's future evaluation process.

4.5 Reflecting on the evaluation process and conclusions

One of our goals for this first round of evaluation, was to learn how a relatively complex family of technology and process innovations should be evaluated with end users. Here are our findings of the evaluation process itself.

Conducting interviews in a rather early phase was useful even though the prototype was not yet fully implemented, and we had to use a more modular approach. Real users - professionals in their own fields, provided useful, fresh and authentic input on the project and the business needs of the daily work routines. This input will help guiding MeMAD in the right direction and helps in designing more relevant and better solutions later in the project.

Even though we only interviewed eight people, the input gave lots of ideas and partial confirmation that many of the original ideas of the MeMAD project are in line with the viewpoints of our interviewees.

The interviewees represented different work roles with different viewpoints on MeMAD's central theme of increasing usage and re-use of content. Since the number of interviewees was only eight, we could not however draw final conclusions on the interviews as such - they are only an indication for further work. Of course, the feedback gathered will be taken into account in the project work plan, as part of new or refined user stories, in defining better exchange format specifications, and even laying out a number of considerations for the future evaluations of technologies developed in the project.

One specific limitation of the interviews was that the participants were all working in the same company, at the Finnish broadcasting company Yle, which affects the results we got from the interviews, because the people are working in the same corporate culture, with similar kinds of tools and part of similar kind of processes.
Despite this limitation, the results are useful as they already gave us valuable feedback to steer the project plan and serve as basis for enquiring a broader group of users. The latter can now be done with incorporated feedback from the first test panel, instead of starting from viewpoints that were conveyed only by the core MeMAD development consortium.

4.6 Future work

In future evaluations of the MeMAD prototype we intend to include in the evaluation people from many other companies and organisations, not only Yle as in this first round of user interviews. For this, we aim to address the external collaborators group and involve them in the next round of evaluations.

In the next stage, we will shift the focus of the evaluation from the project use cases and potential applications to a more extensive evaluation of the interactive prototype (using the first interfaces developed specifically for MeMAD) on one hand, and a more structured approach with respect to assessing the quality of (chained) machine analytics results realized by the project's software components on the other hand.
Using detailed test cases for each relevant use case and end user workflow, we will perform qualitative assessments using both automatable objective metrics and by organizing structured assessments by end user panels.

In the final stage, the evaluation will be expanded to include also workflows and business cases (as identified in deliverable D7.1 - Business and exploitation plan), so that the evaluation can focus on value provided by the prototype to actual users. Evaluating data and current UI gives us a baseline and better sense of use case priorities, but actual value for the users comes from combination of data, data and software development and business development. Evaluating any of these separately provides only partial results and e.g. the value of different data types varies between different business cases. At this stage, we will also measure how production efficiencies are being influenced by the developed prototype.

Appendices

Appendix 1: Data examples

Sähkömoottorit ovat tulleet avuksi pitkille polkupyöräretkille, kuningaskuluttajan kolme kovaa, testaa tänään sähköavusteisen polkupyörän.
- Tavallinen ajokortti on tuhansien eurojen kallis hankinta, mutta autokoulujen vertailu on hankalaa, me selvitämme, mitä eroja, autokouluissa on.
Maaria Vatanen sai ajokortin ensi yrittämällä helmikuussa.
Hän käy kolmivaiheisen kuljettajatutkinnon harjoitteluvaihetta.
Muistot tossa hidastaa sen verta.
Vasemmalle, tuleeko sieltä.
Höyry lämmittämällä siihen.
Oikea.
Marja kerro miten sä oot valinnu auton no.
Mä oon tota.
Mun isän piti alun perin opettaa mua.
Itä-Suomessa, niin ei sinne.
Opettaja ei onnistunut.
Mä lähen vähän kiertelee noita autokouluun ja sit täällä Tampereella.
Taisin käydä kolmessa.
Koulussa ennen kuin mä sit tulin.
Siviili Trafigille ja mä just mä oon just aattelin, et mieti, mitkä ovat tärkeitä, miten mä haluisin siitä autokoulusta, niin mulla oli se, että et sä oisit meen tosi kiva.
Viihdyn siellä.
Et sä kuitenki oon semmonen.
Pidempi aikajakso, mitä käydään sitä korttia, niin ei ois ihan pakko pullaa seiniin.
Paljonko maksaa nyt sitten auton käyntiin?
Tulee yö vähän reilut pari tonnia sitten yhteensä et saa sen pysyvän kortin.
Suomalainen ajokortti maksaa kaikkine vaiheineen keskimäärin 2,5 1 000 €:a monessa muussa Euroopan maassa pääsee huomattavasti halvemmalla.
Kuka sun kortin maksaa.
Vanhemmat kyl se sieltä tulee sitten.

E1

Mycket i dag handlar om återvinning, Jim Lee och Camilla har gett varandra, en riktig utmaning de ska handla grejer åt varandra och som det inte ska vara nog ska de hitta allt material på loppis och det här med att gå på loppis och det kan ju vara lite som lotteri, ibland kommer man hem tomhänt men ibland blir en riktig jackpot, därför är det extra spännande med den här utmaningen dessutom ska vi grilla Härryda men först ska vi nog ta reda ut vad den här lopp.
Han är riktigt handlar om och med andra ord handlar åt varann.
Och det är hur mycket vi måste ha en gräns.
Vi har följts av människor och så får man använder som man vill, men eftersom fem Bueno
Men vem tjänar ingenting.
De reglerna om man köper vad som helst om man gör vad som helst, vad som helst måste man älska.
Men vad vill du att jag skulle vilja göra någonting sådant som till exempel ett serveringsbord eller en serveringsvagn eller någonting sådant så att till exempel vårt fjärde
Det var tydligt hur vill ni höra och se vad ni ska göra färdigt och jag kan inte se vad jag ska göra precis vad som helst.
Det är ett öppet kort när du tar emot i allt vi gör, gör vi självklart.
Har du nån att jag tror att jag kände att jag vill lura in mig på en färg.
Däribland Lasse Vibe
Noll
Och det kan man liksom kärleken och det syns bra.
Vi samarbetar de klubbar som utfördes vid en överraskning
De gör det inte lätt för sig kan man säga att vi följer Liddell tydligare koncept jo som alltid smörgås
För komplicerad men man kan göra en smörgås när vi ska faktiskt göra en halstrad smörgås och till att man kanske lite lätt rökt det som är så häftigt med det här tycker jag är att Hamburg är vi väldigt kända men en grillad macka som är som en riktig måltid och det har vi aldrig gjort innan man kan ju säga att hela måltiden följer den röda tråden för efterrätten blir jag också lidelser och det är ju just att den har så många positiva aspekter med sig men det kanske är ett av de absolut säkraste och bästa och kanske det mest smakrika.

E2

Male2 **00:01:10:22**

Welcomes to show everyone. I have to say it's exciting to visit here. But how is it for a journalist to actually work in the Devil's nest. Kristen.

Female3 **00:01:20:09**

I can't imagine it's even more exciting to work here than to stay here for two days so I'm in Iowa for six years ago. Actually it was so hard for me that I thought every day tomorrow I would go home to Berlin. But then in the mean Vajda they have really nice accents. Interesting stuff and you had this euro crisis and from nationalistic point of view it was a very interesting time to look. Danielle.

Male5 **00:01:43:13**

Or indeed. Kirsten is saying these last two years were crazy. Workwise but also very interesting and very stimulating workwise we all do Euro crises going on but it also us to give to put some things in perspective. Also to understand how little we know about each other in some things and how detached some of the things that are going on go on here are a bit from reality. Now Margolese it's so interesting too because today little can be very interesting I guess and it is very interesting to get the opportunity to challenge the devil.

Male3 **00:02:15:07**

And if you will. I mean it's been forever.

Male2 **00:02:19:07**

Once in a you wanted to draw the EU stars in your flag. You must be a true European if I.

Female9 **00:02:26:15**

It's because I'm Romanian so the flag stands for Romania for all those who are not very familiar with all the yellow flags I now saw that I am Eel's tyres because I work for actually a pan-European English speaking media. So I'm not a Romanian correspondent or.

Male2 **00:02:44:01**

The new MEP is going to be elected on Sunday and one of the most noticeable features of the campaigning has been that since the US economic situation in Europe has been difficult for years there seems to be a need to find guiltiest Daniel if you look at Portugal's media and public discussion who is or who are blamed. The most.

E3

```
  "KeywordPersonNames": [
    {
      "URI": "Lotta/person_names",
      "keyword": "Lotta",
      "source": "person_names",
      "relevance": 0.8008991627805376,
      "frequency": 1,
      "broader": [],
      "path": [],
      "context": [
        "Suonissamme virtasi vapaus munia ",
        "Lotan",
        " reissussa sohvalta ja maasta toiseen jatkui yhä syvemmälle lähin naapuri kylässä oli mykkä mies, joka pelasti."
      ]
    }
  ]

  "KeywordPlaceNames": [
    {
      "URI": "Libanon/place_names",
      "keyword": "Libanon",
      "source": "place_names",
      "relevance": 1.0835694555266098,
      "frequency": 1,
      "broader": [],
      "path": [],
      "context": [
        "",
        "Libanonin",
        " jälkeen otimme suunnaksi Jordanian sinne matkustaminen maanteitse Syyrian lävitse oli mahdotonta, joten turvauduimme lentämiseen."
      ]
    },
    {
      "URI": "Jordania/place_names",
      "keyword": "Jordania",
      "source": "place_names",
      "relevance": 1.0835694555266098,
      "frequency": 1,
      "broader": [],
      "path": [],
      "context": [
        "Libanonin jälkeen otimme suunnaksi ",
        "Jordanian",
        " sinne matkustaminen maanteitse Syyrian lävitse oli mahdotonta, joten turvauduimme lentämiseen."
      ]
```


E4

```
    "KeywordPersonNames": [
     {
       "URI": "Elin/person_names",
       "keyword": "Elin",
       "source": "person_names",
       "relevance": 1.3157629102823118,
       "frequency": 1,
       "broader": [],
       "path": [],
       "context": [
         "Går det smidigt, ",
         "Elin",
         "?"
       ]
     }
    ],
    "KeywordUnclassifiedNames": [
     {
       "URI": "Jag/unclassified_names",
       "keyword": "Jag",
       "source": "unclassified_names",
       "relevance": 1.3157629102823118,
       "frequency": 1,
       "broader": [],
       "path": [],
       "context": [
         "— ",
         "Jag",
         " knäckte en gula av misstag."
       ]
     }
    ]
   }
}
```


E5

NOACTIVITY	0	15.08
Music	15.08	29.78
Female	29.78	32.36
Music	32.38	98.1
NOACTIVITY	98.12	101.66
Female	101.66	148.16
Music	148.18	159.22
Female	159.22	163.14
Music	163.16	179.5
Male	179.5	183.06
Female	183.06	185.54
Music	185.56	221.46
Male	221.46	223.46
Music	223.48	237.16
Female	237.16	241.48
Male	241.48	247.26
Female	247.26	248.88
Male	248.88	252.72
Female	252.72	361.82
Male	361.82	363.08
Female	363.08	373.42
Male	373.42	375.92
Female	375.92	437.84
NOACTIVITY	437.88	438.28
Female	438.28	441.36
NOACTIVITY	441.4	441.98
Female	441.98	447
NOACTIVITY	447.04	447.5
Female	447.5	459.36
Music	459.38	462.04
Female	462.04	468.26
Male	468.26	470.62
Female	470.62	486.32
Male	486.32	488.98
Female	488.98	496.9
Male	496.9	497.94
Female	497.94	521.72
NOACTIVITY	521.76	522.4
Female	522.4	544.86
NOACTIVITY	544.9	545.48
Female	545.48	593.74
NOACTIVITY	593.78	594.28
Female	594.28	622.04
Music	622.06	627.06
Female	627.06	675.68

E6

00:06:12,001 --> 00:06:13,000 "Field recording"--"Male speech--"Echo"

00:06:13,001 --> 00:06:14,000 "Male speech--"Narration--"Field recording"

00:06:14,001 --> 00:06:15,000 "Male speech--"Field recording"--"Narration

00:06:15,001 --> 00:06:16,000 "Echo"--"Male speech--"Narration

00:06:16,001 --> 00:06:17,000 "Echo"--"Male speech--"Narration

00:06:17,001 --> 00:06:18,000 "Guitar"--"Musical instrument"--"Plucked string instrument"

00:06:18,001 --> 00:06:19,000 "Guitar"--"Musical instrument"--"Plucked string instrument"

00:06:19,001 --> 00:06:20,000 "Guitar"--"Musical instrument"--"Plucked string instrument"

00:06:20,001 --> 00:06:21,000 "Musical instrument"--"Guitar"--"Plucked string instrument"

00:06:21,001 --> 00:06:22,000 "Musical instrument"--"Guitar"--"Plucked string instrument"

00:06:22,001 --> 00:06:23,000 "Musical instrument"--"Guitar"--"Plucked string instrument"

00:06:23,001 --> 00:06:24,000 "Musical instrument"--"Techno"--"Guitar"

00:06:24,001 --> 00:06:25,000 "Musical instrument"--"Radio"--"Electronic music"

00:06:25,001 --> 00:06:26,000 "Musical instrument"--"Electronic music"--"Synthesizer"

00:06:26,001 --> 00:06:27,000 "Musical instrument"--"Synthesizer"--"Electronic music"

00:06:27,001 --> 00:06:28,000 "Musical instrument"--"Synthesizer"--"Guitar"

00:06:28,001 --> 00:06:29,000 "Musical instrument"--"Sampler"--"Synthesizer"

00:06:29,001 --> 00:06:30,000 "Echo"--"Male singing"--"Radio"

00:06:30,001 --> 00:06:31,000 "Mantra"--"Echo"--"Male singing"

00:06:31,001 --> 00:06:32,000 "Mantra"--"Male singing"--"Echo"--"Middle Eastern music"

00:06:32,001 --> 00:06:33,000 "Mantra"--"Male singing"--"Pigeon

E7

0:9:26.04	0:9:26.60	a man is cooking food
0:9:26.64	0:9:52.48	a woman is talking about a woman s hair
0:9:52.52	0:9:56.96	a person is folding a paper
0:9:57.00	0:10:1.12	a man is cutting a piece of paper
0:10:1.15	0:10:3.59	a cartoon character is trying to get a fish from the ground
0:10:3.64	0:10:6.03	a woman is putting a baby in her mouth
0:10:6.08	0:10:10.00	a person is making a craft
0:10:10.04	0:10:21.24	a woman is talking about a movie
0:10:21.28	0:10:23.76	a woman is making a cake with a glass
0:10:23.80	0:10:26.96	a person is opening a package
0:10:27.00	0:10:27.76	a man is making a drink from a glass
0:10:27.80	0:10:31.92	a man is talking about a food item
0:10:31.96	0:10:33.00	a man is talking about a project
0:10:33.04	0:10:34.92	a woman is talking to a man
0:10:34.96	0:10:40.24	a man is working on a piece of wood
0:10:40.28	0:11:35.56	a child is playing with dolls
0:11:35.60	0:11:39.92	a man is talking about a car
0:11:39.96	0:11:42.16	a woman is holding a baby
0:11:42.20	0:11:45.28	a man is showing how to make a model
0:11:45.32	0:11:47.16	a man is talking to a woman
0:11:47.20	0:11:52.12	a man is putting a toy in a toy
0:11:52.16	0:11:54.76	a woman is putting a toy in a container
0:11:54.80	0:11:56.92	a woman is putting some candy in a bowl
0:11:56.96	0:11:59.76	a girl is eating a meal
0:11:59.80	0:12:2.84	a woman is talking about a product
0:12:2.88	0:12:8.15	a man is putting a condom on a bottle
0:12:8.20	0:12:9.00	a man is talking about the birds
0:12:9.03	0:12:13.64	a person is doing a cooking show
0:12:13.68	0:12:15.80	a woman is looking at a man
0:12:15.84	0:12:25.40	a person is showing how to make a dish
0:12:25.44	0:12:29.24	a woman is showing how to make a flower
0:12:29.28	0:12:36.96	a woman is talking about her doll
0:12:37.00	0:12:40.44	a woman is talking about her phone
0:12:40.48	0:12:43.28	a man is holding a cup of food
0:12:43.32	0:12:46.20	a woman is talking about her food
0:12:46.24	0:12:56.60	a man is putting a cup of food
0:12:56.64	0:12:58.56	a woman is talking about her hair
0:12:58.60	0:13:0.55	a man is putting a piece of food
0:13:0.59	0:13:2.71	a man is dancing
0:13:2.76	0:13:4.32	a man is surfing
0:13:4.35	0:13:6.35	a woman is talking to a man
0:13:6.40	0:13:9.67	a man is singing a song
0:13:9.71	0:13:13.32	a man is talking about a video game
0:13:13.36	0:13:34.80	a woman is talking to a woman in a kitchen
0:13:34.84	0:13:35.24	a woman is dancing in a music video

E8

Appendix 2: Questionnaire form

(Translated from the Finnish original version.)

Welcome! Thank you for the possibility to interview you!

Yle is participating in an international EU funded project called "MeMAD" together with universities and other companies. The project aims at finding out how automatic content analysis / metadata could help to improve the work methods in the media business.

There are no right or wrong answers. The results are used anonymously in the MeMAD project.

Please talk aloud your thoughts. First reactions and feelings are important!

May I record this interview?

1. Background questions

- What does your work consist of? Tell in your own words what you do.
- What is your workplace?
- How long have you worked in media?

2. Handing out the data examples to the interviewee

- What do you see here?

3. Searching and browsing

- In which work related tasks do you search for or browse media or metadata?
- Which parts of the example data would be useful for searching or browsing? In which work related tasks would these be useful?
- From the data you mentioned, which would be the most useful?
- Looking at the data examples, what types of data are missing? For what work related tasks would you need those?
- From the data you mentioned, which would be the most important?
- How would you grade the overall usefulness of shown data examples from the viewpoint of searching or browsing? (1 to 5, 1 = useless, 5 = very useful)
- Data example specific grades for browsing and searching.
- (Additional comments, if any)

4. Creating data

- Does your work involve creating data or content descriptions of video and audio? (yes / no)
- In which work related tasks do you input data or content descriptions of video or audio?
- Which parts of the example data would be useful when inputting data and content descriptions?
- From the data you mentioned, which would be the most useful?

- Looking at the data examples, what types of data are missing? For what data inputting situations would that be useful for?
- From the data you mentioned, which would be the most useful?
- How would you grade the overall usefulness of shown sample data for creating data and content descriptions? (1 to 5, 1 = useless, 5 = very useful)
- Data example specific grades for data creation.
- (Additional comments, if any)

5. Service Development

- Does your work involve developing or managing online services or applications? (yes / no)
- If yes, what parts of the data examples could be used for improving a service?
- What data is missing, that would be useful for improving a service?
- How should this example data be improved to make it more useful for the service you develop?
- How would you grade the overall usefulness of the example data for service development? (1 to 5, 1 = useless, 5 = very useful)
- Data example specific grades for service development.
- (Additional comments, if any)

6. Subtitles, translations and accessibility

- Does your work involve subtitles, translations or accessibility? (yes / no)
- Which of the data examples would be useful in your work?
- From the data you mentioned, which would be the most useful?
- From the perspective of your work, what useful data is missing from the examples?
- From the data you mentioned, which would be the most important?
- How would you grade the overall usefulness of the example data for subtitling, translating and accessibility? (1 to 5, 1 = useless, 5 = very useful)
- Data example specific grades for subtitles, translating, accessibility
- (Additional comments, if any)

7. Other viewpoints

- Explore the example dataset. How should good data look like?
- What other uses would you find for the example dataset? Should the data be somehow modified for these uses?

8. User interface

- (The MeMAD live prototype is first demonstrated shortly to the interviewee.)
- First reaction?
- Would something like this be useful in your work? (1 to 5, 1 = useless, 5 = very useful)
- Other comments regarding the user interface?
- How would you like to use the data examples presented in printed form on your computer or other device?

9. Other

- Did you forget to say something / free comment for example about this interview?

Thank you for the interview!

Appendix 3: Use case ideas presented during the interviews

- finding the correct segment in a video
- finding a specific program if the name of the program is not remembered
- why does a specific program or live broadcast gain viewers and others not / a richer set of data for viewer analytics purposes
- analysing how many percentages of the program had a certain topic/theme
- processing interview raw video material
- automatic keyword identification for a TV program
- translating the content
- literating podcasts
- finding sound or video effects
- automatically identifying the theme, genre, feeling, mood (e.g. “anger”, “power”)
- identifying the music based on instruments used (music genre too subjective)
- archive material finding
- music and archive material reporting
- automatic marketing and search engine optimisation, recommender systems
- internal navigation of a program, automatic chaptering of a program
- accessing the data via an API
- booking translators for the required languages in a program
- accessibility of the programs - vision impaired, hear impaired
- anything with a timecode
- language identification
- who are the persons present in the program
- what are the main themes / topics in the transcript
- translation would benefit from having both the transcript and the visual description
- interlinking different data elements and different levels of abstraction / abbreviation
- when does the person speaking change (speaker segmentation)
- interpretation of the content: “male3 was talking much but saying little”
- finding the right spot in a four hour recording of a legal proceedings / trial
- finding the old version of the same program / same video segment inside other program
- how to measure the impact of the content?
- how to analyse the dramatic arch of the program? which segments work and why?
- automatic video clip generator that publishes fun stuff on the internet
- automatic identification of the music recording
- automatic identification of the person
- automatically identified keywords interlinked to public linked data sources
- automatic music usage reporting
- automatic speech recognition for songs - for video editors and other analysis
- assisting the writing of the marketing text for a program (transcripts)
- live transcript next to a program while watching a program on Netflix
- live transcript inside the studio, e.g. in the prompter
- automatic quote from a video - for e.g. an online article
- what are the narrative elements of the video (e.g. “montage with music”)

- automatic studio prop - if the presenter says “cheese”, the background would be filled with images of cheeses
- automatic video mixing: cutting into the currently speaking person to the outgoing image
- studio automation based on automatic keyword identifier - if the presented says a specific keyword, for example the picture in picture graphics switches
- increase the uniformity of archiving metadata
- the data should be integrated to (Yle) systems
- copy+paste functionalities to the data - so that it can be easily transported from one application to another
- quickly checking raw material for specific spoken words or topics discussed (e.g. directly in the video camera or audio recorder)
- Associative search: power ⇒ show the building of the parliamentary
- Find material with a certain milieu / what is the milieu of this material?