



**FP6- 027685 MESH**

## **D8.2**

# **Technical Evaluation Plan**

<b>Contractual Date of Delivery:</b>	M8 (October 2006)
<b>Actual Date of Delivery:</b>	December 15, 2006
<b>Workpackage:</b>	<i>WP8 Evaluation and Assessment</i>
<b>Dissemination Level:</b>	Confidential
<b>Nature:</b>	Report
<b>Approval Status:</b>	Final
<b>Version:</b>	Final
<b>Total Number of Pages:</b>	26
<b>Distribution List:</b>	WP8, TMC members, European Commission
<b>Filename:</b>	mesh-wp8-D8.2-20060831-TechEvalPlan.doc
<b>Keyword list:</b>	Performance, Evaluation

### **Abstract**

**This report describes the approach and plans for the evaluation of MESH technology.**

The information in this document reflects only the author's views and the European Community is not liable for any use that may be made of the information contained therein. The information in this document is provided as is and no guarantee or warranty is given that the information is fit for any particular purpose. The user thereof uses the information at its sole risk and liability.

## History

Version	Date	Reason	Revised by
0.1	29-10-2006	Document creation	Franciska de Jong
0.2	25-11-2006	Integration of input requested from WPLEaders	Franciska de Jong
final	15-12-2006	Completion after feed back	Franciska de Jong

## Author list

UT	Franciska de Jong, and contributors from all parties
----	--

## Executive Summary

This report gives an overview of the plans for the performance evaluation of a selected set of MESH components.

## Table of Contents

1. Introduction.....	5
2. Plans for component evaluation. ....	6
2.1. Multimodal Visual Classification .....	6
2.2. Automatic Speech Recognition.....	8
2.3. Textual Information Extraction .....	10
2.4. Natural Language Temporal Reasoner .....	12
2.5. Multimedia Summary Generator.....	14
2.6. Content Retrieval Module.....	16
2.7. Ranking Module .....	18
2.8. Scalable Visual Analysis Module & Annotation Module .....	20
2.9. Recommender / Filtering module .....	22
3. Evaluation Time Line.....	25
Appendix - List of terms and abbreviations .....	26



## 1. Introduction

In this document we will describe the plans for the evaluation of MESH technology. Within MESH we distinguish between three types of technology evaluation:

1. Evaluation of the usability of technology.
2. Evaluation at system level.
3. Evaluation at system component level

Type 1 evaluation will be set up in collaboration with several types of users in a well defined setting. This type of evaluation is not addressed in WP8, and as a consequence also not in this document; it will be considered in WP6, and defined in Milestone MS6.5 (Test & validation plans).

Type 2 evaluation will address the question whether the architectural design and the integration work that is the core of the work in MESH Work Package 6 is adequate in the sense that it allows the foreseen processing of content and whether it supports the envisaged content consumption along the lines of the scenario's selected. This assessment will have to demonstrate the robustness, time efficiency, feasibility and scalability of the MESH approach, according to software engineering best practices. Procedures will be defined not as part of WP8, but as part of the work in WP6.4 (System Integration) and WP6.5. (Testing, validation & training). Those will be elaborated by Milestones MS6.4 (Integration Guidelines) and MS6.5 (Test & validation plans).

Type 3 evaluation is mainly motivated by the ambition of MESH to demonstrate performance quality from a scientific point of view. MESH plans to deliver a complex modular system. For measuring the quality of the architecture and the integration effort, standard software design and engineering principles are the major drivers for success. For some parts of the functionality the MESH consortium aims to integrate state-of-the art technology. It is assumed that high quality components will improve the overall quality of the system, and that monitoring component quality could hand in instruments for system quality monitoring and system optimization. Partly due the experimental nature of the MESH system, not all newly developed components are standardized and mature enough to allow an assessment according to standard scientific performance measures. Instead we have made a selection of components for which both a relatively stable technological basis exists, and for which proper data sets and methodology for setting up a performance evaluation is either readily available or can be developed within MESH.

The major focus in this report is on the plans for performance evaluation of a number of individual components. Information has been gathered for these components in the form of template structure consisting of 9 fields. (The field templates will be presented in section 2. Section 3 will explain the time line for the evaluation planned. This document concludes with a list of abbreviations and pointers to explanations for some technical terms and abbreviations (e.g. recall, precision, ROC-curve, etc.)

## 2. Plans for component evaluation

### 2.1. *Multimodal Visual Classification*

Responsible partner: UAM, Javier Molina, Jose M. Martínez

MESH relevant activity: S2.2.3

#### 2.1.1. Name/ID

Multimodal Visual based Classification

#### 2.1.2. Functionality

The target of this task is to perform video contents classification using, mainly, visual descriptors. The classifying results will be at a shot level, this is, which classes have been found in a certain shot.

After a shot temporal segmentation of the video, we will take some of the frames (keyframe extraction), for carrying out a frame level analysis. For performing this, apart from the features extracted for this aim, we will have access to extracted parameters from other subtasks (S2.2.1 & S2.2.2). A statistical treatment should be given to these extracted characteristics, combining them in order to manage the highest accuracy percent. The parameters which first aim is to help in the classification are of the following natures:

- Identify segments, textures, colours and other low-mid level features.

Basing on a spatial segmentation constructed over these low-level descriptors, we will look for visual descriptors able to discriminate between different issues in an image. These are, presence and location in a frame of: flames, water, vegetation, rocks, persons...

With these inferred concepts we will feed S2.4.2 (Reasoning for Semantic Multimedia)

#### 2.1.3. Evaluation Objective

The evaluation of the results of S2.2.3 should try to perform a quantitative estimation of the extracted visual cues quality. This could be managed at frame level, or at shot level.

#### 2.1.4. Evaluation Steps

Focusing on the evaluation of frame analysis results:

- Spatial Segmentation
- Visual descriptors discrimination capacity.

- Focusing on the evaluation of shot analysis results:
- Measure of the improvements of the use of motion descriptors.

#### **2.1.5. Evaluation implications & Limitations**

It will be difficult to develop a spatial segmentation ground truth.

#### **2.1.6. Data sets**

MESH news MPEG-2 videos.

#### **2.1.7. Measures & Methodology**

- Visual descriptors discrimination. We will first divide the images into spatial segments, manually. We will reunite a significant number of images (cropped of the original ones) of the objective classes (water, flames, rocks, vegetation...). Then we will perform a study of with descriptors and with which quantification they should be used with.
- Measure of spatial segmentation accuracy. Still to be defined.

#### **2.1.8. Evaluation Outcome**

Results will be reported as technical reports (made available via MESH website) or as papers.

#### **2.1.9. Time line**

Evaluation start: M18

## **2.2. Automatic Speech Recognition**

Responsible partner: UT, Roeland Ordelman

MESH relevant activity: 2.3.1

### **2.2.1. MSH-ASR-SP, MSH-ASR-EN, MSH-ASR-GE**

The MESH automatic speech recognition (ASR) modules for English (MSH-ASR-EN), Spanish (MSH-ASR-SP) and German (MSH-ASR-GE) are part of Deliverable D2.2 (MESH Content Analysis Modules) that integrates a 1<sup>st</sup> version of the speech recognition systems (M12).

### **2.2.2. Functionality**

The speech recognition modules produce full text transcripts with word timing-information of those parts in broadcast news shows that contain speech. It is assumed that the modules receive audio segmentation information (speech, music, silence, other).

### **2.2.3. Evaluation Objective**

The evaluation aims at finding the optimal speech recognition parameter settings given the task domain.

### **2.2.4. Evaluation Steps**

The following evaluation steps are distinguished for each of the target languages:

- Phase-0: global baseline evaluation on test set of training data used for initial system training (system check)

In phase-0 all speech recognition modules are initialized using existing acoustic models and language models (English) or existing training resources (Spanish, German) for the broadcast news domain. The systems are evaluated using a small language specific broadcast news evaluation set.

- Phase-1: broadcast news (benchmark) evaluations

The phase-1 evaluations are carried out using the systems initialized in phase-0, but now the evaluations are based on either broadcast news benchmark data (HUB-4) or MESH specific data. The results are indicators of the performance of systems without domain specific tuning and serve as the starting point (baseline) for comparison of tuned systems in successive development steps.

- Phase-2: progress evaluations based on the implementation of new algorithms or new models.

In Phase-2 specific methods for speech recognition in the MESH/broadcast news domain are developed. The evaluations in phase-2 are indicators of the progress

that is made in tuning the baseline speech recognition modules to the MESH/broadcast news domain.

- Phase-3: final evaluation on hitherto unseen MESH data

### **2.2.5. Evaluation implications & Limitations**

Evaluation results on benchmark data are used to relate the performance of the speech recognition modules to other broadcast news speech recognition systems. The results on MESH data give an indication of the quality of the benchmarked systems on MESH specific data.

### **2.2.6. Data sets**

The following external data sets are/ will be used for evaluation:

- English HUB-4 benchmark evaluation corpus
- Spanish HUB-4 benchmark evaluation corpus
- German: to be defined

### **2.2.7. Measures & Methodology**

For the assessment of the speech recognition systems the standard evaluation metric, the word error rate (WER) is used. The word error rate is based upon a comparison of a reference transcription of the test material with the output of the recognizer referred to as the hypothesis transcription. The scoring algorithm searches for the minimum edit distance in words between the hypothesis and the reference and produces the number of substitutions, insertions and deletions that are needed to align the hypothesis with the reference .

The WER alone may however not be a good indicator for the quality of the speech transcript in the context of the requirements for MESH. When the speech recognition system recognised the speech fairly accurately according to the WER metric but in fact missed important names (which are important keywords in search), the system performed badly. Therefore, metrics such as out-of-vocabulary (OOV) rate (e.g., could a name be recognised at all?) and term error rate (TER) will be reported as a complement to the overall word error rate.

### **2.2.8. Evaluation Outcome**

Results will be reported as technical reports (made available via MESH website) or as papers.

### **2.2.9. Time line**

M8 end of phase-0 for English

M10 end of phase-0 for Spanish

M12 end of phase-1 for English, Spanish; end of phase-0 German

## **2.3. Textual Information Extraction**

Responsible partner: DFKI, Thierry Declerck

MESH relevant activity: S2.3.2 (and may be also 2.3.3)

### **2.3.1.Name/ID**

Textual Information Extraction/Part of D2.2

### **2.3.2.Functionality**

Information extraction from text documents and textual parts of multimedia content, possibly including ASR transcripts

### **2.3.3.Evaluation Objective**

Evaluation of coverage and correctness of handling distinct types of information contained in texts accompanying or related to media files, with a special stress on temporal expressions.

### **2.3.4.Evaluation Steps**

We will first need to constitute a corpus of manually annotated data. The annotation schema will have to be defined very soon in cooperation with the relevant partners. Then we run the tools on the same data and compare automatically the annotations generated by the tools and the manually annotated data. ON the base of the first results, we will improve the tools, and start the same procedure with a new set of manually annotated data.

### **2.3.5.Evaluation implications & Limitations**

Following the evaluation metrics described below, we will provide for a clear picture on the accurateness of the textual extraction tools, applied to distinct types of information. On the base of a threshold (to be fixed) we can then decided for which kind of information the tools are working well enough to be included in the automated part of the semantic annotation of the MESH data.

### **2.3.6.Data sets**

We will use for sure all available MESH data (for the time being the data from Deutsche Welle) In close relation with NoE K-Space and SmartWeb project, we can also use integrated data (image/text) on soccer, and some data from INA on news (in French, which is outside the set of languages to be covered by MESH, but this can serve as a proof of concept). This would be done in cooperation with NoE K-Space

### **2.3.7. Measures & Methodology**

We will use the metrics of recall and precision, as usual in Information Extraction. This means that we will also need some manually annotated data for comparing the results of the extraction tools to the so-called gold data represented by the manual annotation.

### **2.3.8. Evaluation Outcome**

Results will be reported as technical reports (made available via MESH website) or as papers.

### **2.3.9. Time line**

Evaluation start: M14.

## **2.4. Natural Language Temporal Reasoner**

Responsible partner: DFKI, Walter Kasper

MESH relevant activity: S3.1.2/S3.3.2

### **2.4.1. Name/ID**

Mesh\_TR/Part of D3.4

### **2.4.2. Functionality**

Special purpose reasoner for temporal information:

- resolution and normalization of temporal expressions
- event ordering

### **2.4.3. Evaluation Objective**

Evaluation of coverage and correctness of handling temporal expressions.

### **2.4.4. Evaluation Steps**

Informal evaluation will be part of the development process. A first formal evaluation will be done after release.

### **2.4.5. Evaluation implications & Limitations**

The formal evaluation of the temporal reasoner indicates missing coverage as well as incorrect or incomplete implementation of the module. It does not evaluate performance or scalability.

### **2.4.6. Data sets**

The TimeBank corpus<sup>1</sup> provides a large set of annotated data for temporal expressions and phenomena. It needs to be checked to what extent it can be used for evaluating the MESH temporal reasoner by identifying annotations that are reasoning results.

Additionally, smaller sets of annotated MESH data will be created.

### **2.4.7. Measures & Methodology**

Evaluation will be done against manual annotated data. Standard measures like precision, recall, F-measure, etc. will be used.

---

<sup>1</sup> Available at <http://www.cs.brandeis.edu/~jamesp/arda/time/timebank.html>

#### **2.4.8. Evaluation Outcome**

Results will be reported as technical reports (made available via MESH website)

#### **2.4.9. Time line**

Start: M18

## **2.5. Multimedia Summary Generator**

Responsible partner: TID, Susana Pérez

MESH relevant activity: S4.2.2 & S4.2.3

### **2.5.1.Name/ID**

Multimedia Summary Generator

### **2.5.2.Functionality**

This component will generate an abridged audiovisual segment that will contain the semantically significant fragments of a news video item, edited so that it appears as a coherent entity, and adapted to a user. Additionally, it will deliver a description of the generated summary which includes the timing relationship between the fragments used and their position in the original content piece.

### **2.5.3.Evaluation Objective**

Assessment that the generated summaries:

- a) are internally coherent and understandable, i.e. the rendering of the summary produces a viewable video.
- b) are semantically meaningful, i.e. really correspond to the (subjective) important pieces of the news item
- c) For personalized summaries, that they have been adapted to the user's profile

### **2.5.4.Evaluation Steps**

Baseline evaluation: corresponds to objective (a): that the summaries, at the very minimum, are usable pieces.

Intermediate & Final evaluation: objectives (b) and (c), to assess that the summaries do capture the essence of the original piece and that, if they are personalized, that they depend on the users' profiles.

### **2.5.5.Evaluation implications & Limitations**

For a summary to be actually valuable, it needs to be significantly shorter than the original piece; therefore the outcome of this module can only be really evaluated when generating summaries from news item of a certain minimum length (e.g. a summary from a 30-second item will not probably mean much).

### **2.5.6.Data sets**

- MESH video test set, with preference for the longest videos.
- It will be investigated whether the TRECVID collections can be used.

In both cases, always taking into account that those videos need to have been processed by WP2 to be able to be used by the module, since we need the extracted semantic information.

### **2.5.7. Measures & Methodology**

Pre-evaluation: automatic video decoding of the summaries to check for stream errors.

Evaluation: subjective assessment of the quality of the summary, in absolute terms (objective a), relative to the original news item (objective b) and relative to the user profile (objective c)

### **2.5.8. Evaluation Outcome**

1. A preliminary report will be included in D4.6 (M16); more data will be made available at a later stage.
2. A database of the generated summaries will be made available online, so that they can be inspected and compared with the original sources.

### **2.5.9. Time line**

S4.2.2 & S4.2.3 start in M6 and go on to M18. Relevant stages are:

- MS4.7 – Off-line semantic summarisation (M15)
- D4.6: Semantic summarisation and syndication, 1st version (M16)

The pre-evaluation will likely take place prior to the delivery of the module in MS4.7, during M12-15, while the evaluation itself will take place M14-M18, with some results included in D4.6 (M16)

## **2.6. Content Retrieval Module**

Responsible partner: Noterik

MESH relevant activity: T4.4.3

### **2.6.1.Name/ID**

Retrieval Module

### **2.6.2.Functionality**

This module is responsible for two main functionalities:

- Semantic retrieval (led by UAM)
  - Run the formal queries against the ontologies and retrieve the semantic entities that fulfil the user request
  - Use the annotations made by WP2 to obtain the media pieces of content where those semantic entities appear.
- Visual retrieval (led by Noterik)
  - Use the visual query of the user (picture, video, ...) to obtain the media pieces of similar content relevant to the user.

### **2.6.3.Evaluation Objective**

(semantic retrieval perspective)

- Evaluate the performance of semantic retrieval with different levels of semantic incompleteness on the domain ontologies and knowledge bases.
- Evaluate the performance of the semantic query considering not just text but also multimedia pieces of content.
- Evaluate the dependence of this retrieval techniques with respect to the quality of semantic annotations

### **2.6.4.Evaluation Steps**

- Obtain a set of formal queries with different kinds of semantic information coverage as result
- Annotate the different kinds of corpus with different annotation techniques
- Perform the retrieval with the different queries and different annotated corpora
- Validate the relevance of results manually by users (ground truth)
- Identity the quality of results depending on the kind of content and the level of knowledge incompleteness in the ontologies and knowledge bases
- Compare the quality of results between the traditional keyword-based search, and traditional techniques of multimedia search

### **2.6.5. Evaluation implications & Limitations**

This methodology can make generalizations about the capabilities of the new techniques of semantic search to overcome the traditional information retrieval techniques

### **2.6.6. Data sets**

It will be investigated whether TREC collections can be used to compare the quality of results with formal datasets

### **2.6.7. Measures & Methodology**

There are several measures to evaluate the quality of retrieval, such as precision and recall.

### **2.6.8. Evaluation Outcome**

Results will be reported as technical reports (made available via MESH website)

### **2.6.9. Time Line**

Start: M18

## **2.7. Ranking Module**

Responsible partner: UAM

MESH relevant activity: T4.4.2

### **2.7.1.Name/ID**

Ranking Module

### **2.7.2.Functionality**

The main functionality of this module is to rank the retrieved content following different criteria

### **2.7.3.Evaluation Objective**

- Evaluate and compare the performance of different ranking algorithms
- Evaluate which criteria are better in which situations to improve the content ranking
- Evaluate the ranking algorithms for keyword based search and for multimedia search

### **2.7.4.Evaluation Steps**

- Obtain different user cases or user information needs
- Decide a set criteria to be considered for ranking
- Validate the relevance of results manually by users
- Identity the quality of results depending on the used criteria and the score returned for each specific criterion during the ranking process
- Compare the quality of results against other ranking techniques used for keyword-based ranking and for multimedia ranking

### **2.7.5.Evaluation implications & Limitations**

n.a.

### **2.7.6.Data sets**

It will be investigated whether TREC collections can be used to compare the quality of results with formal datasets

### **2.7.7.Measures & Methodology**

Standard measures to evaluate the quality of ranking, such as precision and recall

### **2.7.8. Evaluation Outcome**

Results will be reported as technical reports (made available via MESH website)

### **2.7.9. Time line**

Start: M18

## **2.8. Scalable Visual Analysis Module & Annotation Module**

Responsible partner: ITI Athens, Natasa Sofou

MESH relevant activity: S5.3.1, S5.3.2

### **2.8.1.Name/ID**

Component name: Scalable Visual Analysis & Annotation Module

Deliverable id: D5.3

### **2.8.2.Functionality**

According to the functionality, the component can be divided in two subcomponents:

*Subcomponent: scalable visual analysis*

The main functionality of this component will be to adapt a number of image and video processing analysis tasks to less constrained environments, in order to provide close to real-time operations. Main function: temporal segmentation of videos to shots;

*Subcomponent: annotation*

- Clustering of images according to predefined classes.
- The component will also contain a user interface module that will show the classification labels assigned to the shots/keyframes, and enable correction/enhancements of the annotations

### **2.8.3.Evaluation Objective**

The objective here is the evaluation of shot/scene, keyframe detection results and classification of images to predefined (user defined) classes. Additionally, the proposed algorithms will be evaluated in terms of complexity and efficacy.

### **2.8.4.Evaluation Steps**

- Selection of an initial set of test-videos
- Extraction of a ground-truth data:
  - o Scenes
  - o Keyframes
  - o Image classes
- Comparison of the obtained results against the ground-truth values.
- Algorithmic evaluation in terms of complexity, efficacy and processing capabilities.

### **2.8.5. Evaluation implications & Limitations**

n.a.

### **2.8.6. Data sets**

MESH-internal data sets.

### **2.8.7. Measures & Methodology**

The accuracy of the shot, key frame detection and image classification procedures will be evaluated and measured quantitatively through the success and failure rates.

### **2.8.8. Evaluation Outcome**

Results will be reported as technical reports (made available via MESH website)

### **2.8.9. Time line**

Evaluation timeline: Start: M16, End: M20

## **2.9. Recommender / Filtering module**

Responsible partner: UAM, Miriam Fernández

MESH relevant activity: T5.4, S5.4.3 (User profile exploitation)

### **2.9.1.Name/ID**

Recommender / Filtering module

### **2.9.2.Functionality**

This module receives (from WP4) a list of resources identifiers with associated semantic information. As response, it returns a ranked list of resources with an associate rating to each resource that measures its relevance for a given user according to his profile.

The ranking value will be obtained from a combination of different filtering criteria:

- Content-based filtering. Rates an item combining in a smart way long term user interests, stored in the semantic user profile component, and short term interests, extracted from the recent user's activity within the system.
- Collaborative filtering. Rates an item according to the explicit ratings given by users with similar interests.
- Social filtering. Rates an item taking into account the interests of people explicitly linked with the user.

### **2.9.3.Evaluation Objective**

Evaluate the different types of recommendations (content-based, collaborative and social) separately, and possible combinations of some or all of them.

### **2.9.4.Evaluation Steps**

*Evaluation of the content-based filtering module:* measure of the influence of semantic preference spreading strategies through the ontology domain and user profiles, comparison with keyword-based recommendation algorithms, and comparison of personalization enhancements (context-based, etc.) with basic personalization.

*Evaluation of the collaborative filtering module:* exploration of traditional collaborative filtering algorithms, and novel strategies to infer semantic content-based preferences from existent item ratings.

*Evaluation of the social filtering module:* evaluation of social collaborative recommendations.

*Evaluation of hybrid filtering approaches:* examination of novel strategies to combine the outputs of the above recommender modules, and assessment of the policies that determine the weights of combination.

In all of the above steps, evaluations according to different proportions of user preferences, items and semantic annotations should be performed.

Furthermore, the influence of the contextual information during the recommendation processes has also to be studied.

### 2.9.5. Evaluation implications & Limitations

Due to the novelty of the MESH semantic content-based approach, it might be difficult to make comparisons with well-known traditional content-based and collaborative recommenders.

Another difficulty is in evaluating context-based technologies. Formal evaluations as are usual in the IR tradition are difficult to do because of the additional variables involved (ser preferences, contextual information, etc.), but perhaps not impossible. It is also difficult to set up a common scenario for several users. Grounds truth tends to be difficult to define, difficult to get from users, and costly. Other options could be based on external (to the IR system) measurements, more oriented towards user satisfaction or the like (side to side evaluation, etc.). A study of the literature is foreseen to seek reference work on evaluation methodologies for this kind of techniques.

### 2.9.6. Data sets

In order to properly compare the future MESH recommendation algorithms with those existent in the Recommender Systems literature, UAM intends to use the well-known MovieLens movie rating repository<sup>2</sup>. In fact this seems to be the most complete available dataset allowing statistically significant personalization-related experiments.

To enrich this database with domain information, UAM will also provide an ontology generated from the Internet Movie Database<sup>3</sup> (IMDb).

Other public collections include the joke rating dataset from the Jester Joke Recommender System<sup>4</sup>, the Book-Crossing (BX) book rating dataset<sup>5</sup>, the Audioscrobbler's music playlist datasets<sup>6</sup>, and the AOL web search dataset<sup>7</sup>.

The corpora and datasets developed in MESH will probably not support large-scale experiments, and will not be a standard benchmark from a scientific research point of view. But it can be used for demonstrational purposes, i.e. it should be possible to demonstrate/illustrate the capabilities and perceivable effects of the personalization techniques with the MESH content and corpus.

---

<sup>2</sup> <http://www.grouplens.org/>

<sup>3</sup> <http://www.imdb.com/>

<sup>4</sup> <http://www.ieor.berkeley.edu/~goldberg/jester-data/>

<sup>5</sup> <http://www.informatik.uni-freiburg.de/%7ecziegler/BX/>

<sup>6</sup> <http://www.audioscrobbler.net/data/webservices/>

<sup>7</sup> <http://www.gregsadetsky.com/aol-data/>

### **2.9.7. Measures & Methodology**

Related to the Recommender System and Information Retrieval literature, some measures that could be evaluated are the following: precision and recall (ROC curves, F-measure), mean squared error (MAE).

### **2.9.8. Evaluation Outcome**

Results will be reported as technical reports (made available via MESH website)

### **2.9.9. Time line**

Start: M18

### 3. Evaluation Time Line

Most of the MESH tasks aiming at component development will yield results only after M12 or later. As a result most of the evaluation plans described above will be taken up and/or completed in the second phase of MESH. A more detailed planning will be described in the workplan for the phase starting after M18.

<b><i>Component name</i></b>	<b><i>start date</i></b>	<b><i>end date</i></b>
Multimodal Visual Classification	M18	to be specified later
Automatic Speech Recognition – phase 1	M8	M12
Textual Information Extraction	M14	to be specified later
Natural Language Temporal Reasoner	M18	to be specified later
Multimedia Summary Generator	M14	M18
Content Retrieval Module	M18	to be specified later
Ranking Module	M18	to be specified later
Scalable Visual Analysis Module & Annotation Module	M16	M20
Recommender / Filtering module	M18	to be specified later

## Appendix - List of terms and abbreviations

HUB4 NIST/DARPA 1998 evaluation of speech recognition technology on broadcast news in English  
[http://www.nist.gov/speech/tests/bnr/hub4\\_98/hub4e\\_98\\_spec.htm](http://www.nist.gov/speech/tests/bnr/hub4_98/hub4e_98_spec.htm)

TREC Text Retrieval Evaluation Conference  
<http://trec.nist.gov/>

TRECVID Video Retrieval Benchmark Conference  
<http://www-nlpir.nist.gov/projects/trecvid/>

ASR Automated Speech Recognition

For an explanation of recall, precision, F-measure:

[http://en.wikipedia.org/wiki/Information\\_retrieval](http://en.wikipedia.org/wiki/Information_retrieval)

ROC Receiver Operating Characteristic

For an explanation of ROC curves:

[http://en.wikipedia.org/wiki/Receiver\\_operating\\_characteristic](http://en.wikipedia.org/wiki/Receiver_operating_characteristic)

WER Word Error Rate