EUMSSI: a Platform for Multimodal Analysis and Recommendation using UIMA

Jens Grivolla
Universitat Pompeu Fabra
Barcelona, Spain
jens.grivolla@upf.edu

Maite Melero
Universitat Pompeu Fabra
Barcelona, Spain
maite.melero@upf.edu

Toni Badia
Universitat Pompeu Fabra
Barcelona, Spain
toni.badia@upf.edu

Cosmin Cabulea
Deutsche Welle
Bonn, Germany
cosmin.cabulea@dw.de

Yannick Estève
Université du Maine
Le Mans, France
yannick.esteve@lium.univ-lemans.fr

Eelco Herder
L3S Research Center
Hannover, Germany
herder@l3s.de

Jean-Marc Odobez
IDIAP Research Institute
Martigny, Switzerland
odobez@idiap.ch

Susanne Preuß
Gesellschaft zur Förderung der Angewandten Informationsforschung
Saarbrücken, Germany
susannep@iai.uni-sb.de

Raúl Marín
VSN Innovation and Media Solutions
Alicante, Spain
rmarin@vsn.es

Abstract

The EUMSSI project (Event Understanding through Multimodal Social Stream Interpretation) aims at developing technologies for aggregating data presented as unstructured information in sources of very different nature. The multimodal analytics will help organize, classify and cluster cross-media streams, by enriching its associated metadata in an interactive manner, so that the data resulting from analysing one media helps reinforce the aggregation of information from other media, in a cross-modal semantic representation framework. Once all the available descriptive information has been collected, an interpretation component will dynamically reason over the semantic representation in order to derive implicit knowledge. Finally the enriched information will be fed to a hybrid recommendation system, which will be at the basis of two well-motivated use-cases. In this paper we give a brief overview of EUMSSI's main goals and how we are approaching its implementation using UIMA to integrate and combine various layers of annotations coming from different sources.

1 Introduction

Nowadays, a multimedia journalist has access to a vast amount of data from a plurality of types of sources to document a story. In order to put information into context and tell his story from all significant angles, he needs to go through an enormous amount of records with information of very diverse degrees of granularity. At the same time, he needs to reduce the noise of irrelevant content. This is extremely time-consuming, especially when a topic or event is interconnected with multiple entities from different domains. At a different level, many TV viewers are getting used to navigating with their tablets or iPads while watching the TV, the tablet effectively functioning as a second screen, often providing background information on the program or interaction in social networks about what is being watched. Both the

Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT, pages 101–109,
Dublin, Ireland, August 23rd 2014.

journalist and the TV viewer would greatly benefit from a system capable of automatically analysing and interpreting unstructured multimedia data stream and its social background, and, with this understanding, be able of contextualising the data, and contributing with new, related information.

The FP7-ICT-2013-10 STREP project EUMSSI, which started in December 2013, is developing methodologies and techniques for identifying and aggregating data presented as unstructured information in sources of very different nature (video, image, audio, speech, text and social context), including both online (e.g., YouTube) and traditional media (e.g. audiovisual repositories), and for dealing with information of very different degrees of granularity.

This will be accomplished thanks to the integration in a UIMA-based[1] multimodal platform of state-of-the-art information extraction and analysis techniques from the different fields involved (image, audio, text and social media analysis). The multimodal interpretation platform, in an optimized process chain, will analyze a vast amount of multimedia content, aggregate all the resulting information and semantically enrich it with additional metadata layers. The resulting system will be potentially useful for any application in need of cross-media data analysis and interpretation, such as intelligent content management, recommendation, real time event tracking, content filtering, etc. In particular, the EUMSSI project will use the semantically enriched information to make personalized content-based recommendation.

2 Multimodal analytics and Semantic Enrichment

For reasoning with and about the multimedia data, the EUMSSI platform needs to recognize entities, such as actors, places, topics, dates and genres. A core idea is that the process of integrating information coming from different media sources is carried out in an interactive manner, so that the metadata resulting from analyzing one media helps reinforce the aggregation of information from other media. For example, the quality of speech recognition heavily depends on the audio quality and background noise. Existing text, tags and other metadata will be exploited for disambiguation. Further, OCR on video data, speech analysis and speaker recognition mutually reinforce one another. The combined and integrated results of the audio, video and text analysis will significantly enhance the existing metadata, which can be used for retrieval and recommendation. In addition, the extracted entities and other annotations will be exploited for identifying specific video fragments in which a particular person speaks, a new topic begins, or an entity is mentioned. Figure 1 illustrates some of the different layers of analysis that may exist for a video content item.

Once the entities and concepts have been identified in the different modalities, all the information is aggregated and semantically enriched, using general ontologies or structured knowledge bases. Wikipedia categories have been successfully exploited with this purpose in different works: e.g. to describe chemical documents (Köhncke and Balke, 2010), to identify topics of interest for Twitter users (Michelson and Macskassy, 2010), and also to improve Web video categorization (Chen et al., 2010). Moreover, (Hahn et al., 2010) have shown that the structured information gathered from Wikipedia infoboxes can be used to answer complex questions, like "Which Rivers flow into the Rhine and are longer than 50 kilometers?" For this purpose, text documents need to be previously annotated using DBpedia Spotlight (Mendes et al., 2011), which automatically annotates text with links to articles in Wikipedia. The process of semantic enrichment is still largely domain-dependent; therefore, apart from the available general-purpose knowledge bases and ontologies (DBpedia, FOAF, DublinCore...), the EUMSSI platform needs specialized resources for categorizing videos on different dimensions. Linked Data technologies (Heath and Bizer, 2011) and the Linked Open Data cloud[2] provide access to several of these resources, including geodata, movie databases and program information.

3 Content-based Recommendation and the Demonstrators

The semantically enriched information is then used by the EUMSSI system to make personalized content-based recommendation. We propose a novel recommender system that leverages matrix factorization (Koren, 2008) with implicit feedback in order to integrate content-based similarity, usage history

[1] Unstructured Information Management Architecture: http://uima.apache.org/

[2] http://lod-cloud.net/

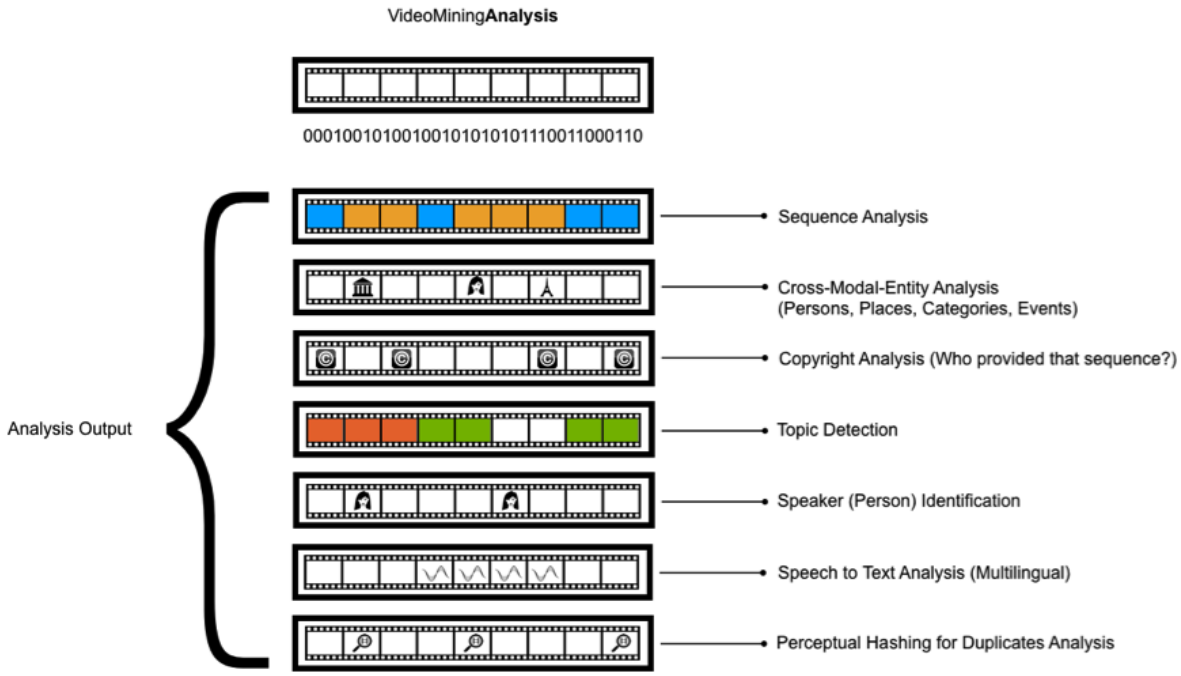

Figure 1: Video Mining Analysis

(i.e. collaborative filtering), as well as user demographics. This integrated approach reduces the cold-start problems typical of collaborative filtering, both for new users and for new content. Recommendation and aggregation of related content in EUMSSI is expected to use varying degrees of personalization, giving more weight in some cases to the individual user's interests, based on his viewing history, but being based primarily on the similarity to the currently shown content in other cases.

On top of the recommender, two demonstrators will be implemented within the EUMSSI project, each catering to a different use-case: (i) a computer-assisted *storytelling* tool integrated in the workflow of a multimedia news editor, empowering the journalist to monitor and gather up-to-date documents related with his investigation, without the need of reviewing an enormous amount of insufficiently annotated records; and (ii) a *second-screen* application for an end-user, able to make relevant suggestions of multimedia content based on what the user is watching, what other people have watched, and what people are saying about these contents in the social networks. Figure 2 shows how both applications build on a common base of multimedia analysis and content aggregation/recommendation algorithms.

4 Architecture overview

All new content coming into the system is first normalized to a common metadata schema (based on schema.org) and stored in a database (MAM/media asset manager, or MongoDB[3]) to make it available for further processing. Analysis results, as well as the original metadata, are stored in CAS format to allow integration of different aligned layers of analysis.

The process flow, pictured in Figure 3, can be summarized as follows:

1. new data arrives (or gets imported)
2. preprocessing stage
   (a) make content available through unique URI (from central MAM)
   (b) create initial CAS with aligned metadata / text content and content URI

[3]it will be developed in parallel as an open source MongoDB based solution, as well as integrated into VSN's proprietary platform

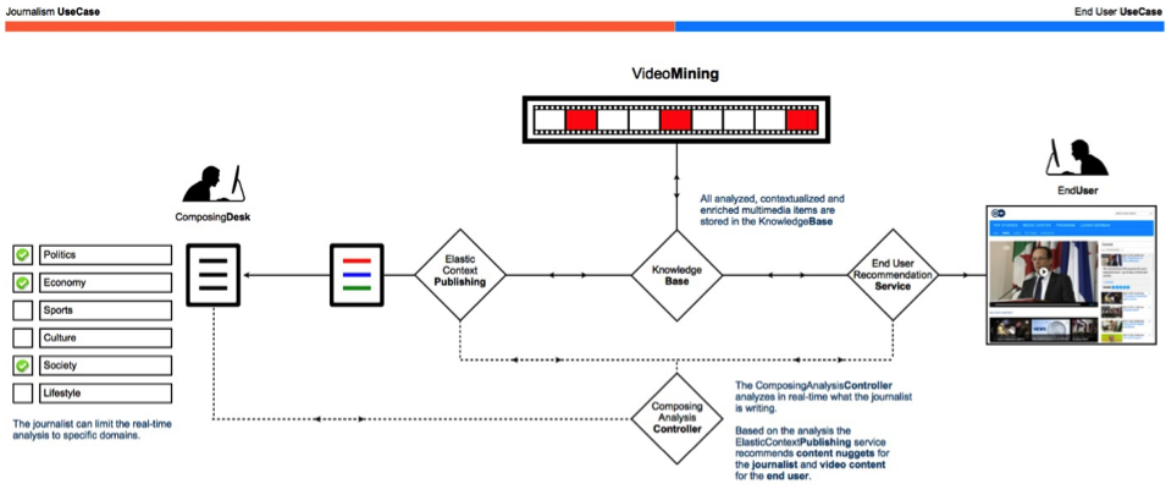

Figure 2: Multimodal platform catering both for the journalist and the end-user's use-cases

(c) add content to processing queues

3. processing / content analysis
   (a) distributed analysis systems query queue when they have processing capacity
   (b) retrieve CAS with existing data (or get relevant metadata from wrapper API)
   (c) retrieve raw content based on content URI
   (d) process
   (e) update CAS (possibly through wrapper API)
   (f) update queues
      i. mark as processed
      ii. add to queues for other processes that depend on previous analysis results
4. indexing when processing is complete for a content item (e.g. with Solr)

Note that this architecture design mainly depicts the data analysis part of the EUMSSI system – the deployment by Web applications is not visible in the figure. These will be built upon the Solr indexes created from the CAS.

5 Aligned data representation

Much of the reasoning and cross-modal integration depends on an aligned view of the different annotation layers, e.g., in order to connect person names detected from OCR with corresponding speakers from the speaker recognition component, or faces detected by the face recognition.

The Apache UIMA[4] CAS (common analysis structure) representation is a good fit for the needs of the EUMSSI project as it has a number of interesting characteristics:

- Annotations are stored "stand-off", meaning that the original content is not modified in any way by adding annotations. Rather, the annotations are entirely separate and reference the original content by offsets

[4]http://uima.apache.org/

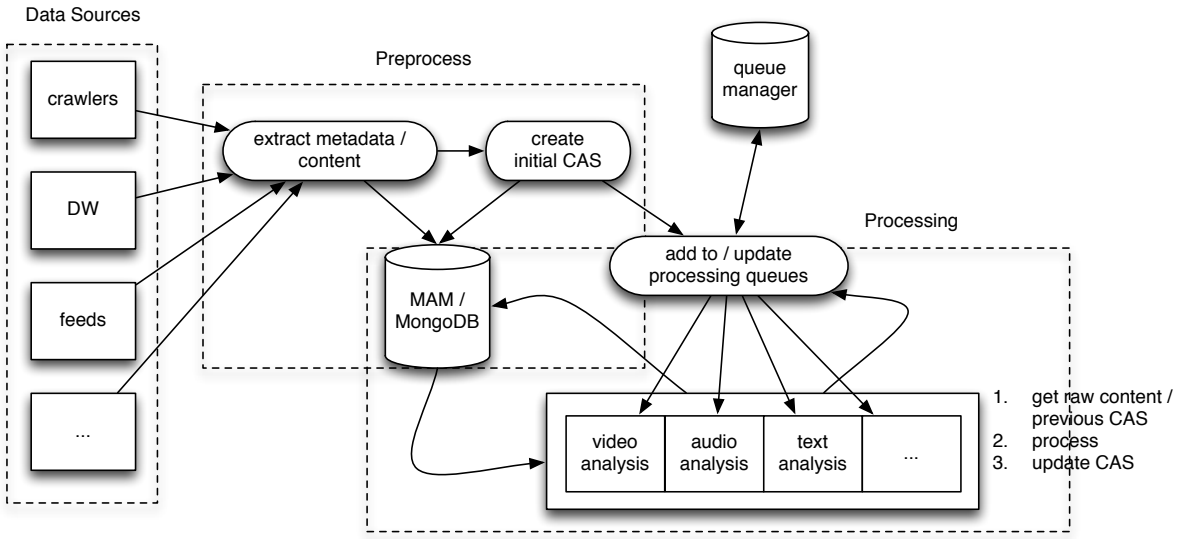

Figure 3: Architecture design

- Annotations can be defined freely by defining a "type system" that specifies the types of annotations (such as *Person, Keyword, Face,* etc.) and the corresponding attributes (e.g. *dbpediaUrl, canonicalRepresentation, ...*)
- Source content can be included in the CAS (particularly for text content) or referenced as external content via URIs (e.g. for multimedia content)
- While each CAS represents one "document" or "content item", it can have several *Views* that represent different aspects of that item, e.g. the video layer, audio layer, metadata layer, transcribed text layer, etc., with separate source content (SofA or "subject of annotation") and separate sets of annotations
- CASes can be passed efficiently in-memory between UIMA analysis engines
- CASes can be serialized in a standardised OASIS format[5] for storage and interchange

In the case of the EUMSSI project, the common base for alignment for different annotation layers referring to multimedia content is timestamps relative to the original content.

Annotations based directly on multimedia content (video and audio) will naturally refer to that content via timestamps, whereas text analysis modules normally work with character offsets relative to the text content. It is therefore fundamental that any textual views created from multimedia content (e.g. via ASR or OCR) refer back to the timestamps in the original content. This will be done by creating annotations, e.g. tokens, that include the original timestamps as attributes in addition to the character offsets.

As an example, we may have a CAS with an audio view on which we apply automatic speech recognition (ASR), providing the transcription as a series of tokens/words with a timestamp for each word. The system then creates a new view in the CAS that has the full plain-text transcription as SofA and a series of *Token* annotations with both character offsets relative to the plain-text SofA, and timestamp offsets relative to the multimedia content.

In this way it is possible to apply standard text analysis modules (that rely on character offsets) on the textual representation, while maintaining the possibility to later map the resulting annotations back onto the temporal scale.

Timestamps will be represented in milliseconds in order to avoid floating point values. In this way, all annotations can be subtypes of the standard UIMA Annotation type[6], which provides access to a number

[5] http://docs.oasis-open.org/uima/v1.0/uima-v1.0.html

[6] otherwise annotations would need to derive from the more generic TOP type

of utility functions that help find sets of overlapping annotations, retrieve annotations in offset order, etc.

SofA-aware UIMA components are able to work on multiple views, whereas "normal" analysis engines only see one specific view that is presented to them. This means that e.g. standard text analysis engines don't need to be aware that they are being applied to an ASR view or an OCR view; they just see a regular text document. SofA-aware components, however, can explicitly work on annotations from different views and can therefore be used to integrate and combine the information coming from different sources or layers, and create new, integrated views with the output from that integration and reasoning process.

6 Flow management

UIMA provides a platform for execution of analysis components (*Analysis Engines* or *AEs*), as well as for managing the flow between those components.

CPE or uimaFIT[7] (Ogren and Bethard, 2009) can be used to design and execute pipelines made up of a sequence of AEs (and potentially some more complex flows), and UIMA-AS[8] (*Asynchronous Scaleout*) permits the distribution of the process among various machines or even a cluster (with the help of UIMA DUCC[9]).

Analysis Engines can either be "natively" written for UIMA or can be wrappers that translate inputs and outputs for existing analysis components so they can be integrated in UIMA. All text analysis components, as well as the integration and reasoning components, will be available as UIMA AEs and can therefore be configured and executed directly within the UIMA environment.

There are some components of the EUMSSI platform, however, that do not integrate easily in this fashion. This is the case of computationally expensive processes that are optimized for batch execution. A UIMA AE needs to expose a *process()* method that operates on a single CAS (= document), and is therefore not compatible with batch processing. This is particularly true for processes that need to be run on a cluster, with significant startup overhead, such as many video and audio analysis tasks.

It is therefore necessary to have an alternative flow mechanism for offline or batch processes, which needs to integrate with the processing performed within the UIMA environment.

The main architectural and integration issues revolve around the data flow, rather than the computation. In fact, the computationally complex and expensive aspects are specific to the individual analysis components, and should not have an important impact on the design of the overall platform.

As such, the design of the flow management is presented in terms of transformations between data states, rather than from the procedural point of view. The resulting system should only rely on the robustness of those data states to ensure the reliability and robustness of the overall system, protecting against potential problems from server failures or other causes. At any point, the system should be able to resume its function purely from the state of the persisted data.

To ensure reliability and performance of the data persistence, we expect to use a well-established and widely used database system such as MongoDB.

Figure 4 shows the general flow of the EUMSSI system, focusing on the data states needed for the system to function.

In order to avoid synchronization issues, the state of the data processing is stored together with the data, and the list of pending tasks can be extracted at any point through simple database queries.

For example in order to retrieve the list of content items that have been crawled or received from feeds, but still need to be converted to the unified EUMSSI schema, it is sufficient to query for items that have a "source_meta:original" but no "source_meta:eumssi".

Similarly, the queues for analysis processes can be constructed directly from the "processing_state" of an item by selecting (for a given queue) all items that have not yet been processed by that queue and that fulfil all prerequisites (dependencies).

[7] https://uima.apache.org/uimafit.html
[8] http://uima.apache.org/doc-uimaas-what.html
[9] http://uima.apache.org/doc-uimaducc-whatitam.html

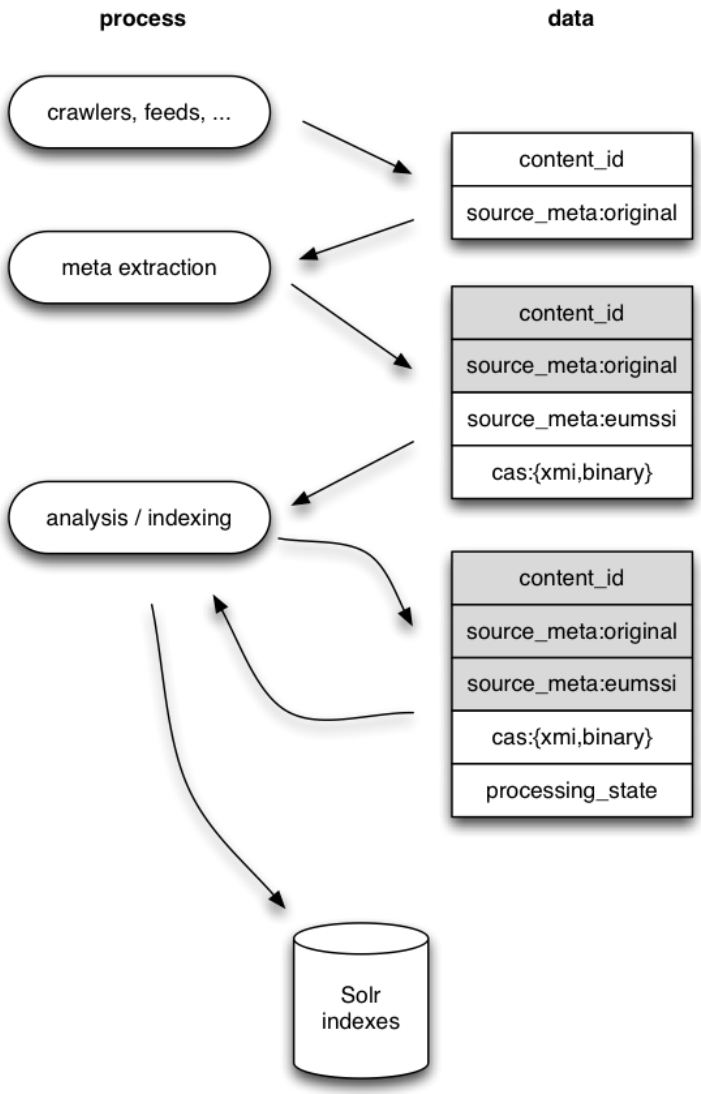

Figure 4: data flow and transformations

As an illustration, each content item has approximately the following structure:

```
{
  "content_id" : UUID,
  "source_meta" : {
    "original" : ORIGINAL_SOURCE_METADATA,
    "eumssi" : EUMSSI_SOURCE_METADATA
  },
  "cas" : {
    "xmi" : XMI_CAS,
    "binary" : BINARY_CAS
  },
  "processing_state" : {
    "queue1" : "done",
    "queue2" : "in_process",
    ...
    "queueN" : "pending"
  },
  "extracted_meta" : METADATA_FROM_CAS
}
```

where:

- **UUID** is a system-wide unique content id, created when first inserting the content into the system
- **ORIGINAL_SOURCE_METADATA** is the metadata as provided from the original content fields
- **EUMSSI_SOURCE_METADATA** is the original metadata mapped to the EUMSSI vocabulary / schema
- **XMI_CAS** is the CAS serialized in XMI format (and possibly compressed)
- **BINARY_CAS** is the CAS serialized in binary format (alternative to XMI_CAS)
- **METADATA_FROM_CAS** is metadata that is generated by EUMSSI analysis processes, using the EUMSSI schema

Normally, the CAS will be stored only in one of the available formats, but potentially different serializations could be used. The "extracted_meta" information can be used for analysis results that are used as inputs to other annotators (such as detected Named Entities as input to speech recognition), to avoid the overhead of extracting that information from the CAS on demand.

MongoDB allows to stored structured information (corresponding to a JSON structure), so that the content of fields like ORIGINAL_SOURCE_METADATA can reflect whatever internal structure the original data had.

The final applications are not expected to use the information stored in MongoDB directly, but rather access Solr indexes created from that information to respond specifically to the types of queries needed by the applications. Those indexes will typically be created from the CAS when all analysis steps have been performed.

It is, however, possible to have indexing processes that only depend on a subset of analyses, and thus make content items (at least partially) accessible to the applications before they have been fully processed (which may take a relatively long time). The indexing processes can be managed in the same way as any analysis process, with their own queues that specify the necessary dependencies, and taking the current state of the CAS as input.

In its simplest form, the processes responsible for the data transitions are fully independent and poll the database periodically to retrieve pending work. Those processes can then be implemented in any language that can communicate comfortably with MongoDB. As an efficiency improvement, in order to reduce the polling load, message queues (such as managed by ActiveMQ[10]) can be used to notify processes of pending work after performing the preceding steps.

7 Conclusions and future work

In this paper, we have presented the main goals and approaches of the EUMSSI project, which aims to innovatively integrate state-of-the-art text and A/V analysis technologies, semantic enrichment and reasoning, social intelligence and collaborative content-based recommendation, in order to build a multimodal, interoperable platform potentially useful for any application in need of automatic cross-media data analysis and interpretation, such as intelligent content management, personalized recommendation, real time event tracking, content filtering, etc.

The project is still in an early stage, and many aspects will need to be defined later on. The different analysis modalities are handled by separate research groups that will each improve the individual types of analysis in their are of expertise. This paper only reports on the platform that will integrate and combine the analysis results.

Additionally, possible interactions between modalities will need to be defined as it becomes clearer what information each analysis can provide or benefit from. We have at this point identified some of the more obvious interactions, such as doing text analysis on speech recognition output, or adding Named Entities from surrounding text to the vocabulary known to the ASR system, but many more may become apparent as the different research groups learn from each other.

[10] http://activemq.apache.org/

One of the main innovative aspects of the project also lies in the combination of the outputs of different analysis layers, and the capacity to perform reasoning or inference over this combined view to create a richer model of the content than can be obtained individually. This is an important research task that has not started yet, and we hope to report on it in the near future. As such, this article is limited to the technological foundation that will enable this work by providing a flexible platform with easy access to all available information layers.

Development of the platform has recently begun and all developments will become publicly available at https://github.com/EUMSSI/.

Acknowledgements

The work presented in this article is being carried out within the FP7-ICT-2013-10 STREP project EUMSSI under grant agreement n° 611057, receiving funding from the European Union's Seventh Framework Programme managed by the REA-Research Executive Agency http://ec.europa.eu/research/rea.

References

Zhineng Chen, Juan Cao, Yicheng Song, Yongdong Zhang, and Jintao Li. 2010. Web video categorization based on Wikipedia categories and content-duplicated open resources. In *Proceedings of the international conference on Multimedia - MM '10*, page 1107, New York, New York, USA, October. ACM Press.

Rasmus Hahn, Christian Bizer, Christopher Sahnwaldt, Christian Herta, Scott Robinson, Michaela Bürgle, Holger Düwiger, and Ulrich Scheel. 2010. Faceted wikipedia search. In *Business Information Systems*, pages 1–11. Springer.

Tom Heath and Christian Bizer. 2011. Linked data: Evolving the web into a global data space. *Synthesis Lectures on the Semantic Web: Theory and Technology*.

Benjamin Köhncke and Wolf-Tilo Balke. 2010. Using Wikipedia categories for compact representations of chemical documents. In *Proceedings of the 19th ACM international conference on Information and knowledge management - CIKM '10*, page 1809, New York, New York, USA, October. ACM Press.

Yehuda Koren. 2008. Factorization meets the neighborhood. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 08*, page 426, New York, New York, USA, August. ACM Press.

Pablo N. Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. 2011. DBpedia spotlight. In *Proceedings of the 7th International Conference on Semantic Systems - I-Semantics '11*, pages 1–8, New York, New York, USA, September. ACM Press.

Matthew Michelson and Sofus A. Macskassy. 2010. Discovering users' topics of interest on twitter. In *Proceedings of the fourth workshop on Analytics for noisy unstructured text data - AND '10*, page 73, New York, New York, USA, October. ACM Press.

Philip V. Ogren and Steven J. Bethard. 2009. Building test suites for UIMA components. *SETQA-NLP '09 Proceedings of the Workshop on Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 1–4, June.