A Platform for Contextual Multimedia Data - Towards a Unified Metadata Model and Querying

Kai Schlegel, Emanuel Berndl,
Michael Granitzer, Harald Kosch
University of Passau, Germany
forename.surname@uni-passau.de

Thomas Kurz
Salzburg Research, Austria
thomas.kurz@salzburgresearch.at

ABSTRACT

Whereas the former Web mostly consisted of information represented in textual documents, nowadays the Web includes a huge number of multimedia documents like videos, photos, and audio. This enormous increase in volume in the private, and above all in the industry sector, makes it more and more difficult to find relevant information. Besides the pure management of multimedia documents, finding hidden semantics and interconnections of heterogeneous cross-media content is a crucial task and stays mostly untouched. To overcome this tendency we see the need for a generic cross-media analysis platform, ranging from extracting relevant features from media objects over representing and publishing extraction results to integrated querying of aggregated findings. In this paper we propose the underlying foundation for a common and contextual multimedia platform in terms of an unified model for publishing multimedia analysis results. The proposed model is based on existing ontologies, adapted and extended to the cross-media environment. Besides the introduction of the already mentioned platform and model, this paper also briefly introduces specific use-case applications as well as possibilities to query the persisted data.

CCS Concepts

•Information systems → Multimedia content creation; Computing platforms; Resource Description Framework (RDF);

Keywords

Cross-Media, Multimedia Analysis, Web Annotation Data Model, Semantic Web, Unified Metadata

1. INTRODUCTION

Due to decreasing costs in media production and viral distribution channels, the amount of multimedia content in the Web and in corporate intranets has increased almost exponentially during the last decade. The pure mass of this data as well as the hidden semantics of raw multimedia content makes it hard to manage and retrieve media assets that satisfy certain information needs. Analysing technology for multimedia content is still not available for everyone, requires expert knowledge, and the processing is costly in terms of time and money. This makes it hard, especially for small and medium-size enterprises, to make use of these technologies. In addition multimedia analysis components typically operate in isolation as standalone applications and therefore do not consider the context of other analysis steps of the same media resource. To address these issues, the open source MICO[1] platform [9] aims at providing an architecture to analyse media in context by orchestrating various analysis components for several media types (video, images, audio, text, link structure, and metadata), with each component depending on each other and contributing to an overall picture of the meaning of the media content. In order to achieve this, MICO provides all necessary technologies for a distributed analysis workflow, such as cross-media extraction, extraction model, component orchestration, metadata publishing, and querying. Instead of publishing the results in proprietary or varying formats, the platform relies on the Resource Description Framework (RDF) [18] to support a common metadata model, which allows semantic interlinking on the level of single resources and comprehensive querying with SPARQL. This combination of Multimedia and Semantic Web is consistent with recent efforts like the W3C Working Groups for Media Annotations[2] or the Group for Fragment URI [23], which are dedicated to make media assets full citizens of the Semantic Web. The W3C Web Annotation Working Group recently issued the Web Annotation Data Model (WADM) [20]. It describes a specification for an interoperable, sharable, and distributed Web annotation architecture. In their point of view, an annotation depicts marginalia or highlights in pictures, videos, audio streams, web pages, or even raw data. In our platform, the Web Annotation Data Model is the primary inspiration for the unified ontology for publishing multimedia analysis results.

The core contributions of this paper includes a **variation of the Web Annotation Data Model in a generic cross-media analysis environment, ranging from representing and publishing extraction results to integrated querying of aggregated findings**. To introduce the environment of the proposed metadata model, the outline of the paper is as follows: Section 2 presents the vi-

i-KNOW '15, October 21 - 23, 2015, Graz, Austria

ISBN 978-1-4503-3721-2/15/10…$15.00

DOI: http://dx.doi.org/10.1145/2809563.2809586

[1] http://www.mico-project.eu/

[2] http://www.w3.org/2008/WebVideo/Annotations/

sion and use-cases of the cross-media analysis platform. The platform itself and its technical background as well as implementation details will be documented in section 3. A short summary of the WADM and our adaption, which resulted in the MICO multimedia metadata model, is covered in section 4. The usage and consumption of the model in form of semantic querying of multimedia analysis results is illustrated in section 5. Section 6 gives insights about related research topics and section 7 concludes the paper with a summary of our contributions.

2. VISION AND FIELD OF APPLICATION

The main vision of the open and extensible MICO platform is bridging the gap between multimedia objects and their hidden and usually interconnected semantics. Given the fact that most of the multimedia content (text, image, audio & video) is integrated in *information units* (spatially related or linked bundles of diverse content formats that are combined to illustrate a certain semantic topic, event, or fact), the platform makes use of all surrounding information to enrich the pure content, align existing and new metadata into a common model and provide access for the emerging *cross-media data*. As examples for such *information units*, imagine an (online) newspaper article with embedded images, a YouTube video combined with the accompanying text documentation, or a music file in conjunction with lyrics, all describing the same semantic content in context. The MICO project developed models, standards and software tools to jointly analyse, query and retrieve information out of connected and related media objects to provide better information extraction for more relevant search and information discovery. The platform allows arbitrary workflow orchestration of different extractors, or custom extensions, each adding its bit of additional information to the final result. Figure 1 shows an example, depicting a workflow with two independent extractors chained to a workflow to perform face detection on important frames of a video.

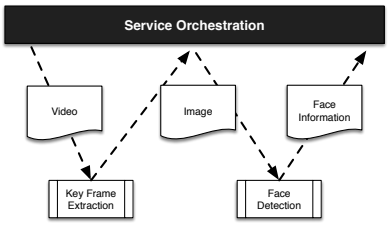

Figure 1: Exemplary multimedia analysis workflow

Extractors for cross-media coherences can be for instance a language detector (identifying the language of a text or audio track), a keyframe detector (identifying relevant images from a video), a face detector (identifying objects that could be faces), a face recogniser (assigning faces to people), an entity linker (assigning objects to specific entities) or a disambiguation component (resolving possible alternatives by choosing the more likely, given the context).

The vision and requirements of the MICO platform were driven by several existing use-cases, which are set in disparate fields of application. The first use case composes the citizen-science crowdsourcing platform Zooniverse[3] with the Snapshot Serengeti project[4]. Here, animal detection and classification is done with active participation of human volunteers. The MICO platform supports the crowdsourcing process with cross-media analysis like:

- to avoid wasting resources by applying visual analysis to filter out empty images to pre-classify images and to estimate the number of animals
- to use semantic or sentiment analysis of text posts and discussions to determine what users are interested in and to notify moderators about controversial topics
- to analyse user behaviour and accuracy scores from past classifications and use this data to recommend next images for volunteers based on preference and expertise

The second use case is a social news platform by InsideOut10[5], which combines image, video, audio, and user-generated content. Users are able to upload related recordings to given news articles. It faces challenges related to automatic quality assessment of video, nudity detection, speech recognition, tampering detection, etc. The scenario includes the need to semantically query videos, along the line of point me to scenes within videos where a specific person is talking about a specific topic, and show me similar content to related topics. As with the crowdsourcing platform, MICO technologies are used to tackle key challenges of the platform:

- combine various extractors for different media types, in this case extraction of tags, language detection, scene detection, automatic speech recognition (ASR), video segmentation, face detection, speech-music discrimination, and subsequent named entity recognition
- bridge the gap between high-level user queries and context-specific low-level queries that include e.g. feature combination and weighting of extractor outputs considering the context
- recommend similar semantic topics, e.g. based on key word co-occurence in different video item transcriptions

Both use cases differ in their inputs and applicable environment, but they make use of generic extractor components which only differ in context-specific, dynamic orchestration and parametrisation, since both the configuration of the analysis setup and the extractor parameters depend on the context. New use cases can easily reuse existing components and a simple plugin API allows to add extractors, whereof the whole platform and possibly other existing workflows can benefit.

3. PLATFORM ARCHITECTURE

After introducing the overall vision and specific use cases of the MICO platform, this section will present the conceptual and technical overview. The platform is an environment that allows to break up the hidden semantics of

[3]https://www.zooniverse.org/
[4]http://www.snapshotserengeti.org/
[5]http://insideout.today/

media in context by orchestrating sets of different components that jointly analyse content, each adding its bit of additional information to the final result. The main workflow implemented by the framework is the analysis of media *information units*, in this context called content items, and subsequent publishing of analysis results, either for search, querying, and recommendations. In the MICO platform, **content items** are representations of media resources together with their subsequent analysis results. It consists of a collection of multiple **content parts**, which are results of analysis components with different media types that are directly related to the same **content item**. In other words, a **content item** is a semantic grouping of information objects (**content parts**) considering a specific multimedia asset.

At its core, the platform uses a distributed Service-Oriented Architecture (SOA) [14], where each analysis process (extractor) is an individually managed component and can be dynamically composed to custom workflows. Figure 2 depicts this high-level architecture, where an orchestration service plays a crucial role.

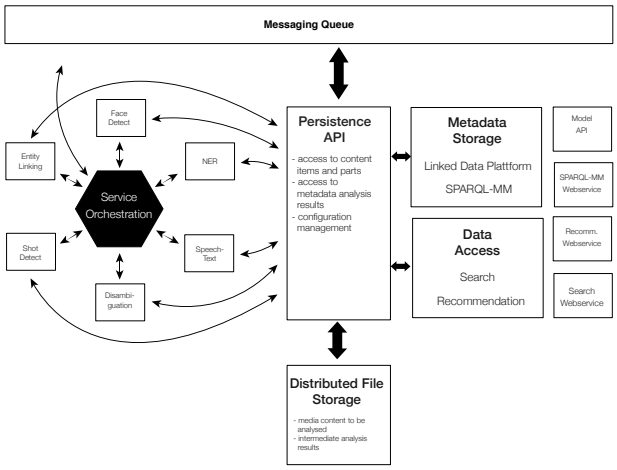

Figure 2: Overview of platform architecture

Execution plans make use of known enterprise integration patterns provided by Apache Camel[6] and are based on declarative service descriptions specifying the input and output dependencies of the services. Each extractor is an independent process, without any restrictions about its resources, neither physical location nor runtime environment. Consequently, extractors might be implemented in different programming languages and run on different servers, even in efficient cluster configurations. The storage process triggers a dynamically orchestrated process across the different extractors. As different analysis services potentially run on different servers, the platform offers shared persistence and communication components. This is supported by a messaging queue for communication (RabbitMQ[7]), a Linked Data Platform [17] for result metadata storage (Apache Marmotta[8]) and an unstructured distributed data store to persist the raw media files (Apache Hadoop HDFS[9]). To summarise, an overall workflow will consider following steps successively:

1. a new content item is loaded into the framework, (temporarily) stored by the shared persistence component and a batch process for analysis is started
2. the service orchestration component builds an execution plan, stores it with the content item metadata in the persistence component, and signals that a new content item is available for analysis
3. analysis components process the content item according to the execution plan, storing additional analysis results as part of the content item metadata in the persistence component, and signal other succeeding components once they are finished
4. the analysed content item is either exported for further processing, or made available for access, querying, or recommendations inside the platform

All results that are produced by the orchestrated analysers (intermediate as well as final results) are made available as Linked Data for further access and follow the specially designed RDF vocabulary (see section 4), which can be produced directly by the analysis service or via Java and C++ APIs. A common persistence API simplifies the use of these different components and provides access to content items, metadata, configuration management, and event management. The platform itself contributes its main software development results as Open Source components to simplify the use of the technology in industrial products. The platform and a detailed documentation can be found online at http://www.mico-project.eu/platform/.

4. A METADATA MODEL FOR MEDIA IN CONTEXT

As extraction and analysis processes in general produce a manifold of different results and output formats, unified ways of storing and requesting the content and its context are required. To enable comprehensive cross-media querying, the underlying data model must be capable to reflect the performed workflow chain including the interaction and mutual dependence between all analysis steps. Metadata as well as provenance can be persisted in close relation to the multimedia content. Arising from the MICO vision and the cross-media use-cases, various requirements originated for the metadata model:

- **Cross-Multimedia Application**: The model needs to be able to support various different multimedia metadata formats in an uniform and query-able fashion.
- **Different Annotation Targets**: Besides the variety of formats, the model needs to support a way of specifying temporal and/or spatial fragments of media assets. Extractors might not utilise the whole asset as input, but rather a subpart of it.
- **Extensibility**: The fields of multimedia and its analysis are quickly evolving, as new extraction techniques

[6]http://camel.apache.org/
[7]https://www.rabbitmq.com/
[8]http://marmotta.apache.org/
[9]https://hadoop.apache.org/

and ways of classifying and analysing multimedia assets are developed further and further. The designed ontology needs to be extensible in order to ingest new extractors with their yet unknown formats.

- **Provenance**: Especially in the process of annotating multimedia assets, provenance is an important feature. Versioning, timestamps, the tracking of workflows, and confidence or trust values are only examples.

The Web Annotation Working Group defines the term *web annotation* as a piece of further description (e.g. marginalia or highlight) for a digital resource like a comment or tag on a single web page or image, or a blog post about a news article. Annotations are used to convey information about a resource or associations between resources. The Web Annotation Working Group also provides an RDF-based Web Annotation Data Model (WADM), which is derived from the Open Annotation Data Model [21]. A basic web annotation consist of two base concepts - the **body** and the **target**. Those two are joined by an annotation entity, which can be further extended by provenance information (see figure 3).

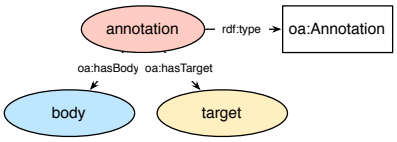

Figure 3: Basic annotation in the OADM

The body of an annotation contains the actual content of the annotation. The target is the "thing" that the annotation is about. The WADM standard states, that the "body is somehow about the target". An example would be a picture as the target of the annotation, while the body could be a text comment given by a user. The definition of a web annotation leaves space for interpretation and application. Making use of the definition as well as the specified ontology, we developed an extension to lift web annotations to another field of application and give the web annotation definition a new facet. The resulting ontology for cross-media metadata publishing allows the design of multimedia extraction workflows in the MICO platform.

At the end of every MICO extraction step, an **annotation** is created, describing the media analysis or intermediary result. It then will be created and appended to its corresponding content part. Annotations can refer to preceding annotations as input. Altogether, the content item will represent an accumulation of extraction results, metadata background, and workflow details of a given multimedia asset.

4.1 Utilised Ontologies

The model has to cover a variety of content, which has to be supported by respective ontology vocabulary. In order to facilitate this, we make use of existing ontologies wherever possible. On instances that could not be covered by existing ones, we made enhancements to the base WADM specification by adding our own ontology.

The base format is posed by the Resource Description Framework RDF [18] and its schema RDFS [10]. In terms of modelling annotations, the Web Annotation Data Model is used. We will describe its core in section 4.2, as it is the most important model component. Because we are mainly dealing with multimedia items and files, an ontology for describing them is needed. Dublin Core DC [4], especially with its subparts for **terms** and **types**, are used here. For provenance modelling, we will utilise the PROV Ontology PROV-O [19], which supports a wide range and variety of possibilities to illustrate provenance information. The FOAF Vocabulary Specification [5] is an ontology allowing to link people and model the relationships between them. We will make use of this ontology in order to depict and further enrich provenance information. For more specific content descriptions, the Representing Content in RDF 1.0 [13] ontology is applied.

4.2 Annotations and Ontology Modules

Our model and the associated ontology with the namespace http://www.mico-project.eu/ns/platform/1.0/schema# are extensions to the specification given by the Web Annotation Working Group. The baseline of three concepts, namely *annotation*, *body*, and *target*, are adopted. In order to implement the workflow chains and constitution of multimedia extractions, we introduce a *composition* module, which covers vocabulary to express the aforementioned features. Extended *provenance* information are applied at every level. Figure 4 shows all the modules of our ontology and their interplay.

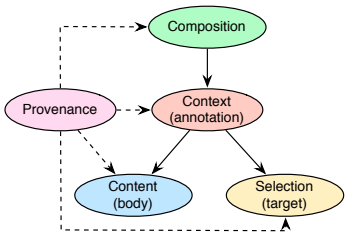

Figure 4: Annotation and model structure

In the following we will show exemplary RDF fragments and graphs about every module. These examples are shown in extracts but specific about the given feature. In all figures, classes will be illustrated with rectangles, while instances are depicted as ovals. We use the namespace abbreviation `entity:` for generic instances of an entity.

Content - The Body.

The content module contains ways of describing and classifying the results or outcomes of the extractors. In general, every extractor has its own output schema. A content feature of an annotation consists of a base node, which is then further specified. The base node is typed as a subclass of `mico:AnnotationBody`, according to the respective extraction process. In addition to this, the base node has various relationships which contain the values of the extraction result. A connection to the annotation node (which connects the content to its selection) is done via the relationship `oa:hasBody`. The current state of the ontology supports a range of extractors that are involved in the MICO context

(e.g. `mico:FaceRecognitionBody` or `mico:NERBody`), but the model allows for easy additions of new extractors. Figure 5 shows an example of the content portion of a MPEG7 color-layout descriptor [12]. It was derived from a corresponding low-level-feature extraction XML result shown in in listing 11.

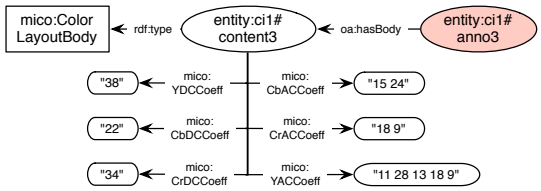

Figure 5: Example content module instance of a Colorlayout annotation

```
<VisualDescriptor xmlns="urn:mpeg:mpeg7:schema:2004"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:type="ColorLayoutType">
  <YDCCoeff>38</YDCCoeff>
  <CbDCCoeff>22</CbDCCoeff>
  <CrDCCoeff>34</CrDCCoeff>
  <YACCoeff5>11 28 13 18 9</YACCoeff5>
  <CbACCoeff2>15 24</CbACCoeff2>
  <CrACCoeff2>18 9</CrACCoeff2>
</VisualDescriptor>
```

Listing 1: Output of a Colorlayout extractor

Selection - The Target.

The specification of what an annotation is about is covered by the selection module. In our case, the target of the annotation is the initial multimedia asset at the start of an extraction workflow, or an intermediary/end result of a workflow. We extend the usage of the target concept of the WADM specification in the way that we always make use of a `oa:SpecificResource`. A specific resource is used in order to further describe the target, and therefore it is connected to the base annotation node via the relation `oa:hasTarget` and to the actual source with `oa:hasSource`. From this point, we distinguish between two selection types. The *basic selection* does not require any further details or features, while the *extended selection* is refined by the utilisation of a `oa:Selector`. Figure 6 shows an example.

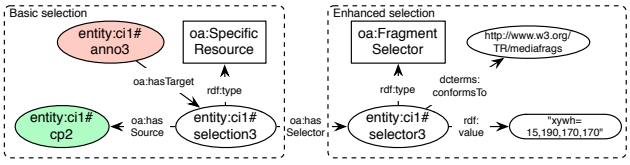

Figure 6: Example selection module instance of a Media Fragments based annotation

On the left side, one can see the basic assembly of annotation and its specific resource, which then points to another content part (`entity:ci#cp2`) specifying the input for this extraction. The addition of a selector (`entity:ci#selector3`) on the right side of figure 6 reflects the enhanced selection.

In this example, a `oa:FragmentSelector` is used in order to select a spatial fragment from the corresponding input asset. The actual fragment is specified by the relation `rdf:value`, the selection is made conform to the W3C Media Fragments [23]. The example would select a rectangle with height and width of 170 pixels, starting with its top left corner at the pixel with position 15, 190 of the given input. The WADM already supports various different selector types.

Context - The Annotation.

The actual annotation is constituted by the context module, representing the semantic recombination of content and selection. The annotation node on its behalf connects the corresponding nodes. As derived from the WADM specification, an annotation is always of the type `oa:Annotation`. The relationships are also inherited. `oa:hasTarget` points to the selection node, `oa:hasBody` towards the content node. Figure 7 shows a small example.

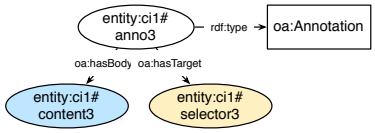

Figure 7: Exemplary context module instance with content and selection

Composition and Provenance - Joining Annotations and Creating Workflows.

The key adaption our of model in contrast to the WADM is given by the composition and provenance elements. They allow to bundle the annotations to represent whole workflow chains and join the intermediate results of various extractors to form a combined metadata background for the original multimedia asset. An example can be seen in figure 8.

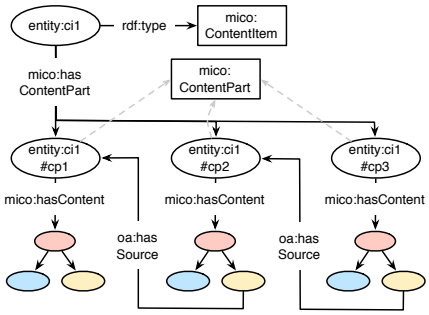

Figure 8: Exemplary composition module instance with content and selection

Three annotations (which represent direct results of extractions) can be seen on the bottom. For each of the annotations, a `mico:ContentPart` is created and then joined via the `mico:hasContent` relation. A content part represents a workflow step. All content parts for one ingested multimedia asset are subsumed under an instance of `mico:ContentItem`, which represents the whole analysis and resulting metadata of the asset. A workflow pattern of successive steps will cre-

ate "backlinks" (`oa:hasSource`) from annotations to their respective input. This can be seen at the second and third annotation in figure 8. It shows that both the extractors as well as their results are dependent on each other.

As stated in the requirements, provenance, which basically refers to the documented history and origin of data, is an important feature that has to be supported for many facets of the multimedia model. Figure 4 shows that provenance will take part in every of the other four modules. Provenance is an important feature in order to trace back resulting extraction workflows, assign confidence or trust values to given results (e.g. you can trust the result of a process only to a given degree, or you can give a quality mark to different workflows with the same task), and simply track who and when a specific result has been processed. Timestamps will be stored for content items and content parts (when they are created at the platform) as well as for annotations. The latter constitutes the point of time when an annotation has been created as the result of an extraction step. The agents taking part in the platform, in this case mostly software agents, although it is possible to introduce humans as annotating agents, will also be backed up with provenance metadata. Basic data for the extractors will contain a unique id, a timestamp of generation and invalidation, its version, and its input and output in a descriptive manner. For users personal data is stored.

5. MODEL QUERYING

The value of connecting media to its semantic metadata is limited to the provided access methods specialised for media assets and fragments. Hence, the proposed model conforms with the recently issued SPARQL-MM specification [15]. SPARQL-MM extends SPARQL, the standard query language for the Semantic Web with media specific concepts and functions to unify the access to media results. SPARQL-MM supports various multimedia specific features like **Query-by-spatial-relationship** (object A is left beside object B), **Query-by-temporal-relationship** (object A appears after object B) or **Query-by-resource-similarity** (face recognition function, concept detection, etc.) and is thus perfectly suitable for our needs. SPARQL-MM defines some core classes that are necessary to describe spatial-temporal properties as well as relational and aggregational functions. On the one hand, SPARQL-MM classes can be used to make more specific statements about the *Selection* module of our model (e.g. face *a* is right beside face *b* on a picture) and on the other hand Apache Marmotta, the triple store used in our platform, already supports backend specific implementations for SPARQL-MM which allow media fragment extraction (e.g. cropping a image for a specified spatial media fragment) or query optimisation by using inherent SPARQL-MM function characteristics. Listing 2 shows how SPARQL-MM can be used to query semantic relations between persisted annotations. In this example, we query for media assets where "Barack Obama" stands left besides "Angela Merkel". Therefore we select annotations with face recognition bodies and compare their targets with a SPARQL-MM filter. The filter implementation directly compares the corresponding media sources and existing spatial media fragments. To crop the media result, we can bind the bounding box of both face recognitions with the result.

All the aforementioned features combined with SPARQL-MM functionality allow our model to meet the requirements defined in section 4. Cross-media usage is enabled and storage as well as querying of multimedia files in context with their metadata is possible in a unified way. The model allows a rich and extensible way of defining the annotations both on the content and selection side. The whole workflow chain as well as participating agents are backed up by provenance information. This permits the platform to be utilised in a cross-cutting multimedia context and rich fields of application.

```
1  SELECT ?result WHERE {
2    ?anno1    oa:hasBody     ?faceRec1;
3              oa:hasTarget   ?target1.
4    ?anno2    oa:hasBody     ?faceRec2;
5              oa:hasTarget   ?target2.
6
7
8    ?faceRec1 a              mico:FaceRecognitionBody;
9              rdf:value      "Barack Obama".
10   ?faceRec2 a              mico:FaceRecognitionBody;
11             rdf:value      "Angela Merkel".
12
13   FILTER mm:leftBeside(?target1,?target2)
14   BIND (mm:boundingBox(?target1,?target2) AS ?result)
15 }
```

Listing 2: Exemplary SPARQL-MM query

6. RELATED WORK

Various literature explore the management of multimedia assets or metadata, partially in the context of the Semantic Web, and therefore can be considered as related to our topic. One of the de-facto standards in the fields of describing multimedia content information is MPEG-7 [7], which has been developed by the Moving Pictures Expert Group MPEG[10]. It covers representing metadata of still pictures, graphics, 3D models, audio, speech, video, and combinations of them using a set of descriptors and description schemes. It supports a XML-based language (MPEG-7 Description Definition Language) as well as schemes (MPEG-7 Description Schemes) in order to implement those descriptors. Using these, metadata (e.g. low-level features like colour moments, histograms or shot detections) about a multimedia file can be stored together or separately with the multimedia data.

Although the possibilities to describe and query multimedia items in conjunction with their metadata are manifold [22, 16, 1], there are still certain difficulties in systems where users of different domains meet. Yu [26] tries to close the perceived gap between users' interpretation and the richness of the multimedia content. The author states, that users from different domains might see multimedia content from different points of view and consequently they also create diverse metadata when tagging or annotating multimedia content. This circumstance can also interfere in terms of querying. In addition to this problem, most multimedia systems do only support one representation of its multimedia content, so users experience difficulties when trying to define their information needs or consume the content. The process described in [26] combines an ontology for MPEG-7 [11] with other existing descriptive ontologies. It supports a semantic web vocabulary that enables users to tag specific parts of the multimedia content with the expression tools of other or own ontologies. Doing this, one content can be

[10]http://mpeg.chiariglione.org/

tagged by different ontologies and accordingly have different representations.

Also based on MPEG-7 and referred to by Yu, Hunter [11] used reverse engineering and top-down processes in order to design an RDF equivalent ontology of the MPEG-7 standard. This enables MPEG-7 to be utilised in the Semantic Web to be more flexible and interconnected.

Specialised for describing media resources published on the Web, the W3C Ontology for Media Resource [6] intends to bridge different descriptions formats and defines a interoperable set of metadata properties for media resources, along with their mappings to elements from a set of existing metadata formats like Dublin Core, EXIF, or MPEG-7. The W3C Ontology for Media Resource provide a Semantic Web compatible ontology using RDF/OWL and facilitates cross-community data integration of multimedia metadata.

Another semantic web ontology approach is presented in [3, 2], where Arndt et al. describe COMM, a core ontology for multimedia. They claim, that the MPEG-7 standard is very well designed, but in the annotation world of today, lacks some capabilities. Amongst these are *Fragment Identification* (MPEG-7 supports a variety of identifiers, but a unified and agreed upon way is essential), *Semantic annotation* (again, MPEG-7 provides a range of descriptors, but they have to be interpretable at different agents), *Web interoperability* (MPEG-7 excludes web capabilities), and the *Embedding into compound documents*. They also define several requirements that should be met by the design of a multimedia ontology. They state that an ontology and its results must be both semantically and syntactically interoperable, which means, that the extractors need to have a common syntax that is shareable as well as understandable and interpretable at other locations. Furthermore, the ontology needs to support a separation of concerns, so that there is a clear distinction between the actual extraction information and structural knowledge. Modularity as well as extensibility are also marked as key features of a well designed ontology. Besides all these facts, the MPEG-7 standard [7] is mentioned with its importance in the fields of the multimedia world. An ontology for multimedia should express the whole standard, or it should at least be possible to integrate MPEG-7 on its own. The result is an RDF based ontology, that implements the MPEG-7 semantics and allows for other extensions. It is open and integrate-able with the web as well as domain independent, allowing it to be used in a very wide context, connecting semantic annotations with multimedia assets or only subparts of that asset.

Besides the pure modelling of multimedia content, the creation and extraction of this metadata is major issue and sometimes even needs to be assisted by the user. Different approaches from web-tools to semi- and full-automated platforms cover the landscape of todays annotation scenery. The most related approach to the platform described in this paper is given by Verborgh et al. [25], who propose a platform that connects its analysis process with the Semantic Web. They claim that most extractors on their own are unaware of their context, which has impact on its extraction quality as well as great interference when it comes to workflow composition of several different extractors. By retrieving available information about the extractor via the Semantic Web, the context for its process is determined. This allows to combine the services in various workflows that can generate a rich background for the injected multimedia items. Additionally, the platform itself is able to find alternative extraction routes in case of an error or unexpected behaviour of a specified workflow. The extractors or services are embedded as SPARQL endpoints [24], its input and output is formulated as RDF statements. Using OWL-S [19], the algorithm as well as the relationships between the extractors can be described. The platform itself implements a blackboard architectural pattern [8], which is composed by several components. The blackboard serves as a institution that collects and holds received information. Services are the working instances of the whole platform. A supervisor in combination with a service composer triggers services accordingly to the requested processes and finds alternative problem solutions in case of errors.

7. CONCLUSION

In this paper we presented an extensible model for bridging the gap between multimedia object analysis and the Semantic Web. Using a flexible service orchestration the platform will enable cross-media analysis along the loosely coupled distributed analysis chain. At its core, the platform features the highlighted harmonised metadata model for unified publishing, homogeneous processing, and integrated semantic querying of multimedia analysis results. Based upon Semantic Web technologies, the model allows analysis results of videos, images, audio, and text to be lifted in an overall interconnected environment including a traceable chain of origins and transformations. As shown in the related work, MPEG-7 is an important baseline for multimedia ontologies. The presented model does not utilise the standard by default, but rather offers the possibility to easily integrate all of its features. The WADM was extended with the possibility to recombine web annotations to depict a workflow chain. Provenance features enable full traceability of origins, and they enhance the workflow structure of the results. The model is extensible and is already integrated in the MICO platform. Actively ongoing use cases like the citizen-science crowdsourcing platform Zooniverse and the social news platform by InsideOut10 demonstrate and validate the application of the model in the real world and will influence the next iteration of the model ontology. The initial configuration of the platform includes a multitude of different multimedia analyser components like face detection, temporal video segmentation, speech-music discrimination, speech-to-text, or sentiment analysis. Furthermore, the platform provides a Java API to read and write the MICO data model, which we also published as Open Source library Anno4j[11] for generic Web Annotations.

8. ACKNOWLEDGEMENTS

The presented work was developed within the MICO project partially funded by the EU Seventh Framework Programme, grant agreement number 610480.

9. REFERENCES

[1] M. Arias Gallego, O. Corcho, J. D. Fernández, M. A. Martínez-Prieto, and M. C. Suárez-Figueroa. Compressing semantic metadata for efficient multimedia retrieval. In *Advances in Artificial*

[11] https://github.com/anno4j/anno4j

Intelligence, volume 8109 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2013.

[2] R. Arndt, R. Troncy, S. Staab, and L. Hardman. Comm: A core ontology for multimedia annotation. In *Handbook on Ontologies*. Springer, 2009.

[3] R. Arndt, R. Troncy, S. Staab, L. Hardman, and M. Vacura. *COMM: designing a well-founded multimedia ontology for the web*. Springer, 2007.

[4] D. C. U. Board. DCMI metadata terms. Dcmi recommendation, DMCI, June 2012. http://dublincore.org/documents/dcmi-terms/.

[5] D. Brickley and L. Miller. FOAF vocabulary specification. Technical report, xmlns, Jan. 2014. http://xmlns.com/foaf/spec/.

[6] P.-A. Champin, T. Bürger, T. Michel, J. Strassner, W. Lee, W. Bailer, J. Söderberg, F. Stegmaier, J.-P. EVAIN, V. Malaisé, and F. Sasaki. Ontology for media resources 1.0. W3C recommendation, W3C, Feb. 2012.

[7] S.-F. Chang, T. Sikora, and A. Purl. Overview of the mpeg-7 standard. *Circuits and Systems for Video Technology, IEEE Transactions on*, 11(6):688–695, 2001.

[8] D. D. Corkill. Blackboard systems. *AI expert*, 6(9), Sep. 1991.

[9] S. Fernández, S. Schaffert, and T. Kurz. Mico: Towards contextual media analysis. In *Proceedings of the 24th International Conference on World Wide Web Companion*, WWW '15 Companion, 2015.

[10] R. Guha and D. Brickley. RDF vocabulary description language 1.0: RDF schema. W3C recommendation, W3C, Feb. 2004. http://www.w3.org/TR/2004/REC-rdf-schema-20040210/.

[11] J. Hunter. *Adding multimedia to the Semantic Web-Building and applying an MPEG-7 ontology*. Wiley, 2005.

[12] E. Kasutani and A. Yamada. The mpeg-7 color layout descriptor: a compact image feature description for high-speed image/video segment retrieval. In *Image Processing, 2001. Proceedings. 2001 International Conference on*, volume 1. IEEE, 2001.

[13] J. Koch, C. A. Velasco, and P. Ackermann. CNT representing content int rdf 1.0. W3C recommendation, W3c, May 2011. http://www.w3.org/TR/Content-in-RDF10/.

[14] D. Krafzig, K. Banke, and D. Slama. *Enterprise SOA: service-oriented architecture best practices*. Prentice Hall Professional, 2005.

[15] T. Kurz, K. Schlegel, and H. Kosch. Enabling access to linked media with sparql-mm. In *Proceedings of the 24nd international conference on World Wide Web (WWW2015)*, LIME15, 2015.

[16] S. Laborie, A. Manzat, and F. Sèdes. Managing and querying efficiently distributed semantic multimedia metadata collections. *MultiMedia, IEEE*, PP(99), 2009.

[17] A. Malhotra, J. Arwe, and S. Speicher. Linked data platform 1.0. Candidate recommendation, W3C, June 2014. http://www.w3.org/TR/2014/CR-ldp-20140619/.

[18] F. Manola and E. Miller. RDF primer. W3C recommendation, W3C, Feb. 2004. http://www.w3.org/TR/2004/REC-rdf-primer-20040210/.

[19] D. Martin, M. Burstein, J. Hobbs, and O. Lassila. OWL-S semantic markup for web services. W3C recommendation, W3C, Nov. 2004. http://www.w3.org/TR/prov-o/.

[20] R. Sanderson and P. Ciccarese. WADM web annotation data model. Community draft, W3C, Dec. 2014. http://www.w3.org/TR/annotation-model/.

[21] R. Sanderson, P. Ciccarese, and H. V. de Sompel. OADM open annotation data model. Community draft, W3C, Feb. 2013. http://www.openannotation.org/spec/core/.

[22] F. Stegmaier, M. Döller, H. Kosch, A. Hutter, and T. Riegel. Air: Architecture for interoperable retrieval on distributed and heterogeneous multimedia repositories. In *Analysis, Retrieval and Delivery of Multimedia Content*. Springer, 2013.

[23] R. Troncy, D. V. Deursen, E. Mannens, and S. Pfeiffer. Media fragments URI 1.0 (basic). W3C recommendation, Sept. 2012. http://www.w3.org/TR/2012/REC-media-frags-20120925/.

[24] R. Verborgh, D. Van Deursen, J. De Roo, E. Mannens, and R. Van de Walle. Sparql endpoints as front-end for multimedia processing algorithms. In *Proceedings of the Fourth International Workshop on Service Matchmaking and Resource Retrieval in the Semantic Web (SMR2 2010)*, 2010.

[25] R. Verborgh, D. Van Deursen, E. Mannens, C. Poppe, and R. Van de Walle. Enabling context-aware multimedia annotation by a novel generic semantic problem-solving platform. *Multimedia Tools and Applications*, 61(1), 2012.

[26] C.-L. Yu. A multimedia access platform based on multi-ontology. *International Journal of Advancements in Computing Technology*, 5(3), 2013.