M3O: The Multimedia Metadata Ontology

Carsten Saathoff and Ansgar Scherp

ISWeb, University of Koblenz-Landau, Universitätsstr. 1, Koblenz 56070, Germany
{saathoff,scherp}@uni-koblenz.de

Abstract. We propose the Multimedia Metadata Ontology (M3O), a framework for integrating the central aspects of multimedia metadata. These central aspects are the separation of the information conveyed by multimedia items and their realization, the annotation with both semantic and low-level metadata, and the decomposition of multimedia content. M3O bases on Semantic Web technologies and provides the means for rich semantic annotation using further, possibly domain-specific ontologies. Moreover, it can be used to represent other existing metadata models and metadata standards. We introduce the M3O and present its application at the example of a SMIL presentation.

1 Introduction

Multimedia metadata and semantic annotation of multimedia content is the key-enabler for improved services on multimedia content. The archiving, retrieval, and management of multimedia content becomes very hard if not even practicably infeasible if no or only limited metadata and annotations are provided. Looking at the existing metdata models and metadata standards, we find a huge number and variety serving different purposes and goals. In addition, the models are of different scope and level of detail. Typically, the existing models cannot be combined with each other. For example, image descriptions using EXIF [1] can not be combined with MPEG-7 [2] descriptors. In addition, the existing models are semantically ambiguous, i.e., they do not provide a well-defined interpretation of the metadata. For example, in IPTC [3] the location fields are defined to contain the locations the content is "focusing on". However, it remains unclear what this "focusing on" actually means. For instance, consider an image from the atomic bombing of Nagasaki in Japan in 1945. This image is about Nagasaki since it documents an event taking place in that city. But it is also about the world as a whole since the atomic bombing is of global importance. Distinguishing these different roles a location can play is impossible with IPTC. In general, support for semantic annotations using formally defined background knowledge is hardly found. Finally, the models are typically focused on a single media type, ignoring the type's relation to other media types or their context within a true multimedia presentation. As a consequence of this, providing interoperability between different applications that deal with the storage, retrieval, and delivery of multimedia content and single media assets annotated with today's models becomes very hard. However, this is required in many multimedia application scenarios, in particular in the open world of the Web.

What is missing is a representation of the data structures that underlie today's multimedia metadata models and metadata standards. We aim at extracting the common patterns underlying existing metadata models and metadata standards. We provide these patterns as a set of *ontology design patterns (ODPs)* [4]. It provides a comprehensive modeling framework for representing arbitrary multimedia metadata and is called the Multimedia Metadata Ontology (M3O). Basing the M3O on Semantic Web and ontologies particularly provides support for the rich semantic annotation of multimedia content.

2 Annotating Structured Multimedia Content

Using a simple scenario, we show the different requirements that need to be considered when annotating structured multimedia content. We assume that we need to give a lecture on discussing the advantages and disadvantages of nuclear energy. For this lecture, we have prepared a multimedia presentation shown in Figure 1 to start discussions. Both for later retrieval and for descriptive purposes, we would like to annotate the presentation. The multimedia content of our multimedia presentation consists of different single media assets. These media assets are combined in a coherent, structured way. This means that the content provides a spatial layout and a temporal course and also includes interactivity. The multimedia content is encoded using the multimedia presentation format SMIL[1] and rendered using the RealPlayer[2].

(a) (b)

Fig. 1: An image of Albert Einstein combined with an image of the Times Square and an image of the nuclear bomb cloud expressing contrary views on *nuclear energy*.

Our SMIL presentation discussing the advantages and disadvantages of nuclear energy consists of two parts. The first part depicted in Figure 1a shows a picture of Albert Einstein[3] and a photo of the Times Square in New York.

[1] Synchronized Multimedia Integration Language, `http://www.w3.org/TR/2008/REC-SMIL3-20081201/`

[2] RealNetworks, Inc., 2009, `http://www.real.com/realplayer/`

[3] `http://en.wikipedia.org/wiki/File:Einstein1921_by_F_Schmutzer_4.jpg`, from Wikipedia. The image is in the public domain.

This part of the presentation serves as a metaphor for the achievements reached by the discovery of nuclear energy in which Einstein played a central role. By the peaceful use of nuclear energy, it can serve large cities like New York with electricity.

In the second part of our SMIL presentation depicted in Figure 1b, we replace the photo of the Times Square by a picture showing the atomic bombing of the city of Nagasaki[4] in Japan in 1945. The picture of Einstein remains unchanged. However, the contextual use in which the picture of Einstein is shown is completely different. By this change of contextual use, the media assets composed transmit a totally different message and express a different semantics [5]. Instead of showing the advantages of nuclear energy, this part of the presentation serves as metaphor for the risks and the potential destructive power of nuclear energy.

For providing a comprehensive semantic description of this multimedia presentation, there are different kinds of annotations involved. These different annotations put requirements to the metadata model used to represent the semantics of the multimedia content shown in the presentation. We discuss these requirements in the following section.

3 Requirements on a Multimedia Metadata Model

From the scenario above, we can derive three principal requirements that need to be supported for annotating rich, structured multimedia content such as the SMIL presentation in the scenario. These requirements are the separation between information objects and information realization, multimedia annotation, and multimedia decomposition. They need to be reflected by a multimedia metadata ontology.

Separation between Information Objects and Information Realizations. On the conceptual level, multimedia content conveys information to the consumer. As such, the multimedia content plays the role of a message that is transmitted to a recipient. Such a message can be understood as an abstract information object [6]. Examples of information objects are stories, stage plays, or narrative structures. The information object can be realized by different so-called information realizations [6]. The narrative structure of our scenario above is, e.g., realized in a SMIL presentation. The following requirements of multimedia annotation and multimedia decomposition can be applied on both levels of information objects and information realization.

Annotation of Information Objects and Information Realizations. The model needs to support the annotation of multimedia content. This can be annotations in the style of typed key-value pairs as provided, e.g., by EXIF or semantic annotation, i.e., the use of semantic background knowledge for describing the multimedia content. In our example, we could annotate the picture from the

[4] `http://commons.wikimedia.org/wiki/File:Nagasakibomb.jpg`, from Wikimedia Commons. The image is in the public domain.

Times Square with the geo-coordinates where it was taken or annotate the whole presentation with the general topic it discusses. Please note that low-level metadata, such as EXIF, typically is attached to the realization, while the semantic annotation rather applies to the information object.

Decomposition of Information Objects and Information Realizations. Multimedia content can be decomposed into its constituent parts. The SMIL presentation above can, e.g., be decomposed into the two parts it consists of. The parts can be decomposed into the images they contain. The realization of the presentation can be decomposed into the realizations of the contained images. Decomposition can be applied arbitrarily often, i.e., we can can create a hierarchy of parts.

4 Related Work

In research and industry, numerous metadata models and metadata standards have been proposed so far. These models come from different backgrounds and with different goals set. They vary in various aspects such as the domain for which they have been designed. The models can be domain-specific or designed for general purpose. The existing metadata models also focus on a specific single media type such as image, text, or video. In addition, the metadata models differ in the complexity of the data structures they provide. With standards like EXIF [1], XMP [7], and IPTC [3] we find metadata models that provide (typed) key-value pairs to represent metadata of the image media type. Harmonization efforts like in the case of image metadata pursued by the Metadata Working Group[5] are very much appreciated. However, they remain on the same technological level and do not extend their effort beyond the single media type of image. Another metadata model like Dublin Core[6] and its extension for multimedia content[7] support hierarchical modeling of key-value pairs. It can be used to describe almost any resources. However, only entire documents and not parts of it. With MPEG-7 [2], we find a comprehensive metadata standard that aims at covering mainly decomposition and description of low-level features of audiovisual media content. MPEG-7 also provides basic means for semantic annotation. Several approaches have been published providing a formalization of MPEG-7 as an ontology, e.g., by Hunter [8] or the Core Ontology on Multimedia (COMM) [9]. However, although these ontologies provide clear semantics and an integration with Semantic Web standards, they still focus on MPEG-7 as the underlying metadata standard. As a consequence, they do not provide a generic framework for the integration of different metadata standards and metadata models. Furthermore, most metadata models also lack in supporting structured multimedia content. Structured multimedia content means that the content is organized in different discrete media assets such as images and text and continuous media assets like videos and audio. It has a coherent spatial layout, temporal course,

[5] `http://www.metadataworkinggroup.org/`
[6] `http://dublincore.org/`
[7] `http://dublincore.org/documents/dcmi-type-vocabulary/`

and some interaction with the user. Annotation of such structured multimedia content is in principle possible with MPEG-7 using separate media signals for the invidual media assets. However, actually doing it for a complex structured multimedia presentation is not very practical due to the complexity involved with this MPEG-7 annotation. In addition, various studies have shown the need in image retrieval for semantic annotation and conceptual queries [10–12].

This list of metadata models and metadata standards is very far from being complete and is beyond the scope of this work. Some overview of multimedia metadata models and standards can be found in a report [13] by the W3C Multimedia Semantics Incubator Group or in the overview[8] of the current W3C Media Annotations Working Group. The examples mentioned have been selected to show the variety of the different multimedia metadata models that exist today.

5 Multimedia Metadata Ontology

For defining our Multimedia Metadata Ontology (M3O), we leverage Semantic Web technologies and follow a pattern-oriented ontology design approach. We identified five core patterns required to express metadata for multimedia content. These patterns model the basic structural elements of existing metadata formats and conceptual models. In order to realize a specific metadata standard or metadata model in M3O, these patterns need to be specialized. The patterns base on the foundational ontology DOLCE+DnS Ultralight[9] and are formalized using Description Logics [14]. By this, we provide a clear semantics of the patterns and their elements. We achieve an improved formal representation of the metadata compared to existing models. In addition, such a generic model is not limited to a single media type such as images, video, text, and audio but provides support for structured multimedia content as it can be created with today's multimedia presentation formats such as SMIL, SVG[10], and Flash[11].

Furthermore, implementing the M3O using Semantic Web technologies is a promising approach, as it allows for representing rich metadata and multimedia semantics. Thus, it provides the infrastructure to represent both high-level semantic annotation with background knowledge as well as the annotation with low-level features extracted from the multimedia content. In addition, existing standardized multimedia presentation formats such as SMIL and SVG explicitly define the use of the Semantic Web standard RDF [15] for modeling the annotations. Semantic Web technologies ease the use of formal domain ontologies, leverage the employment of reasoning services, and provide the means to exploit the growing amount of Linked Open Data[12] available on the web.

[8] http://www.w3.org/2008/WebVideo/Annotations/drafts/ontology10/WD/mapping_table.html
[9] http://ontologydesignpatterns.org/wiki/Ontology:DOLCE+DnS_Ultralite
[10] http://www.w3.org/Graphics/SVG/
[11] http://www.adobe.com/de/products/flashplayer/
[12] http://linkeddata.org/

In the following, we introduce three basic patterns from DOLCE+DnS Ultralight that we use for our model. Subsequently, we present two patterns provided by M3O for multimedia annotation and multimedia decomposition.

5.1 DOLCE+DnS Ultralight Patterns

The Descriptions and Situation Pattern allows for the representation of contextualized views on the relations of a set of individuals and is depicted in Figure 2a. It provides a formally defined mechanism to view relations among individuals within a context, and assign roles or types that are only valid within this context.

The pattern consists of a Situation that satisfies a Description. The Description defines the roles and types present in a context, called Concepts. Each Concept classifies an Entity. The entities are the individuals that are relevant in a given context. Each Entity is connected to the situation via the hasSetting relation. Furthermore, the concepts can be related to other concepts by the isRelatedToConcept relation in order to express their dependency. The Descriptions and Situations Pattern therefore expresses an n-ary relation among a set of entities. The concepts determine the roles that the entities play within this context.

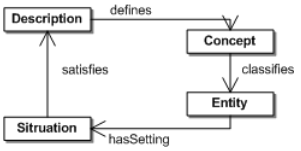
The information realization pattern in Figure 2b models the distinction between information objects and information realizations. An example is the lecture from our scenario and its realization as a SMIL presentation. The lecture would be the information object, while the SMIL presentation is the information realization. The same information can be realized in different ways. The pattern consists of the InformationRealization that is connected to the InformationObject by the realizes relation. Both are subconcepts of InformationEntity, which will make presentation of our M3O patterns easier.

With ontologies, we can use abstract concepts and clearly identifiable individuals to represent data and to perform inferencing over the data. However, at a certain point one will need to represent concrete data values, such as strings or numerical values. The Data Value Pattern (depicted in Figure 2c) assigns a concrete data value to an attribute of that entity. The attribute is represented by the concept Quality and is connected to the Entity by the hasQuality property. The Quality is connected to a Region by the hasRegion relation. The Region models the data space the value comes from. We attach the concrete value to the Region using the relation hasRegionDataValue. The data value is encoded using typed literals, i.e., the datatype can be specified using XML Schema Datatypes [16]. Using the hasPart relation, we can also express structured data values, such as present in MPEG-7.

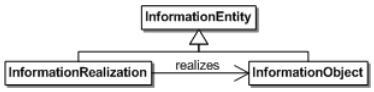
5.2 Annotation Pattern

Annotation denotes the description of some entity in terms of a note or an explanation[13]. In the context of a computer system, annotation usually refers to the description of some document stored on the computer. An example might be the

[13] Merriam-Webster Online, `http://www.merriam-webster.com`.

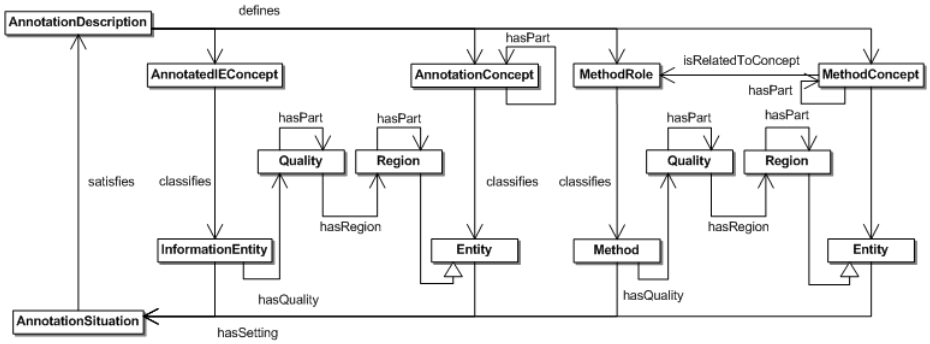


(a) Description and Situation Pattern

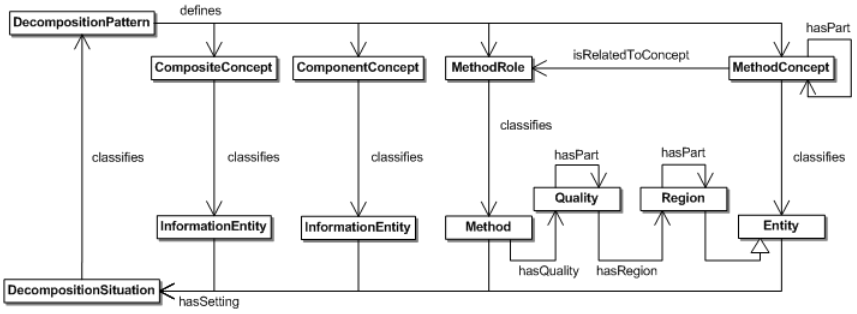


(b) Information Realization Pattern

(c) Data Value Pattern



(d) Annotation Pattern



(e) Composition Pattern

Fig. 2: Ontology Patterns of the Multimedia Metadata Ontology (M3O)

tagging of images on Flickr[14]. More generally speaking, we can define annotation as the attachment of metadata to an information entity on a computer system.

As we have discussed in Section 4, metadata comes in a various forms, such as low-level descriptors obtained by automatic methods, non-visual information covering authorship and technical details, or semantic annotation, aiming at a formal and machine-understandable representation of the contents. We identified that the underlying basic structure of annotation is always the same. Our annotation pattern models this basic structure and allows for assigning arbitrary annotations to information entities, while providing the means for modeling provenance and context.

The Annotation Pattern depicted in Figure 2d is a specialization of the Descriptions and Situations pattern and consists of an AnnotationSituation that satisfies an AnnotationDescription. The description defines at least one AnnotatedIEConcept that classifies each InformationEntity that is annotated by an instance of this pattern. The InformationEntity has as setting the AnnotationSituation. Each metadata item is represented by an Entity that is classified by an AnnotationConcept. Furthermore, we can express provenance and context information using the second part of the pattern. A Method that is classified by some MethodRole might specify how this annotation was produced. An example could be an agorithm or a manual annotation. We can further describe details, such as parameters, of the applied Method using a number of entities included in the IEAnnotationSituation that are classified by MethodConcepts, which are related to the MethodRole. In case of concrete data values for the metadata or the parameters, the Data Value Pattern is used. Please note that in the case of structured data values, also the MethodConcepts might have parts. This is expressed by the hasPart relation that classifies the parts of the Region.

5.3 Decomposition Pattern

Our Decomposition Pattern models the decomposition of information entities, e.g., the decomposition of a SMIL presentation into its logical parts or the segmentation of an image. After a decomposition, there is a whole, the composite, and there are the parts, the components. We decided to call this pattern Decomposition Pattern, since from a metadata point of view we decompose the media into parts, which we want to annotate further. Obviously, the same pattern can also be viewed as a composition of media elements and might be used like that.

The Decomposition Pattern consists of an IEDecompositionDescription that defines exactly one CompositeConcept and at least one ComponentConcept. The CompositeConcept classifies an InformationEntity, which is the whole. Each ComponentConcept classifies an InformationEntity, which are the parts. We can further specify a Method which generated the composition, and which is classified by a MethodRole. The Method can further be described by entities that are classified by Concepts, providing the means to model parameters or more abstract reasons for this decomposition. This part of the pattern is similar to the Annotation Pattern. All classified entities have the IEDecompositionSituation as setting.

[14] `http://flickr.com/`

It is important to note that in cases of structured multimedia content there is already composition information available in the media itself. A SMIL file, e.g., contains information about how single media assets are arranged. However, with M3O we aim at representing metadata about parts of the media that are not necessarily equal to or included in the physical structure defined in the SMIL file.

6 Application of M3O

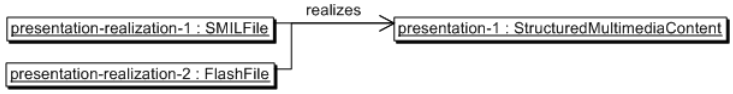
We demonstrate the application of our Multimedia Metadata Ontology at the example of the scenario in Section 2. For reasons of brevity, we present the core aspects of our model, namely the information realization, decomposition, and annotation of multimedia. Decomposition and annotation are only demonstrated on the information object level. More elaborate examples, up-to-date documentation, and discussions will be available from our wiki[15]. In the following, we use the term individual when we refer to concrete objects and the term concept when we refer to concepts of the M3O ontology. Please note that within an instantiation of a pattern only individuals appear. Additionally, we use terms like image or presentation in order to refer to the information object, and terms like image file or SMIL presentation when we refer to their realization.

We start with an example of how to apply the Information Object Pattern in order to represent the two basic levels of our model, i.e., the information object and the information realization. In this example, we consider two realizations of our presentation, namely one based on SMIL and one based on Flash.

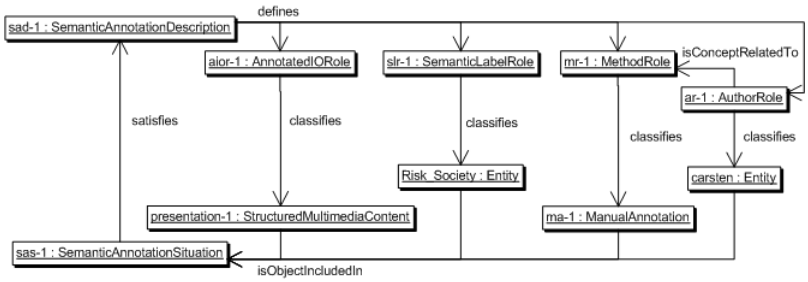
In Figure 3a, we can see that there is one individual presentation-1 of type StructuredMultimediaContent, which is a subclass of InformationObject. The files are represented by the individuals presentation-realization-1 and presentation-realization-2, which realize the presentation. They are of type SMILFile and FlashFile, which are subclasses of InformationRealization. Further information about the realization such as storage location, size, access rights, and others can be added using the annotation pattern.

In Figure 3b, the application of the Annotation Pattern is shown. The description defines four roles. The first two roles are an AnnotatedIORole and a SemanticLabelRole. The former classifies the individual presentation-1 and expresses that this is the information object being annotated. This individual is the same used in the Information Realization pattern in Figure 3a. The latter classifies the individual Risk_Society from DBpedia, which thus represents the semantic label. We exemplify the support of our patterns for context and provenance by including information about the author. The MethodRole classifies a ManualAnnotation, and thus expresses that this image was labeled manually. We specify the author of this annotation by classifying some individual carsten using the AuthorRole. The AuthorRole isConceptRelatedTo MethodeRole, expressing that carsten is the author of this manual annotation.

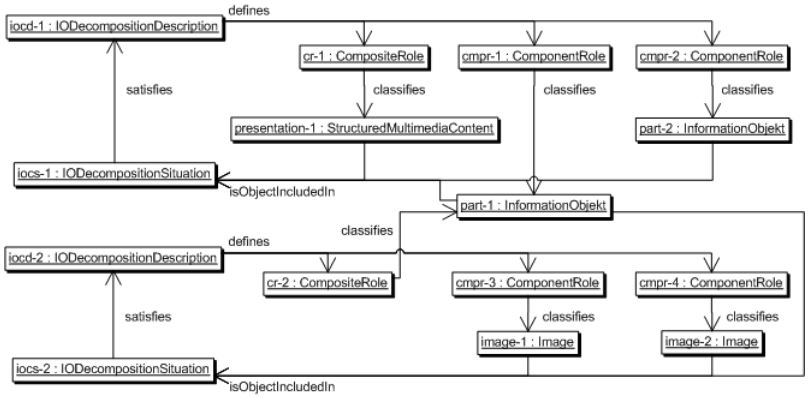
[15] http://semantic-multimedia.org/index.php/M3O:Documentation



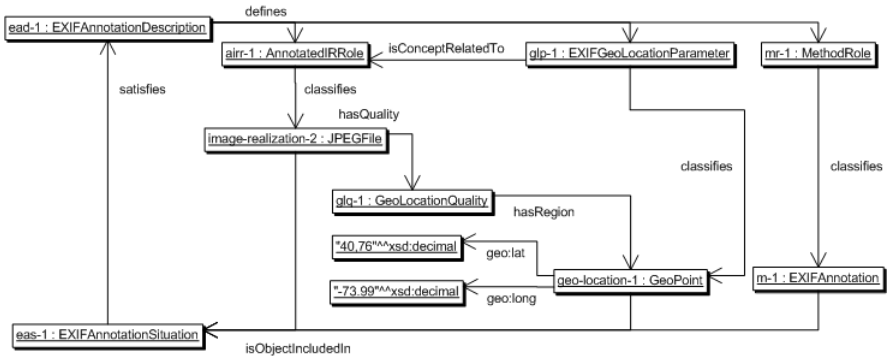
(a) An Example of Information Realization



(b) The Semantic Annotation of the Presentation



(c) A Two-Layered Decomposition



(d) Annotation with Geo-Coordinates based on EXIF

Fig. 3: Example Instantiations of our Patterns Based on the Scenario in Section 2.

Subsequently, we present the decomposition of the presentation into logical components that we want to annotate further. We can describe the decomposition both on the information object level and on the information realization level. However, in this paper we focus on the information object level. In Figure 3c we show the logical decomposition of the presentation into two parts representing the positive and negative aspects of nuclear energy, respectively. We further demonstrate the decomposition of the first part into the two images of Albert Einstein and the Times Square.

The upper part of Figure 3c shows the first composition, the lower half the second one. We see that the IODecompositionDescription defines the CompositeRole and two ComponentRoles. The CompositeRole classifies the individual presentation-1, which is again the information object representing our presentation. The ComponentRoles classify the two InformationObjects named part-1 and part-2, representing the two logical parts of the presentation. The lower half shows how the first part of the image, represented by part-1, which is further decomposed into the two images present in this part, represented by image-1 and image-2. The individual part-1 plays the ComponentRole in the first composition and the CompositeRole in the second one.

Finally, we demonstrate the annotation of an image file with EXIF metadata. Please note that we attach the EXIF descriptor to the realization image-realization-2, which represents the JPEG file realizing the image from the Times Square. The basic pattern is the same as in the example of the semantic annotation. Annotating an information entity with low-level or semantic metadata follows the same underlying structure and only the kind of metadata is different. We use an EXIFAnnotationSituation that satisfies the EXIFAnnotationDescription in order to represent that this annotation is an EXIF descriptor. The description defines a EXIFGeoParameter that parametrizes a GeoPoint, which is the Region. In order to represent the coordinates, we employ the Data Value Pattern, attaching latitude and longitude using the WGS84 vocabulary, i.e., geo:lat and geo:long [17] and use a GeoLocationQuality as the quality of the image.

7 Conclusions and Future Work

In this paper, we presented the Multimedia Metadata Ontology (M3O) that aims at capturing the structural elements of today's multimedia metadata models and metadata standards. The M3O introduces core ontology patterns for annotations and decomposition of multimedia content. It clearly distinguishes between the information object and its realization. It supports both the representation of high-level semantic annotation with background knowledge as well as the annotation with low-level features extracted from the multimedia content. With the M3O, we can better describe multimedia content and integrate the metadata provided with today's models. The current patterns presented are available in OWL at `http://m3o.semantic-multimedia.org/ontology/2009/09/16/`.

Future work is to demonstrate the general applicability and support for the different aspects of today's metadata models by providing a set of default modules covering, e.g., the well established EXIF standard and rich semantic anno-

tation. We also need to integrate further aspects of existing conceptual models [10–12]. It is also aimed at supporting new requirements that may occur in future.

Acknowledgements: We thank Frank Nack for discussing the features and concepts of the MPEG-7 metadata standard. This research has been co-funded by the EU in FP6 in the X-Media project (026978) and FP7 in the WeKnowIt project (215453).

References

1. Technical Standardization Committee on AV & IT Storage Systems and Equipment: Exchangeable image file format for digital still cameras: Exif Version 2.2. Technical Report JEITA CP-3451 (April 2002)
2. MPEG-7: Multimedia content description interface. Technical report, Standard No. ISO/IEC n15938 (2001)
3. International Press Telecommunications Council: "IPTC Core" Schema for XMP Version 1.0 Specification document (2005)
4. Gangemi, A., Presutti, V.: Ontology Design Patterns. In: Handbook on Ontologies. 2nd edn. Springer (2009)
5. Scherp, A., Jain, R.: An ecosystem for semantics. IEEE MultiMedia **16**(2) (2009) 18–25
6. Borgo, S., Masolo, C.: Foundational choices in DOLCE. In: Handbook on Ontologies. 2nd edn. Springer (2009)
7. Adobe Systems Incorporated: XMP – Adding Intelligence to Media (September 2005)
8. Hunter, J.: Enhancing the semantic interoperability of multimedia through a core ontology. IEEE Transactions on Circuits and Systems for Video Technology **13**(1) (January 2003) 49–58
9. Arndt, R., Troncy, R., Staab, S., Hardman, L., Vacura, M.: COMM: designing a well-founded multimedia ontology for the web. In: The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15, 2007. (2007) 30–43
10. Markkula, M., Sormunen, E.: End-user searching challenges indexing practices in the digital newspaper photo archive. Information Retrieval **1**(4) (January 2000) 259–285
11. Hollink, L., Schreiber, A.T., Wielinga, B.J., Worring, M.: Classification of user image descriptions. International Journal of Human-Computer Studies **61**(5) (November 2004) 601 – 626
12. Hollink, L., Schreiber, G., Wielinga, B.: Patterns of semantic relations to improve image content search. Web Semantics: Science, Services and Agents on the World Wide Web **5**(3) (2007) 195–203
13. Boll, S., Bürger, T., Celma, O., Halaschek-Wiener, C., Mannens, E., Troncy, R.: Multimedia Vocabularies on the Semantic Web. Multimedia Semantics Incubator Group Report (XGR) (July 2007)
14. Baader, F., Calvanese, D., McGuinness, D.L., Nardi, D., Patel-Schneider, P.F., eds.: The Description Logic Handbook: Theory, Implementation, and Applications. Cambridge University Press (2003)
15. Manola, F., Miller, E.: RDF Primer (February 2004)
16. Biron, P.V., Malhotra, A.: XML Schema Part 2: Datatypes Second Edition, W3C Recommendation. (October 2004)
17. Brickley, D.: Basic Geo (WGS84 lat/long) Vocabulary (2006)