Unlocking the Semantics of Multimedia Presentations in the Web with the Multimedia Metadata Ontology

Carsten Saathoff
WeST, University of Koblenz-Landau
http://west.uni-koblenz.de/
saathoff@uni-koblenz.de

Ansgar Scherp
WeST, University of Koblenz-Landau
http://west.uni-koblenz.de/
scherp@uni-koblenz.de

ABSTRACT

The semantics of rich multimedia presentations in the web such as SMIL, SVG, and Flash cannot or only to a very limited extend be understood by search engines today. This hampers the retrieval of such presentations and makes their archival and management a difficult task. Existing metadata models and metadata standards are either conceptually too narrow, focus on a specific media type only, cannot be used and combined together, or are not practically applicable for the semantic description of rich multimedia presentations.

In this paper, we propose the Multimedia Metadata Ontology (M3O) for annotating rich, structured multimedia presentations. The M3O provides a generic modeling framework for representing sophisticated multimedia metadata. It allows for integrating the features provided by the existing metadata models and metadata standards. Our approach bases on Semantic Web technologies and can be easily integrated with multimedia formats such as the W3C standards SMIL and SVG. With the M3O, we unlock the semantics of rich multimedia presentations in the web by making the semantics machine-readable and machine-understandable. The M3O is used with our SemanticMM4U framework for the multi-channel generation of semantically-rich multimedia presentations.

Categories and Subject Descriptors

E.4 [**Data**]: Coding and Information Theory; H.1.m [**Information Systems**]: Miscellaneous

General Terms

Design, Languages, Management

Keywords

rich multimedia presentations, multimedia metadata, semantic annotation

1. INTRODUCTION

Multimedia metadata and semantic annotation of multimedia is a key-enabler for improved services on multimedia content. If there is no or only limited metadata and annotations provided, the archival, retrieval, and management of multimedia content becomes very hard if not practicably infeasible. Rich, structured multimedia content is encoded by the combination of at least one continuous media asset like audio and video and one discrete media asset such as text and image [33]. The media assets are arranged in time and space into a coherent multimedia presentation such as SMIL [37], SVG [38], and Flash [1]. Annotation of such rich, structured multimedia content is the association of metadata to the structured content and its media assets [24].

In the web of today, rich, structured multimedia presentations constitute a "black box". They cannot or only to a limited extend be understood by search engines. Multimedia formats such as the W3C standards SMIL and SVG forsee the use of Semantic Web technologies for annotating the content using the Resource Description Framework (RDF) [36]. However, there is currently no appropriate model provided or best practice available that explains how to describe and annotate such rich, structured multimedia content in the web. The existing metadata models such as [17, 3, 22, 15, 16] and metadata standards like [21, 2, 19, 25] are either conceptually too narrow, semantically ambiguous, focus on a specific media type only, cannot be used and combined with each other, or are not practically applicable for the semantic description of rich multimedia presentations in the web. For example, image descriptions using EXIF [21] cannot be combined with MPEG-7 [25] descriptors. In IPTC [19], the location fields are defined to contain the locations the content is "focusing on". However, it remains unclear what "focusing on" actually means. For instance, consider an image from the atomic bombing of the city of Nagasaki in Japan in 1945. This image is about the city of 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 of the city of Nagasaki is of global importance. Distinguishing these different roles a location can play is impossible with IPTC and others. Here, support for semantic annotations with formally defined background knowledge needs to be provided. However, this is hardly found in the existing models. In addition, none of the existing models and standards explicitly support the distinction between information objects and information realizations [8]. An information object such as an image is often available in different formats and resolutions, i.e., in different information realizations. However, this feature is often requested by conceptual multimedia metadata models [20, 15]. In addition, most metadata models focus on a single media type only. Thus, they ignore the type's relation to other media types and the media assets' context within a rich, structured multimedia presentation. These metadata models are not designed to be combined with each

WWW 2010, April 26–30, 2010, Raleigh, North Carolina, USA.
ACM 978-1-60558-799-8/10/04.

other. This results in a disconnectedness of today's models. However, this combination is required by multimedia applications that integrate, e.g., image data and video data, in particular in the open world of the web. In addition, most existing models do not support representing both high-level semantic annotation with background knowledge as well as the annotation with low-level features extracted from the multimedia content.

This situation is very unfortunate as the authoring of rich, structured multimedia content can be quite expensive and providing support for annotating rich content not only improves archival, retrieval, and management of the content, but also allows for better reuse. With the Multimedia Metadata Ontology (M3O), we propose an approach for annotating rich, structured multimedia content in the web and unlocking its semantics by making it machine-readable and machine-understandable. The M3O allows for a sophisticated semantic description of rich, structured multimedia content. It provides a generic modeling framework that can accomodate and integrate the features provided by the different multimedia metadata models and metadata standards we find today. The M3O bases on Semantic Web technologies and thus can be easily integrated with today's presentation formats like SMIL and SVG. For designing the M3O, we conducted an analysis of related work and extracted the common data structures that underlie the existing metadata models and metadata standards. We represent these common data structures in form of *ontology design patterns* (ODPs) [13] based on the formal upper-level ontology DOLCE+DnS Ultralight [8]. By employing ontology design patterns and basing on a formal foundational ontology, the M3O can serve as a reference modeling framework for annotating rich, structured multimedia content.

Implementing the M3O using Semantic Web technologies is a promising approach as it allows for representing sophisticated multimedia annotations. Semantic Web technologies ease the use of formal domain ontologies, leverage the employment of reasoning services, and provide the means to exploit the rapidly growing amount of Linked Open Data (`http://linkeddata.org/`) available on the web. The distinguishing features of our Multimedia Metadata Ontology are:

- The explicit distinction of information objects and information realizations.
- Support for annotating both information objects and information realizations.
- Representing high-level annotations as well as low-level annotations.
- Support for decomposing the rich, structured multimedia presentations into its single media assets. Like annotation, this can be applied on information objects and information realizations.
- Capturing of provenance information for the annotations, decompositions, and the origins of the media assets themselves.

The M3O is agnostic to the source of the annotations and decompositions. They may be generated by automatic processes or manually created by humans.

It is important to note that we do not propose yet another model on multimedia metadata. Rather than replacing any of the existing models, the M3O aims at integrating and representing the metadata and data structures that underlie the existing approaches. This is achieved by the formal nature of the M3O and by following a pattern-oriented design approach. The problem of annotating rich, structured multimedia presentations such as SMIL, SVG, and Flash is not new. However, until today it remains an unsolved problem and there is no appropriate model or best practice available that can be used to describe and annotate structured content in the web. With the M3O, we aim at providing a model for annotating structured multimedia content and to unlock its semantics for a better archiving, retrieval, and management of the content. The M3O continues our prior work on multimedia annotation and bases on the experiences gained with developing the Core Ontology on Multimedia [3].

The remainder of the paper is organized as follows: In the next section, we motivate the need for the M3O by a concrete scenario. The requirements to the M3O are presented in Section 3 and related work is reviewed in Section 4. The ontology design patterns of the M3O are introduced in Section 5. In Section 6, we discuss the scenario from Section 2 again and demonstrate the application of M3O to the concrete annotation problems the scenario rises. In addition, we demonstrate how annotations in M3O are integrated into SMIL, SVG, and Flash. The use and implementation of the M3O in our SemanticMM4U framework for the multichannel generation of semantically-rich multimedia presentations is described in Section 7, before we conclude the paper.

2. SCENARIO

In this section, we introduce a small scenario that demonstrates the added benefit of rich semantic annotations for structured multimedia content. The scenario involves John, a nuclear physicist, who was asked by his son's school principal to give a talk about the history of nuclear energy for the schools anniversary celebration. He prepares a multimedia presentation in SMIL (cf. Figure 1) about the history of nuclear energy, but also mentions the downsides of this technology. The presentation is rendered using the RealPlayer[1].

Among others, the presentation contains two parts that discuss and visualize the positive effects of nuclear energy and the risks. The first part (cf. Figure 1a) shows a picture of Albert Einstein and a photo of the Times Square in New York. 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, which is one of the basic supplies for a high quality of living.

In the second part of our SMIL presentation (cf. Figure 1b), we replace the photo of the Times Square by a picture showing the atomic bombing of the city of Nagasaki 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. He is now used as a scientist who contributed to the invention of such a terrifying weapon. Instead of showing the advantages of nuclear en-

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

(a) (b)

Figure 1: An image of Albert Einstein [31] combined with an image of the Times Square and an image of a nuclear bomb cloud [34] expressing contrary views on *nuclear energy*.

ergy, this part of the presentation serves as metaphor for the risks and the potential destructive power of nuclear energy.

By the change of contextual use, the media assets transmit a totally different message and express different semantics [30]. John uses these two parts to discuss with the class that scientific results can often be used to do positive and negative things, independently of what the scientists originally wanted to achieve. He hopes to trigger some discussion and reflection about this topic by using one of the best known scientists in such a contradictory manner.

John usually publishes all of his writings and presentations in the web using different platforms and formats. In order to support searching for his presentation on the web, he annotates his works. Everything of potential interest about the presentation shall be annotated. For providing such a comprehensive semantic description of this multimedia presentation, there are different kinds of annotations involved: (i) John would like to annotate the different parts of the presentation individually. (ii) He wants to express that the two parts discussed above are about the positive and negative aspects of nucelar energy, respectively. (iii) In addition, he would like to annotate the individual media assets of the presentation such as the pictures of Einstein, the Times Square, and the atomic cloud with background information. For example, he likes to annotate the picture showing the atomic cloud with the historic event of the bombing of the city of Nagasaki in 1945. In addition, he may want to annotate text assets used in the presentation with their bibliography (this is not further considered in this paper for reasons of brevity). (iv) Furthermore, John would like to add metadata to the images that represent among others the place of capture or the creator. (v) John usually publishes his presentations in SMIL and Flash on the web. He wants the metadata to reflect that all the different files are realizations of the same presentation. (vi) John reused some images from Wikipedia and his personal photo collection, so he would like to point to the locations where these images can be accessed individually. (vii) Finally, he adds provenance information about himself to the presentation and the metadata such that other people can assess whether they trust the statements made in the presentation. He uses the built-in functionality of his favorite multimedia authoring tool to create these annotations using Linked Open Data vocabularies and appropriate ontologies, and publishes the presentation in SMIL and Flash including the annotations on the web.

3. REQUIREMENTS

From the scenario above, we can derive five principal requirements that need to be supported for annotating rich, structured multimedia content such as the SMIL presentation in the scenario and making both the media and its annotations available and usable on the web. These need to be reflected by our multimedia metadata ontology.

REQ-1: Identification of Resources: In the scenario, the presentation is published in different formats on different platforms (cf. (v) in Section 2). Furthermore, the presentation uses other media elements that are available separately (cf. (vi)). In order to be able to link these different resources and to coherently integrate the metadata, a universal indentification mechanism is required that allows for identifying resources on the web. Only such a mechanism guarantess that the SMIL version of Johns presentation can be linked to the Flash version of the same presentation in a way that allows to infer the fact that both files realize the same presentation.

REQ-2: Separation of Information Objects and 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 [8]. Examples of information objects are stories, stage plays, or narrative structures. Johns presentation, e.g., could be seen as a narrative structure telling the history of nuclear energy. Each information object is realized by different so-called information realizations [8]. Only a realization brings something abstract such as a message into the real world and makes it perceivable by humans (or any kind of agent). The presentation in our scenario above is, e.g., realized as a SMIL and a Flash presentation (cf. (v)), but both convey the same message. This separation between information objects and information realizations is important, since it provides a clean distinction between the semantics and the data.

REQ-3: Annotation of Information Objects and Realizations: The model needs to support the annotation of multimedia content (cf. (ii—iv)). This can be 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 like DBpedia (`http://dbpedia.org`). In our example, the picture from the Times Square would be annotated with the geocoordinates it was taken at. The first part of the presentation is annotated with some concepts that represent the positive aspects of nuclear energy. Specifically the attachment of low-level metadata such as geo coordinates, shutter time, color histograms, and others, require means to represent arbitrary, possibly complex data values. Some low-level metadata such as color information or the file size are typically attached to the realization, since they depend on the concrete realization. The realization of an image as a JPG will very likely have another file size than the realization as a PNG.

REQ-4: Decomposition of Information Objects and Realizations: Multimedia content can be decomposed into its constituent parts. The presentation above can be decomposed into, e.g., the two parts discussing the chances and risks of nuclear energy (cf. (i)). Each part can be decomposed into the images it contains. The decomposition is important to refer to only the relevant parts of the pre-

sentation when applying some annotation. In the example above, e.g., only the first part is about the positive aspects of nuclear energy and the second about the negative ones. It is required to attach the annotations to the corresponding parts. If we annotated the presentation globally, we would not be able to relate the use of Albert Einstein and the Times Square in the context of the positive aspects of nuclear energy. A global annotation would just express the fact that the presentation is about positive and negative effects and that Albert Einstein, the Times Square, and the Nagasaki Bombing are depicted. It would not provide the same semantics as if we annotated the presentation parts individually. Decomposition can be applied arbitrarily often, i.e., one can create a hierarchy of parts. A clear separation between the information object and information realization is also important for the decomposition. For example, addressing a component is depending on the realization. Physically addressing the first part is different in SMIL and in Flash, whereas this does not matter for the information object.

REQ-5: Representation of Provenance Information: Specifically on the web, provenance is of crucial importance in order to judge the reliability of information (cf. (vii)). This is also true for the metadata. One might only be interested in media discussing the risks of nuclear power created by experts in the field. When John adds the annotation about the positive and negative aspects of nuclear energy, he includes information about himself. With additional knowledge about John, e.g., coming from his FOAF [10] file, a user can judge whether he has the required expertise.

4. RELATED WORK

Numerous metadata models and metadata standards have been proposed in research and industry. These models come from different backgrounds and with different goals set. They vary in the domain for which they have been designed and can be domain-specific or for general purpose. The existing metadata models also focus on a specific single media type such as image, text, or video and are not designed for annotating rich, structured multimedia presentations such as SMIL, SVG, and Flash. In addition, the metadata models differ in the complexity of the data structures they provide. With standards like EXIF [21], XMP [2], and IPTC [19] we find metadata models that provide (typed) key-value pairs to represent metadata of the media type image. Harmonization efforts like the Metadata Working Group[2] 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. Similar limitations occur with metadata models for audio files such as ID3 [26]. Like EXIF, ID3 provides a predefined list of key-value-pairs to annotate audio files and allows for defining custom metadata fields.

Other metadata models like Dublin Core [12] support hierarchical modeling of key-value pairs. It can be used to describe arbitrary resources. However, it is designed to annotate only entire documents and not parts of it. In addition, as it is very generic it only covers a small fraction of the metadata needed to sufficiently annotate rich, structured multimedia content.

With MPEG-7 [25], 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 [11], e.g., by Hunter [17] or the Core Ontology on Multimedia (COMM) [3]. Although these ontologies provide clear semantics for the multimedia annotations, they still focus on MPEG-7 as the underlying metadata standard.

From the existing metadata standards and metadata models, only the Functional Requirements for Bibliographic Records (FRBR) [18] and COMM consider the separation of information objects and information realizations [8], i.e., the separation of an information object like an image and its multiple realizations. However, it is not fully supported in FRBR as the annotations can only be applied on the information objects. Also COMM does not fully support the separation of information objects and information realizations. The decomposition and annotation can only be applied on information object level. Thus, it is not possible to individually annotate, e.g., the different realizations of the same image information object. This is very unfortunate, as the decomposition of information realizations depends on the realization's resolution and others.

The existing metadata models and metadata formats also hardly integrate high-level and low-level features, i.e., the integration of representing both the features that can be extracted from the media assets as well as the annotation with semantic background knowledge. This is unfortunate, as studies have shown the need for semantic annotation and conceptual queries, e.g., in image retrieval [22, 15, 16]. Furthermore, the advantage of semantic annotations have been shown as well [14, 32].

Finally, most metadata models lack in supporting structured multimedia content. Annotation of such structured multimedia content is in principle possible with MPEG-7 by considering the multimedia content as a media stream that can be decomposed. However, conducting such a decomposition for a complex structured multimedia presentation is not very practical in MPEG-7 due to the nature of annotations in MPEG-7 and the complexity involved with these annotations. For example, the multimedia presentation from Section 2 rendered using the Real Player can be temporally decomposed into some time-variant streams of the first and second part of the presentation. In addition, the streams can be spatially decomposed into the left and right image. Each stream is then annotated in MPEG-7 with the appropriate metadata at the time when it is rendered to the users. This approach is not very practicable, as the annotations have to be associated to the content each time the presentation is rendered. However, as the multimedia annotations are available a-priori to the rendering of the presentations, they can already be associated with and stored together with the multimedia content beforehand. This approach is followed by today's presentation formats such as SMIL and SVG and is implemented with the M3O.

This list of metadata models and metadata standards is 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 [6] of the W3C Multimedia Semantics Incubator Group or in the overview [23] of the current W3C Media Annotations Working Group. The examples have been selected as representative to show the variety of the different metadata models and metadata formats for multimedia content that exist today.

[2] http://www.metadataworkinggroup.org/

5. MULTIMEDIA METADATA ONTOLOGY

For defining our *Multimedia Metadata Ontology (M3O)*, we leverage Semantic Web technologies and follow a pattern-oriented ontology design approach [13]. Since the goal is not to provide an ontological representation of a specific metadata standard or conceptual model, we analysed the existing standards and models (cf. Section 4). As an example, every metadata standard can assign metadata to some media item. These standards are limited and different with respect to the type of media and the kind of metadata that they support. However, common to them is that they assign *some metadata* to *some media*. Therefore, we provide a *pattern* that allows to accomplish exactly the assignement of arbitrary metadata to arbitrary media.

From our analysis, 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 such as provenance, structure, annotation, information realization, and complex data values. 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 (DUL) [8] and are formalized using Description Logics [4]. By this, we provide 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, and Flash.

The ontology is represented in OWL [35]. The annotations can therefore be represented in RDF, which can be serialized in different formats. The best known is probably RDF/XML, which is also recommended by the W3C. This allows us to directly embed RDF metadata within formats such as SMIL or SVG, which already provide appropriate means for embedding XML-based metadata. However, in general, the representation of M3O based metadata is independent of a specific serialization format like RDF/XML.

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. We conclude this section by comparing the ontology against the requirements from Section 3.

5.1 DOLCE+DnS Ultralight (DUL) Patterns

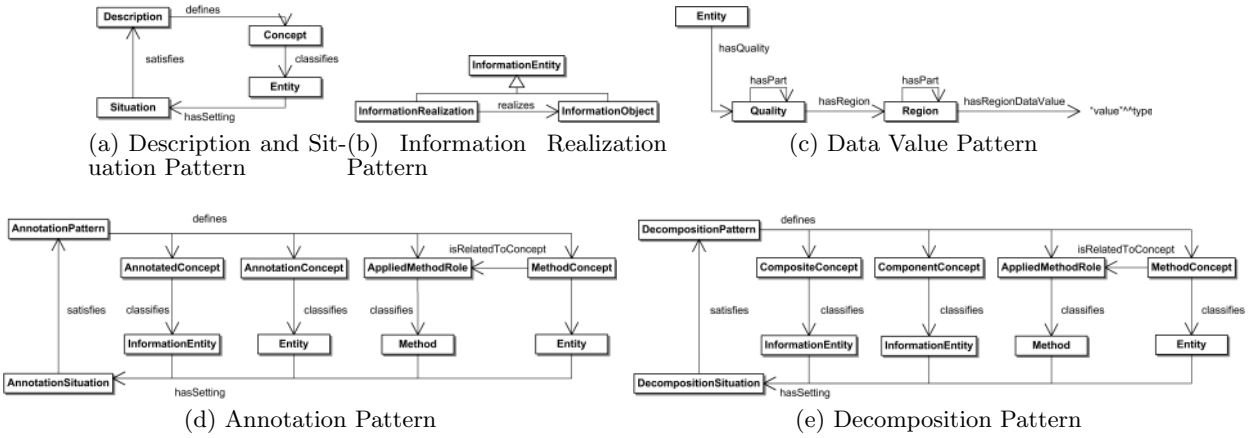
The Descriptions and Situation Pattern (D&S) [13] allows for the representation of contextualized views on the relations of a set of individuals and is depicted in Figure 2a. Since annotations might only be valid, interesting, or trusted within certain contexts, we have to consider the annotation itself within a context. This modeling requires reification, i.e., we have to be able to make statements about predicates. The D&S pattern gives us a formally sound reification mechanism. 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 D&S 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 Concept can be seen as the type of the Entity that is true only within the actual context. 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 D&S pattern therefore expresses an n-ary relation among a set of entities. The concepts determine the roles that the entities play within this context.

As an example, we consider a semantic annotation expressing a view on the relation of Albert Einstein and the development of a nuclear bomb. One might be interested in the source of this annotation in order to judge whether it is trustworthy or not. Using a simple relation would not suffice since additional information about the context in which this annotation is asserted needs to be provided. Using Descriptions and Situations, we express the fact that a relation between the media and its annotation is valid within the context of the user, i.e., that the user who provides the annotation asserts this relation. We can then restrict a search only to certain users or types of users we trust.

The information realization pattern in Figure 2b models the distinction between information objects and information realizations [8]. A digital image like the Times Square might be stored on the hard disk in several formats and resolutions. The image is the information object and each file one information realization that realizes the image. Another example is the presentation from our scenario and its realization as a SMIL file. The presentation is the information object, while the SMIL file is the information realization. The information realization pattern therefore represents the difference between information as an abstract concept and its concrete 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 allows treating information in a general sense. We will use this in our annotation and decomposition patterns, since the structure of an annotation or a decomposition is the same on both the information object and information realization levels.

With ontologies, we can use abstract concepts and clearly identifiable individuals to represent data and to perform inferencing over the data. However, we also need the means to represent concrete data values such as strings or numerical values. In DUL there exists the concept Quality in order to represent intrinsic attributes of an Entity, i.e., attributes that only exist together with the Entity. Regions are used in order to represent the values of Qualities and the data space they come from. In DUL, there are different ways to encode concrete data values using Qualities and Regions. However, this discussion is beyond the scope of this paper. With the Data Value Pattern, we propose one of these options to be used in M3O. We argue that it is important to have a single, well defined way of representing concrete data values in order to reduce the risk of ambiguities. 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 [5].



(a) Description and Situation Pattern (b) Information Realization Pattern (c) Data Value Pattern

(d) Annotation Pattern (e) Decomposition Pattern

Figure 2: Ontology Patterns of the Multimedia Metadata Ontology (M3O)

As an example, we like to represent the EXIF metadata ISOSpeed:200 of an image. We model ISOSpeed as a quality since it is an attribue of the image. As region, we use the real numbers or more abstract, the space of all possible ISO speeds. We attach the number 200 to the region using the hasRegionDataValue.

Using the hasPart relation, we can also express structured data values such as supported in MPEG-7. We can, e.g., represent the dominant color of an image as a Quality having a value from the Region RGBColorSpace. The RGBColorSpace has as part the individual color spaces, e.g., the RedColorSpace. Finally, we attach the color index to these subregions.

5.2 Annotation Pattern

Annotations are understood in M3O as the attachment of metadata to an information entity (see Section 1). Thus, annotations are metadata of information objects and information realizations, respectively. As we have discussed in Section 4, metadata comes in 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 AnnotationPattern. The description defines at least one AnnotatedConcept that classifies each InformationEntity that is annotated by an instance of this pattern. The InformationEntity has the AnnotationSituation as its setting. 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 AppliedMethodRole might specify how this annotation was produced. An example could be an algorithm or a manual annotation. We can describe further details such as parameters of the applied Method using a number of entities included in the AnnotationSituation that are classified by MethodConcepts, which are related to the MethodRole. In general, the M3O makes no assumption about the source of the annotation and both manually and automatically created annotations are supported. We also support both low-level and high-level annotations. Low-level annotations use the Data Value Pattern in order to represent metadata such as color histograms, while the high-level annotations reuses concepts and individuals from arbitrary domain ontologies.

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, called the composite, and there are the parts, called the components. We call this pattern Decomposition Pattern, as from a metadata point of view we decompose the (multi-)media into parts, which we want to annotate further.

The Decomposition Pattern (cf. Figure 1b) consists of an DecompositionPattern that defines exactly one CompositeConcept and at least one ComponentConcept. The CompositeConcept classifies an InformationEntity, expressing that it is the whole. Each ComponentConcept classifies an InformationEntity, asserting that they are parts of the whole. We can further specify the Method that created the composition, which is classified by an AppliedMethodRole. The decomposition can be automatically generated or manually created by a human. The Method can further be described by entities that are classified by MethodConcepts, providing the means to model the parameters of the Method or the general provenance of this decomposition. This part of the pattern is similar to the Annotation Pattern. All classified entities have the DecompositionSituation as setting.

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.

5.4 Summary

We have presented the five patterns underlying our Multimedia Metadata Ontology M3O. We now compare the requirements discussed in Section 3 with the provided patterns, to verify that our ontology supports all required features.

REQ-1 is supported by the use of Semantic Web technologies. In the Semantic Web resources are identified by URIs, which can be used as universal identifiers. Typically http-URIs are used. They are dereferencable and provide besides identification also an access mechanism. *REQ-2* is covered by the Information Realization Pattern, which models the distinction between the information object and it realizations. The annotation pattern addresses *REQ-3*. It provides the means to attach arbitrary metadata to both information objects and information realizations. Complex data values can be represented using the Data Value Pattern. The decomposition as formulated in *REQ-4* is provided by the Decomposition Pattern. It provides decomposition on both levels. Finally, *REQ-5* is satisfied with the use of the Descriptions and Situations pattern, which provides a formalized means for representing context and thus is applicable to represent provenance.

Besides these functional requirements, the M3O also provides a number of non-functional features [3, 29]. These non-functional features are rich axiomatization, partly due to the formal basis of DUL, modularity, extensibility, reusability, and separation of concerns. The M3O can be classified as a core ontology, which means that the ontology is modelled independently of a specific domain, but focussed on an aspect orthogonal to many domains, namely media annotation. We clearly separate concerns by using small and reusable patterns such as annotation and separation of information objects and their realizations. An important aspect is the independence of specific domain ontologies. We can incorporate arbitrary domain ontologies such as DBpedia.

6. REVISITING THE SCENARIO

Having introduced our M3O Multimedia Metadata Ontology, we now show its application to the scenario in Section 2. We first show how to apply the single patterns based on some selected examples from the scenario, and then how to embed the resulting M3O annotations into SMIL presentations based on RDF.

6.1 Modeling the Scenario with M3O

We present the core aspects of our model, namely the information realization, decomposition, and annotation of multimedia. The concrete objects are referred to as individuals, which is common in the context of ontologies and the Semantic Web. Each individual has a type that refers to a concept of some ontology. Within the diagrams, each box represents an individual and its type. For example, the presentation-realization-1:SMILFile in Figure 3a refers to an individual presentation-realization-1 of type SMILFile. Both the ontology and the concrete annotations are represented using RDF. Concepts and individuals are identified by URIs. However, for easier presentation we will omit the namespace completely.

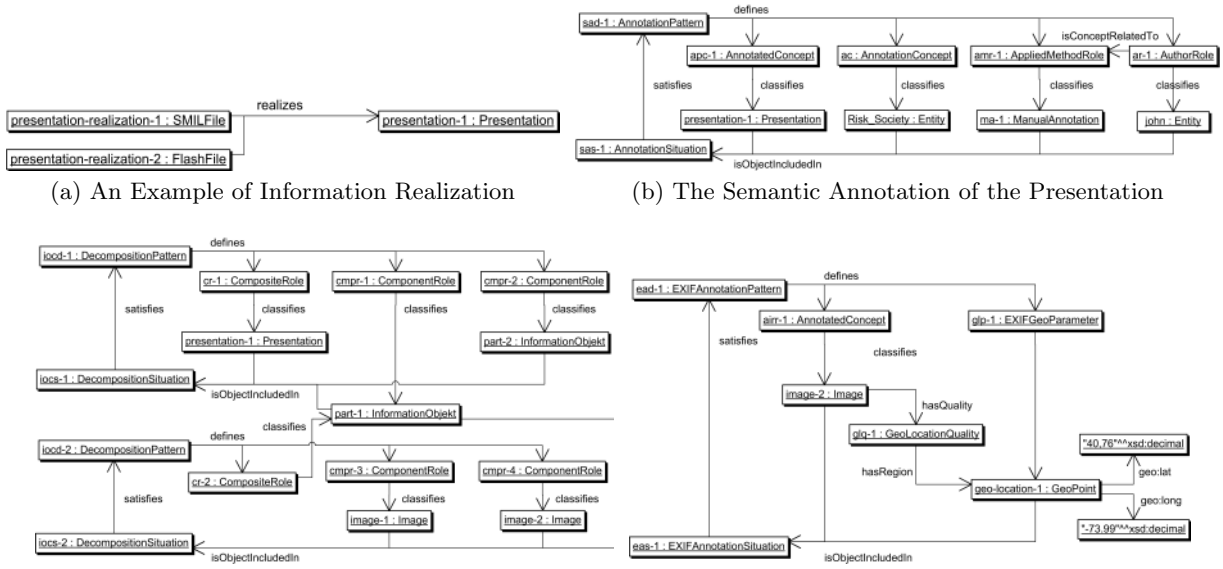
We start with an example of how to apply the Information Object Pattern in order to represent the two basic levels of our model, i.e., the information object and the information realization. In this example, we consider two realizations of our presentation, namely one in SMIL and one in Flash. Therefore, we represent the fact that the presentation is realized by a SMIL file and also by a Flash file. In Figure 3a, we can see that there is one individual presentation-1 of type Presentation, 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. Ideally, the full URI of the realization are derefereancable, i.e., a client can directly retrieve the respective realizations.

In the next step, we annotate the whole presentation with its general topic, which is in this case represented by a Wikipedia article on the risk society using the Annotation Pattern. In Figure 3b, the application of the Annotation Pattern is shown. The AnnotatedConcept classifies the individual presentation-1 and expresses that this is the information object being annotated. The AnnotationConcept classifies the individual Risk_Society from DBPedia, which represents the semantic label. We are not limited to using DBPedia, but can use any domain ontology.

The pattern shows the benefit of using a Descriptions and Situations based approach. We cannot only express the annotation as a relation between the information object and the label, but we can treat this relation within a context. We exemplify the support of our patterns for context and provenance by including information about the creator of the annotation. The AppliedMethodRole classifies a ManualAnnotation, and thus expresses that this image was labeled manually. We specify the author of this annotation by classifying some individual john using the AuthorRole. The AuthorRole isConceptRelatedTo the AppliedMethodRole, expressing that john is the author of this manual annotation. Please note that the concepts such as ManualAnnotation and AuthorRole are subconcepts of the DOLCE+DnS Ultralight concepts Method and Entity and should be provided by some specialization of the M3O core patterns.

Subsequently, we present the decomposition of the presentation into logical components that we want to annotate further. 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 DecompositionPattern defines the CompositeRole and two ComponentRoles. The CompositeRole classifies the individual presentation-1, i.e., the information object representing our presentation. This relation represents the fact that the presentation is the Composite, i.e., the *whole* in this decomposition. The ComponentRoles classify the two InformationObjects named part-1 and part-2, representing the two logical *part*s of the presentation. The lower part of Figure 3c shows how part-1 is further decomposed into the two images, represented by image-1 and image-2. Here, we see that the individual part-1 plays the ComponentRole in the first composition and the CompositeRole in the second one. Being a component or a composite is therefore depending on the context. Thus, from a modeling point of view our M3O approach is advantegous as it considers properties such as being a component or a composite only within a specific context.



(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

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

Annotating an information entity with low-level metadata follows the same underlying structure as the semantic annotation. We demonstrate this by the geo-annotation of an image file with EXIF metadata in Figure 3d. In order to represent the coordinates, we employ the Data Value Pattern. The description defines a EXIFGeoParameter that parametrizes a GeoPoint. This is a Region that represents the data space of all geo coordinates. We attach latitude and longitude using the WGS84 vocabulary, i.e., geo:lat and geo:long [9] and use a GeoLocationQuality as the quality of the image. Please note that the Region, the Quality, and the WGS84 relations are not specific to the EXIF descriptor, and could be reused in other annotations that represent geo locations.

Depending on the exact metadata it might be more appropriate to attach the information to the information object or the information realization. In this case, e.g., we want to represent the location on which the image was taken. We attach this information to the information object, since the location is independent of the format of the image, i.e., the information realization. The capturing location is therefore a property of the information object. Also, this shows that a direct one-to-one mapping from an existing metadata standard into M3O is not always appropriate. Existing standards may be ambiguous or not cleanly modelled in some aspects. Thus, a refactoring might be appropriate. However, also a one-to-one mapping is possible with M3O.

6.2 Embedding M3O in Multimedia Presentations

In this section, we describe how the metadata represented in M3O are embedded into rich, structured multimedia presentations. As our M3O annotations are represented in RDF, they can be easily serialized into XML. With the Metainformation Module, SMIL explicitly forsees the integration of XML-based metdata to describe the SMIL presentation [37] . The XML-serialized M3O is embedded into the SMIL presentation's <header> by using the <metadata>-tag (cf. lines 2-4 of Listing 1).

The embedded RDF has to represent both the information object and the information realization levels. As the example in Listing 1 shows, the embedded RDF uses the xml:base attribute to set the base URI of the RDF part to http://example.com/john/nuclear representing the information object level (cf. line 6). The SMIL file itself has the URI http://example.com/john/nuclear.smil. This URI also denotes the location from which the file can be retrieved. Within the RDF part, we use abbreviated URIs, such as #presentation-1 (eg. in line 11). Using the xml:base this is concatenated to the full URI http://example.com/john/nuclear# presentation-1. On the information realization level, we address parts of the SMIL file using the URI of the file and adding the value of the respective xml:id attribute using a hash sign (cf. line 17 for a RDF snippet referencing a SMIL element and line 29 for its definition). The URI http://example.com/john/nuclear.smil# scientificAchievements identifies the first part of the SMIL document with the xml:id scientificAchievements.

Listing 1: Embedding M3O into SMIL as RDF/XML.

```
<smil xmlns="http://www.w3.org/2006/SMIL30/...">
<head>
<!-- Metadata -->
<metadata id="meta-rdf">
<rdf:RDF
  xml:base="http://example.com/john/nuclear"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-
      syntax-ns#"
  xmlns:dul="http://www.loa-cnr.it/ontologies/DUL.
      owl#"
  xmlns:m3odec="http://m3o.semantic-multimedia.org
      /ontology/decomposition.owl#"
  xmlns:m3oann="http://m3o.semantic-multimedia.org
      /ontology/annotation.owl#"
  xmlns:m3osmil="http://m3o.semantic-multimedia.
      org/ontology/smil.owl#">

<dul:InformationObject rdf:about="#presentation-1"
    >
  <dul:isObjectIncludedIn rdf:resource="#ds-1"/>
```

```
15 </dul:InformationObject>
   <m3osmil:SMILFile rdf:about="http://example.com/
       john/nuclear.smil">
     <dul:realizes rdf:resource="#presentation-1"/>
   </m3osmil:SMILFile>
   <m3osmil:SMILElement rdf:about="http://example.com
       /john/nuclear.smil#scientificAchievements">
20   <dul:realizes rdf:resource="#part-1"/>
   </m3osmil:SMILElement>
   <!-- more, e.g. #part-1, #image-1, ... -->
   </rdf:RDF>
   </metadata>
25 <!-- layout -->
   </head>
   <body>
     <!-- Presentation content -->
     <seq id="mainPresentation">
30     <par xml:id="scientificAchievements">...</par>
       <par xml:id="scientificRisks">...</par>
     </seq>
   </body>
   </smil>
```

The decomposition pattern is shown in Listing 2. We see again the abbreviated URIs for the whole presentation #presentation-1 in line 4 as well as the URIs of the two parts in lines 9 and 14. This decomposition is on the information object level. The information objects are linked to the realizations, i.e., the SMIL file itself and the two elements of the SMIL presentation using the information realization pattern. This is shown in Listing 1 in lines 14 and 18.

Listing 2: Embedding M3O Decomposition into SMIL.

```
<m3odec:DecompositionPattern rdf:about="#dp-1">
  <dul:defines>
    <m3odec:CompositeRole rdf:about="#cr-1">
      <m3odec:classifies rdf:resource="#
          presentation-1" />
5   </m3odec:CompositeRole>
  </dul:defines>
  <dul:defines>
    <m3odec:ComponentRole rdf:about="#cmpr-1">
      <m3odec:classifies rdf:resource="#part-1" />
10  </m3odec:ComponentRole>
  </dul:defines>
  <dul:defines>
    <m3odec:ComponentRole rdf:about="#cmpr-2">
      <m3odec:classifies rdf:resource="#part-2" />
15  </m3odec:ComponentRole>
  </dul:defines>
</m3odec:DecompositionPattern>
<dul:InformationObject rdf:about="#part-1">
  <dul:isObjectIncludedIn rdf:resource="#ds-1" />
20 </dul:InformationObject>
<dul:InformationObject rdf:about="#part-2">
  <dul:isObjectIncludedIn rdf:resource="#ds-1" />
</dul:InformationObject>
<m3odec:DecompositionSituation rdf:about="#ds-1">
25 <dul:satisfies rdf:resource="#dp-1" />
</m3odec:DecompositionSituation>
```

In Listing 3, we see how to represent the fact that image-1 is realized by some file from Wikipedia. Please note that within the SMIL body even a local copy might be used. But from a metadata perspective it might be more appropriate to link to the original version on the web. Even both could be included.

Listing 3: Referring to Media Assets with M3O in SMIL.

```
<m3osmil:JPEGFile rdf:about="http://en.wikipedia.
    org/wiki/File:Einstein1921_by_F_Schmutzer_4.
    jpg">
  <dul:realizes rdf:resource="#image-1" />
</m3osmil:JPEGFile>
```

As a final example, we demonstrate the annotation of the first part with the individual Nuclear_power from DBPedia in Listing 4. As discussed, our modelling approach allows for the use of arbitrary background knowledge [29].

Listing 4: Embedding M3O Semantic Annotations into SMIL.

```
<m3odec:AnnotationConcept rdf:about="#cmpr-1">
  <m3odec:classifies rdf:resource="http://dbpedia.
      org/resource/Nuclear_power"/>
</m3odec:AnnotationConcept>
```

Embedding the M3O metadata into other presentation formats like SVG works in principle similar to the integration into SMIL. SVG also provides a <metadata>-tag that can be used to embed XML-serialized RDF in the SVG-header. As with SMIL, the individual parts of the SVG presentation such as the media assets can be annotated.

Finally, the frame-based and binary presentation format Flash does not allow for integrating metadata with the presentation [28]. However, we can use the M3O to annotate the frames in Flash. To publish the metadata an additional file or http content negotiation is required.

7. INTEGRATION IN SEMANTICMM4U

The M3O is used and currently implementeted in the SemanticMM4U framework as a generic annotation model for rich, structured multimedia presentations. The SemanticMM4U framework provides for the multi-channel generation of multimedia presentations in formats like SMIL, SVG, Flash, and others [27, 30]. The framework uses existing multimedia metadata and allows to derive new semantics while the multimedia presentations are created. It has been successfully applied in various domains including personalized sports news, context-aware tourist guides, and the generation and semantic enrichment of personal photo albums [7]. Although SemanticMM4U allows for the multi-channel generation of semantically-rich multimedia presentations, a proper model that describes how the generated presentations shall be annotated has still been desparately missing. This gap of a generic model and a reference framework for semantic annotation of rich, structured multimedia presentations is now being filled by the M3O. As the SemanticMM4U framework is available in open source, we also plan to release the M3O-extended version of the framework for the use of the community and general public.

8. CONCLUSIONS AND FUTURE WORK

In this paper, we have presented the Multimedia Metadata Ontology (M3O) as a generic modeling framework for rich, structured multimedia presentations. Unlike existing metadata models, the M3O is not bound to a specific media type and allows for integrating the features of the different models and standards we find today. The M3O strictly separates information objects from their realizations and supports annotation and decomposition of the multimedia presentations on both levels. 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. In addition, it allows to capture and represent provenance information about the annotations and decompositions.

We do not propose yet another model on multimedia metadata, but rather aim at providing a general modeling framework for multimedia metadata that comprises the features

of today's metadata models and metadata standards. Due to its formal nature and pattern-based approach it is well suited for this task and provides the basis needed to host and integrate the different existing metadata approaches. The M3O is available in OWL at `http://m3o.semantic-multimedia.org/ontology/2009/09/16/` and is formalized using Description Logics [4].

The M3O bases on the foundational ontology DOLCE+ DnS Ultralight and makes use of its rich axiomatization. Using Semantic Web technologies, the M3O is a promising approach for representing the metadata of rich multimedia presentations and unlocking their semantics for the web. As RDFa has been uptaken by Google and Yahoo! beginning of 2009 and integrated into their core business of search engines, it shows that Semantic Web technologies are of interest for the industry. Thus, we assume that an efficient gathering and processing of rich multimedia presentations described with M3O is also possible by their search engines.

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

9. REFERENCES

[1] Adobe. Flash file format, July 2008. http://www.adobe.com/licensing/developer/.

[2] Adobe Systems Incorporated. XMP – Adding Intelligence to Media, September 2005. http://www.adobe.com/products/xmp/.

[3] R. Arndt, R. Troncy, S. Staab, L. Hardman, and M. Vacura. COMM: designing a well-founded multimedia ontology for the web. In *ISWC+ASWC*, pages 30–43, 2007.

[4] F. Baader, D. Calvanese, D. L. McGuinness, D. Nardi, and P. F. Patel-Schneider, editors. *The Description Logic Handbook*. Cambridge University Press, 2003.

[5] P. V. Biron and A. Malhotra. XML Schema Part 2: Datatypes Second Edition, W3C Recommendation. October 2004. http://www.w3.org/TR/xmlschema-2/.

[6] S. Boll, T. Bürger, O. Celma, C. Halaschek-Wiener, E. Mannens, and R. Troncy. Multimedia Vocabularies on the Semantic Web. Multimedia Semantics Incubator Group Report (XGR), July 2007.

[7] S. Boll, P. Sandhaus, A. Scherp, and U. Westermann. Semantics, content, and structure of many for the creation of personal photo albums. In *ACM MULTIMEDIA*, pages 641–650, 2007.

[8] S. Borgo and C. Masolo. *Handbook on Ontologies*, chapter Foundational choices in DOLCE. Springer, 2009.

[9] D. Brickley. Basic Geo (WGS84 lat/long) Vocabulary, 2006.

[10] D. Brickley and L. Miller. The Friend Of A Friend (FOAF) vocabulary specification, November 2007. `http://xmlns.com/foaf/spec/`.

[11] S. Dasiopoulou, V. Tzouvaras, I. Kompatsiaris, and M. G. Strintzis. Enquiring MPEG-7 based multimedia ontologies. Oct. 2009.

[12] Dublin Core Metadata Initiative. DCMI Metadata Terms, Jan. 2008. http://dublincore.org/documents/dcmi-terms/.

[13] A. Gangemi and V. Presutti. *Handbook on Ontologies*, chapter Ontology Design Patterns. Springer, 2009.

[14] L. Hollink, G. Nguyen, G. Schreiber, J. Wielemaker, B. Wielinga, and M. Worring. Adding spatial semantics to image annotations. In *Knowledge Markup and Semantic Annotation*, 2004.

[15] L. Hollink, A. T. Schreiber, B. J. Wielinga, and M. Worring. Classification of user image descriptions. *International Journal of Human-Computer Studies*, 61(5):601 – 626, November 2004.

[16] L. Hollink, G. Schreiber, and B. Wielinga. Patterns of semantic relations to improve image content search. *Web Semantics: Science, Services and Agents on the World Wide Web*, 5(3):195–203, 2007.

[17] J. Hunter. Enhancing the semantic interoperability of multimedia through a core ontology. *IEEE Transactions on Circuits and Systems for Video Technology*, 13(1):49–58, January 2003.

[18] Int. Federation of Library Associations and Institutions. Functional requirements for bibliographic records. Technical report, IFLA, 2009.

[19] International Press Telecommunications Council. "IPTC Core" Schema for XMP Version 1.0 Specification document, 2005. http://www.iptc.org/.

[20] A. Jaimes and S.-F. Chang. A conceptual framework for indexing visual information at multiple levels. In *IS&T/SPIE Internet Imaging*, volume 3964, 2000.

[21] JEITA. Exchangeable image file format for digital still cameras, April 2002.

[22] M. Markkula and E. Sormunen. End-user searching challenges indexing practices in the digital newspaper photo archive. *Information Retrieval*, 1(4):259–285, January 2000.

[23] Media Annotations Working Group. Mapping table, 2008. `http://www.w3.org/2008/WebVideo/Annotations/drafts/ontology10/WD/mapping_table.html`, draft status.

[24] Merriam-Webster, Inc. Metadata, 2009. http://www.m-w.com/dictionary/metadata.

[25] MPEG-7. Multimedia content description interface. Technical report, Standard No. ISO/IEC n15938, 2001.

[26] M. Nilsson and M. Mutschler. ID3, 2009. http://www.id3.org/.

[27] A. Scherp. Canonical processes for creating personalized semantically rich multimedia presentations. *Multimedia Syst.*, 14(6):415–425, 2008.

[28] A. Scherp. Semantics support for personalized multimedia content. In *Internet and Multimedia Systems and Applications*, pages 57–65. IASTED, Mar. 2008.

[29] A. Scherp, T. Franz, C. Saathoff, and S. Staab. F—A Model of Events based on the Foundational Ontology DOLCE+ Ultralight. In *Knowledge Capturing*, 9 2009.

[30] A. Scherp and R. Jain. An ecosystem for semantics. *IEEE MultiMedia*, 16(2):18–25, 2009.

[31] F. Schmutzer. Albert Einstein, 1921. Public Domain, `http://commons.wikimedia.org/wiki/File:Einstein1921_by_F_Schmutzer_2.jpg`.

[32] G. Schreiber, I. Blok, D. Carlier, W. van Gent, J. Hokstam, and U. Roos. A mini-experiment in semantic annotation. pages 404–408, 2002.

[33] R. Steinmetz and K. Nahrstedt. *Multimedia Systems*. Springer, 2004.

[34] U.S. Federal Government. Atomic Bombing of Nagasaki, 1945. Public Domain, `http://commons.wikimedia.org/wiki/File:Nagasakibomb.jpg`.

[35] Owl web ontology language overview, January 2004.

[36] W3C. RDF Primer, Feb. 2004. http://www.w3.org/TR/REC-rdf-syntax/.

[37] W3C. SMIL 3.0, Dec. 2008. http://www.w3.org/TR/SMIL/.

[38] W3C. SVG, Apr. 2009. http://www.w3.org/TR/SVG/.