COMM: A Core Ontology for Multimedia Annotation

Richard Arndt[1], Raphaël Troncy[2], Steffen Staab[1], and Lynda Hardman*[2]

[1] ISWeb, University of Koblenz-Landau, Germany,
{rarndt|staab}@uni-koblenz.de
[2] CWI, Amsterdam, The Netherlands, {Raphael.Troncy|Lynda.Hardman}@cwi.nl

Summary. In order to retrieve and reuse non-textual media, media annotations must explain how a media object is composed of its parts and what the parts represent. Annotations need to link to background knowledge found in existing knowledge sources and to the creation and use of the media object. The representation and understanding of such facets of the media semantics is only possible through a formal language and a corresponding ontology. In this chapter, we analyze the requirements underlying the semantic representation of media objects, explain why the requirements are not fulfilled by most semantic multimedia ontologies and present COMM[3], a core ontology for multimedia, that has been built re-engineering the current defacto standard for multimedia annotation, i.e. MPEG-7, and using DOLCE as its underlying foundational ontology to support conceptual clarity and soundness as well as extensibility towards new annotation requirements.

1 Introduction

Multimedia objects are ubiquitous, whether found via web search (e.g., Google[4] or Yahoo![5] images), or via dedicated sites (e.g., Flickr[6] or YouTube[7]) or in the repositories of private users or commercial organizations (film archives, broadcasters, photo agencies, etc.). The media objects are produced and consumed by professionals and amateurs alike. Unlike textual assets, whose content can be searched for using text strings, media search is dependent on processes that have either cumbersome requirements for feature comparison (e.g. color or texture) or rely on associated, more easily processable descriptions, selecting aspects of an image or video and expressing them

[*] Lynda Hardman is also affiliated with the Technical University of Eindhoven.
[3] This chapter is a revised and extended version of [1].
[4] http://images.google.com/
[5] http://images.search.yahoo.com/
[6] http://www.flickr.com/
[7] http://www.youtube.com/

as text, or as concepts from a predefined vocabulary. Individual annotation and tagging applications have not yet achieved a degree of interoperability that enables effective sharing of semantic metadata and that links the metadata to semantic data and ontologies found on the Semantic Web.

MPEG-7 [12, 13] is an international standard that specifies how to connect descriptions to parts of a media asset. The standard includes descriptors representing low-level media-specific features that can often be automatically extracted from media types. Unfortunately, MPEG-7 is not fully suitable for describing multimedia content, because *i)* it is not open to standards that represent knowledge and make use of existing controlled vocabularies for describing the subject matter and *(ii)* its XML Schema based nature has led to design decisions that leave the annotations conceptually ambiguous and therefore prevent direct machine processing of semantic content descriptions.

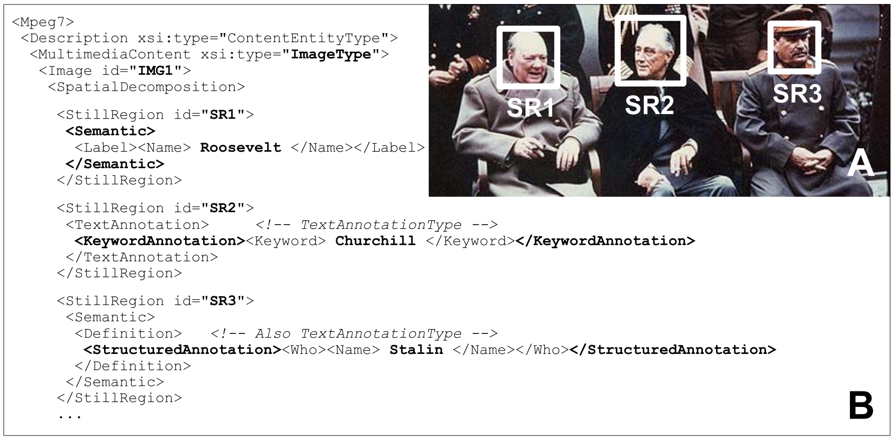
In order to avoid such problems, we advocate the use of Semantic Web languages and a core ontology for multimedia annotations, which is built based on rich ontological foundations provided by an ontology such as DOLCE (cf. Chapter 16) and sound ontology engineering principles. The result presented in this chapter is COMM, a core ontology for multimedia.

In the next section, we illustrate the main problems when using MPEG-7 for describing multimedia resources on the web. In section 3, we review existing multimedia ontologies and show why previous proposals are inadequate for semantic multimedia annotation. Subsequently, we define the requirements that a multimedia ontology should meet (section 4) before we present COMM, an MPEG-7 based ontology, and discuss our design decisions based on our requirements (section 5). In section 6, we demonstrate the use of the ontology with the scenario from section 2 and then conclude.

2 Annotating Multimedia Assets on the Web

Let us imagine that Nathalie, a student in history, wants to create a multimedia presentation of the major international conferences and summits held in the last 60 years. Her starting point is the famous “Big Three” picture, taken at the Yalta (Crimea) Conference, showing the heads of government of the United States, the United Kingdom, and the Soviet Union during World War II. Nathalie uses an MPEG-7 compliant authoring tool for detecting and labeling relevant multimedia objects automatically. On the web, she finds three different face recognition web services which provide very good results for detecting Winston Churchill, Franklin D. Roosevelt and Josef Stalin, respectively. Having these tools, she would like to run the face recognition web services on images and import the extraction results into the authoring tool in order to automatically generate links from the detected face regions to detailed textual information about Churchill, Roosevelt and Stalin (image in Fig. 1-A).

Nathalie would then like to describe a recent video from a G8 summit, such as the retrospective *A history of G8 violence* made by Reuters[8]. She uses again an MPEG-7 compliant segmentation tool for detecting the seven main sequences of this 2'26 minutes report: the various anti-capitalist protests during the Seattle (1999), Melbourne (2000), Prague (2000), Gothenburg (2001), Genoa (2001), St Petersburg (2006), Heiligendamm (2007) World Economic Forums, EU and G8 Summits. Finally, Nathalie plans to deliver her multimedia presentation in an Open Document Format (ODF) document embedding the image and video previously annotated. This scenario, however, causes several problems with existing solutions.



```
<Mpeg7>
 <Description xsi:type="ContentEntityType">
  <MultimediaContent xsi:type="ImageType">
   <Image id="IMG1">
    <SpatialDecomposition>

     <StillRegion id="SR1">
      <Semantic>
       <Label><Name> Roosevelt </Name></Label>
      </Semantic>
     </StillRegion>

     <StillRegion id="SR2">
      <TextAnnotation>     <!-- TextAnnotationType -->
       <KeywordAnnotation><Keyword> Churchill </Keyword></KeywordAnnotation>
      </TextAnnotation>
     </StillRegion>

     <StillRegion id="SR3">
      <Semantic>
       <Definition>   <!-- Also TextAnnotationType -->
        <StructuredAnnotation><Who><Name> Stalin </Name></Who></StructuredAnnotation>
       </Definition>
      </Semantic>
     </StillRegion>
     ...
```

Fig. 1. MPEG-7 annotation example (Image adapted from Wikipedia), `http://en.wikipedia.org/wiki/Yalta_Conference`

Fragment identification. Particular regions of the image need to be localized (anchor value in [6]). However, the current web architecture does not provide a means for uniquely identifying sub-parts of multimedia assets, in the same way that the fragment identifier in the URI can refer to part of an HTML or XML document. Indeed, for almost any other media type, the semantics of the fragment identifier has not been defined or is not commonly accepted. Providing an agreed upon way to localize sub-parts of multimedia objects (e.g. sub-regions of images, temporal sequences of videos or tracking moving objects in space and in time) is fundamental[9] [5]. For images, one can use either MPEG-7 or SVG snippet code to define the bounding box coordinates of specific regions. For temporal locations, one can use MPEG-7

[8] `http://www.reuters.com/news/video/summitVideo?videoId=56114`

[9] See also the forthcoming W3C Media Fragments Working Group `http://www.w3.org/2008/01/media-fragments-wg.html`.

code or the TemporalURI RFC[10]. MPEG-21 specifies a normative syntax to be used in URIs for addressing parts of any resource but whose media type is restricted to MPEG [11]. The MPEG-7 approach requires an indirection: an annotation is *about* a fragment of an XML document that *refers* to a multimedia document, whereas the MPEG-21 approach does not have this limitation [21].

Semantic annotation. MPEG-7 is a natural candidate for representing the extraction results of multimedia analysis software such as a face recognition web service. The language, standardized in 2001, specifies a rich vocabulary of multimedia descriptors, which can be represented in either XML or a binary format. While it is possible to specify very detailed annotations using these descriptors, it is not possible to guarantee that MPEG-7 metadata generated by different agents will be mutually understood due to the lack of formal semantics of this language [7, 18]. The XML code of Fig. 1-B illustrates the inherent interoperability problems of MPEG-7 : several descriptors, semantically equivalent and representing the same information while using different syntax can coexist [19]. As Nathalie used three different face recognition web services, the extraction results of the regions `SR1`, `SR2` and `SR3` differ from each other even though they are all syntactically correct. While the first service uses the MPEG-7 `SemanticType` for assigning the `<Label>` *Roosevelt* to still region `SR1`, the second one makes use of a `<KeywordAnnotation>` for attaching the keyword *Churchill* to still region `SR2`. Finally the third service uses a `<StructuredAnnotation>` (which can be used within the `SemanticType`) in order to label still region `SR3` with *Stalin*. Consequently, alternative ways for annotating the still regions render almost impossible the retrieval of the face recognition results within the authoring tool since the corresponding XPath query has to deal with these syntactic variations. As a result, the authoring tool will not link occurrences of Churchill in the images with, for example, his biography as it does not expect semantic labels of still regions as part of the `<KeywordAnnotation>` element.

Web interoperability. Nathalie would like to link the multimedia presentation to historical information about the key figures of the Yalta Conference or the various G8 summits that is already available on the web. She has also found semantic metadata about the relationships between these figures that could improve the automatic generation of the multimedia presentation. However, she realizes that MPEG-7 cannot be combined with these concepts defined in domain-specific ontologies because of its closing to the web. As this example demonstrates, although MPEG-7 provides ways of associating semantics with (parts of) non-textual media assets, it is incompatible with (semantic) web technologies and has no formal description of the semantics encapsulated implicitly in the standard.

Embedding into compound documents. Finally, Nathalie needs to compile the semantic annotations of the images, videos and textual stories into

[10] `http://www.annodex.net/TR/URI_fragments.html`

a semantically annotated compound document. However, the current state of the art does not provide a framework which allows the semantic annotation of compound documents. MPEG-7 solves only partially the problem as it is restricted to the description of audiovisual compound documents. Bearing the growing number of multimedia office documents in mind, this limitation is a serious drawback.

3 Related Work

In the field of semantic image understanding, using a multimedia ontology infrastructure is regarded to be the first step for closing the, so-called, semantic gap between low-level signal processing results and explicit semantic descriptions of the concepts depicted in images. Furthermore, multimedia ontologies have the potential to increase the interoperability of applications producing and consuming multimedia annotations. The application of multimedia reasoning techniques on top of semantic multimedia annotations is also a research topic which is currently investigated [15]. A number of drawbacks of MPEG-7 have been reported [14, 17]. As a solution, multimedia ontologies based on MPEG-7 have been proposed.

Hunter [7] provided the first attempt to model parts of MPEG-7 in RDFS, later integrated with the ABC model. Tsinaraki et al. [22] start from the core of this ontology and extend it to cover the full Multimedia Description Scheme (MDS) part of MPEG-7, in an OWL DL ontology. A complementary approach was explored by Isaac and Troncy [10], who proposed a core audio-visual ontology inspired by several terminologies such as MPEG-7, TV Anytime and ProgramGuideML. Garcia and Celma [4] produced the first complete MPEG-7 ontology, automatically generated using a generic mapping from XSD to OWL. Finally, Simou et al. [2] proposed an OWL DL Visual Descriptor Ontology[11] (VDO) based on the Visual part of MPEG-7 and used for image and video analysis.

These ontologies have been recently compared with COMM according to three criteria: *i)* the way the multimedia ontology is linked with domain semantics, *ii)* the MPEG-7 coverage of the multimedia ontology, and *iii)* the scalability and modeling rationale of the conceptualization [20]. Unlike COMM, all the other ontologies perform a one to one translation of MPEG-7 types into OWL concepts and properties. This translation does not, however, guarantee that the intended semantics of MPEG-7 is fully captured and formalized. On the contrary, the syntactic interoperability and conceptual ambiguity problems illustrated in section 2 remain. Although COMM is based on a foundational ontology, the annotations proved to be no more verbose than those in MPEG-7.

Finally, general models for annotations of non-multimedia content have been proposed by librarians. The Functional Requirements for Bibliographic

[11] `http://image.ece.ntua.gr/~gstoil/VDO`

Records (FRBR)[12] model specifies the conventions for bibliographic description of traditional books . The CIDOC Conceptual Reference Model (CRM)[13] defines the formal structure for describing the concepts and relationships used in cultural heritage documentation (cf. Chapter 19) . Hunter has described how an MPEG-7 ontology could specialize CIDOC-CRM for describing multimedia objects in museums [8]. Interoperability with such models is an issue, but interestingly, the design rationale used in these models are often comparable and complementary to foundational ontologies approach.

4 Requirements for Designing a Multimedia Ontology

Requirements for designing a multimedia ontology have been gathered and reported in the literature, e.g. in [9]. Here, we compile these and use our scenario to present a list of requirements for a web-compliant multimedia ontology.

MPEG-7 compliance. MPEG-7 is an existing international standard, used both in the signal processing and the broadcasting communities. It contains a wealth of accumulated experience that needs to be included in a web-based ontology. In addition, existing annotations in MPEG-7 should be easily expressible in our ontology.

Semantic interoperability. Annotations are only re-usable when the captured semantics can be shared among multiple systems and applications. Obtaining similar results from reasoning processes about terms in different environments can only be guaranteed if the semantics is sufficiently explicitly described. A multimedia ontology has to ensure that the intended meaning of the captured semantics can be shared among different systems.

Syntactic interoperability. Systems are only able to share the semantics of annotations if there is a means of conveying this in some agreed-upon syntax. Given that the (semantic) web is an important repository of both media assets and annotations, a semantic description of the multimedia ontology should be expressible in a web language (e.g. OWL, RDF/XML or RDFa).

Separation of concerns. Clear separation of subject matter (i.e. knowledge about depicted entities, such as the person Winston Churchill) from knowledge that is related to the administrative management or the structure and the features of multimedia documents (e.g. Churchill's face is to the left of Roosevelt's face) is required. Reusability of multimedia annotations can only be achieved if the connection between both ontologies is clearly specified by the multimedia ontology.

Modularity. A complete multimedia ontology can be, as demonstrated by MPEG-7, very large. The design of a multimedia ontology should thus be made modular, to minimize the execution overhead when used for multimedia annotation. Modularity is also a good engineering principle.

[12] `http://www.ifla.org/VII/s13/frbr/index.htm`

[13] `http://cidoc.ics.forth.gr/`

Extensibility. While we intend to construct a comprehensive multimedia ontology, as ontology development methodologies demonstrate, this can never be complete. New concepts will always need to be added to the ontology. This requires a design that can always be extended, without changing the underlying model and assumptions and without affecting legacy annotations.

5 Adding Formal Semantics to MPEG-7

MPEG-7 specifies the connection between semantic annotations and parts of media assets. We take it as a base of knowledge that needs to be expressible in our ontology. Therefore, we re-engineer MPEG-7 according to the intended semantics of the written standard. We satisfy our semantic interoperability not by aligning our ontology to the XML Schema definition of MPEG-7, but by providing a formal semantics for MPEG-7. We use a methodology based on a foundational, or top level, ontology as a basis for designing COMM (cf. Chapter 6). This provides a domain independent vocabulary that explicitly includes formal definitions of foundational categories, such as processes or physical objects, and eases the linkage of domain-specific ontologies because of the shared definitions of top level concepts. We briefly introduce our chosen foundational ontology in section 5.1, and then present our multimedia ontology, COMM, in sections 5.2 and 5.3. Finally, we discuss why our ontology satisfies all our stated requirements in section 5.4.

COMM is available at `http://multimedia.semanticweb.org/COMM/` .

5.1 DOLCE as Modeling Basis

Using the review in [16], we select the Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE) (cf. Chapter 16) as a modeling basis . Our choice is influenced by two of the main design patterns: *Descriptions & Situations* (D&S) and *Ontology of Information Objects* (OIO) [3]. The former can be used to formalize contextual knowledge, while the latter, based on D&S, implements a semiotics model of communication theory. We consider that the annotation process is a *situation* (i.e. a reified context) that needs to be described.

5.2 Multimedia Patterns

The patterns for D&S and OIO need to be extended for representing MPEG-7 concepts since they are not sufficiently specialized to the domain of multimedia annotation. This section introduces these extended multimedia design patterns, while section 5.3 details two central concepts underlying these patterns: digital data and algorithms (cf. Chapter 10). In order to define design patterns, one has to identify repetitive structures and describe them at an abstract level. The two most important functionalities provided by MPEG-7 are: the *decomposition* of media assets and the (semantic) *annotation* of their parts, which we include in our multimedia ontology.

Decomposition. MPEG-7 provides descriptors for spatial, temporal, spatio-temporal and media source decompositions of multimedia content into segments. A segment is the most general abstract concept in MPEG-7 and can refer to a region of an image, a piece of text, a temporal scene of a video or even to a moving object tracked during a period of time.

Annotation. MPEG-7 defines a very large collection of descriptors that can be used to annotate a segment. These descriptors can be low-level visual features, audio features or more abstract concepts. They allow the annotation of the content of multimedia documents or the media asset itself.

In the following, we first introduce the notion of multimedia data and then present the patterns that formalize the decomposition of multimedia content into segments, or allow the annotation of these segments. The decomposition pattern handles the structure of a multimedia document, while the media annotation pattern, the content annotation pattern and the semantic annotation pattern are useful for annotating the media, the features and the semantic content of the multimedia document respectively.

Multimedia Data.

This encapsulates the MPEG-7 notion of multimedia content and is a subconcept of digital-data[14] (introduced in more detail in section 5.3). multimedia-data is an abstract concept that has to be further specialized for concrete multimedia content types (e.g. image-data corresponds to the pixel matrix of an image). According to the OIO pattern, multimedia-data is realized by some physical media (e.g. an image). This concept is needed for annotating the physical realization of multimedia content.

Decomposition Pattern.

Following the D&S pattern, we consider that a decomposition of a multimedia-data entity is a situation[15] (a segment-decomposition) that satisfies a description, such as a segmentation-algorithm or a method (e.g. a user drawing a bounding box around a depicted face), which has been applied to perform the decomposition, see Fig. 2-B. Of particular importance are the roles that are defined by a segmentation-algorithm or a method. output-segment-roles express that some multimedia-data entities are segments of a multimedia-data entity that plays the role of an input segment (input-segment-role). These data entities have as setting a segment-decomposition situation that satisfies the roles of the applied segmentation-algorithm or method. output-segment-roles as well as segment-decompositions are then specialized according to the segment and

[14] Sans serif font indicates ontology concepts.
[15] Cf. Chapter 16.

decomposition hierarchies of MPEG-7 ([12], part 5, section 11). In terms of MPEG-7, unsegmented (complete) multimedia content also corresponds to a segment. Consequently, annotations of complete multimedia content start with a root segment. In order to designate multimedia-data instances that correspond to these root segments the decomposition pattern provides the root-segment-role concept. Note that root-segment-roles are not defined by methods which describe segment-decompositions. They are rather defined by methods which cause the production of multimedia content. These methods as well as annotation modes which allow the description of the production process (e.g. [12], part 5, section 9) are currently not covered by our ontology. Nevertheless, the prerequisite for enhancing the COMM into this direction is already given.

The decomposition pattern also reflects the need for localizing segments within the input segment of a decomposition as each output-segment-role requires a mask-role. Such a role has to be played by one or more digital-data entities which express one localization-descriptor. An example of such a descriptor is an ontological representation of the MPEG-7 `RegionLocatorType`[16] for localizing regions in an image (see Fig. 2-C). Hence, the mask-role concept corresponds to the notion of a mask in MPEG-7.

The specialization of the pattern for describing image decompositions is shown in Fig. 2-F. According to MPEG-7, an image or an image segment (image-data) can be composed into still regions. Following this modeling, the concepts output-segment-role and root-segment-role are specialized by the concepts still-region-role and root-still-region-role respectively. Note, that root-still-region-role is a subconcept of still-region-role *and* root-segment-role. The MPEG-7 decomposition mode which can be applied to still regions is called `StillRegionSpatialDecompositionType`. Consequently, the concept still-region-spatial-decomposition is added as a subconcept of segment-decomposition. Finally, the mask-role concept is specialized by the concept spatial-mask-role.

Analogously, the pattern can be used to describe the decomposition of a video asset or of an ODF document (see Fig. 3).

Content Annotation Pattern.

This formalizes the attachment of metadata (i.e. annotations) to multimedia-data (Fig. 2-D). Using the D&S pattern, annotations also become situations that represent the state of affairs of all related digital-data (metadata and annotated multimedia-data). digital-data entities represent the attached metadata by playing an annotation-role. These roles are defined by methods or algorithms. The former are used to express manual (or semi-automatic) annotation while the latter serve as an explanation for the attachment of automatically computed features, such as the dominant colors of a still region. It is mandatory that the multimedia-data entity being annotated plays an annotated-data-role.

[16] Italic type writer font indicates MPEG-7 language descriptors.

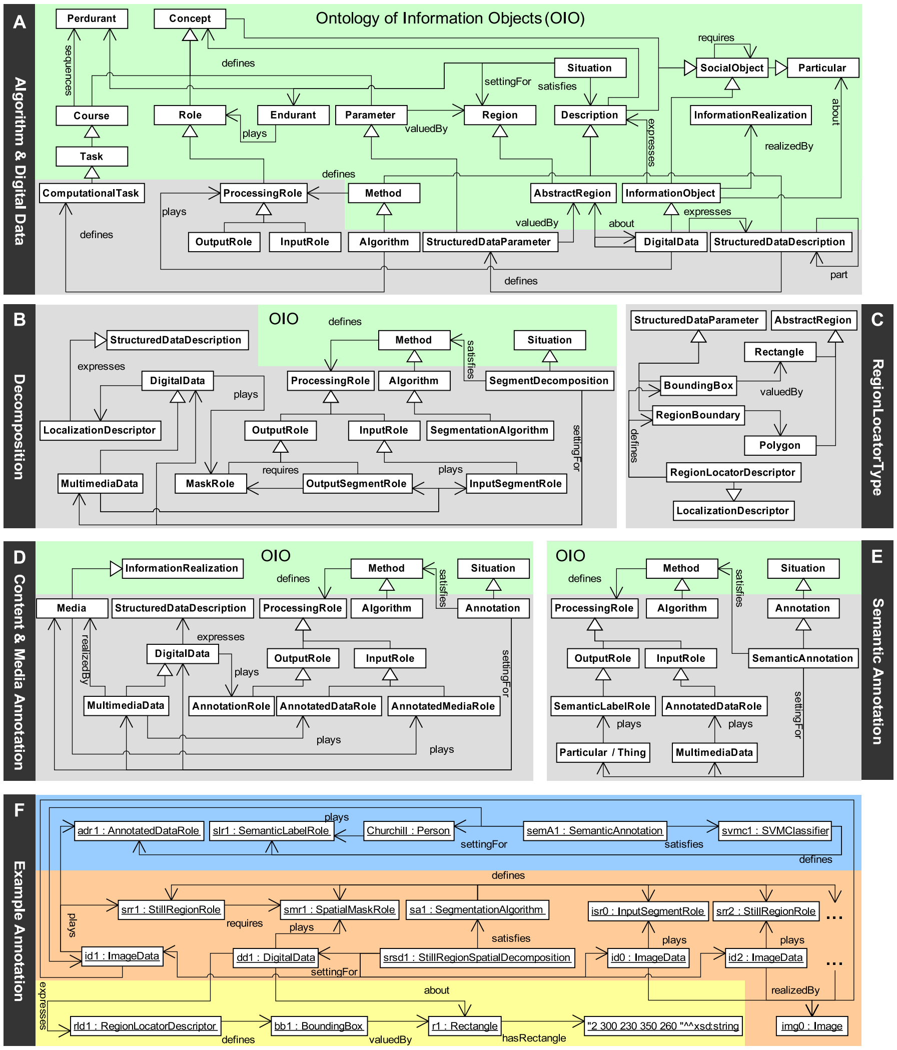


Fig. 2. COMM: Design patterns in UML notation: Basic design patterns (A), multimedia patterns (B, D, E) and modeling examples (C, F).

The actual metadata that is carried by a digital-data entity depends on the structured-data-description that is expressed by it. These descriptions are formalized using the digital data pattern (see section 5.3). Applying the content annotation pattern for formalizing a specific annotation, e.g. a dominant-color-annotation which corresponds to the connection of a MPEG-7 `DominantColorType` with a segment, requires only the specialization of the concept annotation, e.g. dominant-color-annotation. This concept is defined by being a setting for a digital-data entity that expresses one dominant-color-descriptor (a subconcept of structured-data-description which corresponds to the DominantColorType).

Media Annotation Pattern.

This forms the basis for describing the physical instances of multimedia content (Fig. 2-D). It differs from the content annotation pattern in only one respect: it is the media that is being annotated and therefore plays an annotated-media-role.

One can thus represent that some visual content (e.g. the picture of a digital camera) is realized by a JPEG image with a size of 462848 bytes, using the MPEG-7 MediaFormatType. Using the media annotation pattern, the metadata is attached by connecting a digital-data entity with the image. The digital-data plays an annotation-role while the image plays an annotated-media-role. An ontological representation of the MediaFormatType, namely an instance of the structured-data-description subconcept media-format-descriptor, is expressed by the digital-data entity. The tuple formed with the scalar "462848" and the string "JPEG" is the value of the two instances of the concepts file-size and file-format respectively. Both concepts are subconcepts of structured-data-parameter.

Semantic Annotation Pattern.

Even though MPEG-7 provides some general concepts (see [12], part 5, section 12) that can be used to describe the perceivable content of a multimedia segment, independent development of domain-specific ontologies is more appropriate for describing possible interpretations of multimedia—it is useful to create an ontology specific to multimedia, it is not useful to try to model the real world within this. An ontology-based multimedia annotation framework should rely on domain-specific ontologies for the representation of the real world entities that might be depicted in multimedia content. Consequently, this pattern specializes the content annotation pattern to allow the connection of multimedia descriptions with domain descriptions provided by independent world ontologies (Fig. 2-E).

An OWL Thing or a DOLCE particular (belonging to a domain-specific ontology) that is depicted by some multimedia content is not directly connected to it but rather through the way the annotation is obtained. Actually, a manual annotation method or its subconcept algorithm, such as a classification algorithm, has to be applied to determine this connection. It is embodied through a semantic-annotation that satisfies the applied method. This description specifies that the annotated multimedia-data has to play an annotated-data-role and the depicted Thing/ particular has to play a semantic-label-role. The pattern also allows the integration of features which might be evaluated in the context of a classification algorithm. In that case, digital-data entities that represent these features would play an input-role.

5.3 Basic Patterns

Specializing the D&S and OIO patterns for defining multimedia design patterns is enabled through the definition of basic design patterns, which formalize the notion of digital data and algorithm.

Digital Data Pattern.

Within the domain of multimedia annotation, the notion of digital data is central — both the multimedia content being annotated and the annotations themselves are expressed as digital data. We consider digital-data entities of arbitrary size to be information-objects, which are used for communication between machines. The OIO design pattern states that descriptions are expressed by information-objects, which have to be about facts (represented by particulars). These facts are settings for situations that have to satisfy the descriptions that are expressed by information-objects. This chain of constraints allows the modeling of complex data structures to store digital information. Our approach is as follows (see Fig. 2-A): digital-data entities express descriptions, namely structured-data-descriptions, which define meaningful labels for the information contained by digital-data. This information is represented by numerical entities such as scalars, matrices, strings, rectangles or polygons. In DOLCE terms, these entities are abstract-regions. In the context of a description, these regions are described by parameters. structured-data-descriptions thus define structured-data-parameters, for which abstract-regions carried by digital-data entities assign values.

The digital data pattern can be used to formalize complex MPEG-7 low-level descriptors. Fig. 2-C shows the application of this pattern by formalizing the MPEG-7 RegionLocatorType, which mainly consists of two elements: a Box and a Polygon. The concept region-locator-descriptor corresponds to the RegionLocatorType. The element Box is represented by the structured-data-parameter subconcept BoundingBox while the element Polygon is represented by the region-boundary concept.

The MPEG-7 code example given in Fig. 1 highlights that the formalization of data structures, so far, is not sufficient — complex MPEG-7 types can include nested types that again have to be represented by structured-data-descriptions. In our example, the MPEG-7 SemanticType contains the element Definition which is of complex type TextAnnotationType. The digital data pattern covers such cases by allowing a digital-data instance dd1 to be about a digital-data instance dd2 which expresses a structured-data-description that corresponds to a nested type (see Fig. 2-A). In this case the structured-data-description of instance dd2 would be a part of the one expressed by dd1.

Algorithm Pattern.

The production of multimedia annotation can involve the execution of algorithms or the application of computer assisted methods which are used to produce or manipulate digital-data. The recognition of a face in an image region is an example of the former, while manual annotation of the characters is an example of the latter.

We consider algorithms to be methods that are applied to solve a computational problem (see Fig. 2-A). The associated (DOLCE) situations represent the work that is being done by algorithms. Such a situation encompasses digital-data[17] involved in the computation, regions that represent the values of parameters of an algorithm, and perdurants[18] that act as computational-tasks (i.e. the processing steps of an algorithm). An algorithm defines roles which are played by digital-data. These roles encode the meaning of data. In order to solve a problem, an algorithm has to process input data and return some output data. Thus, every algorithm defines at least one input-role and one output-role which both have to be played by digital-data.

5.4 Comparison with Requirements

We discuss now whether the requirements stated in section 4 are satisfied with our proposed modeling of the multimedia ontology.

The ontology is **MPEG-7 compliant** since the patterns have been designed with the aim of translating the standard into DOLCE. It covers the most important part of MPEG-7 that is commonly used for describing the structure and the content of multimedia documents. Our current investigation shows that parts of MPEG-7 that have not yet been considered (e.g. navigation & access) can be formalized analogously to the other descriptors through the definition of further patterns. The technical realization of the basic MPEG-7 data types (e.g. matrices and vectors) is not within the scope of the multimedia ontology. They are represented as ontological concepts, because the about relationship which connects digital-data with numerical entities is only defined between concepts. Thus, the definition of OWL data type properties is required to connect instances of data type concepts (subconcepts of the DOLCE abstract-region) with the actual numeric information (e.g. xsd:string). Currently, simple string representation formats are used for serializing data type concepts (e.g. rectangle) that are currently not covered by W3C standards. Future work includes the integration of the extended data types of OWL 1.1.

Syntactic and semantic interoperability of our multimedia ontology is achieved by an OWL DL formalization[19]. Similar to DOLCE, we provide a rich axiomatization of each pattern using first order logic. Our ontology can be linked to any web-based domain-specific ontology through the semantic annotation pattern.

[17] digital-data entities are DOLCE endurants, i.e. entities which exist in time and space.

[18] Events, processes or phenomena are examples of perdurants. endurants participate in perdurants.

[19] Examples of the axiomatization are available on the COMM website.

A clear **separation of concerns** is ensured through the use of the multimedia patterns: the decomposition pattern for handling the structure and the annotation pattern for dealing with the metadata.

These patterns form the core of the **modular** architecture of the multimedia ontology. We follow the various MPEG-7 parts and organize the multimedia ontology into modules which cover *i)* the descriptors related to a specific media type (e.g. visual, audio or text) and *ii)* the descriptors that are generic to a particular media (e.g. media descriptors). We also design a separate module for data types in order to abstract from their technical realization.

Through the use of multimedia design patterns, our ontology is also **extensible**, allowing the inclusion of further media types and descriptors (e.g. new low-level features) using the same patterns. As our patterns are grounded in the D&S pattern, it is straightforward to include further contextual knowledge (e.g. about provenance) by adding roles or parameters. Such extensions will not change the patterns, so that legacy annotations will remain valid.

6 Expressing the Scenario in COMM

The interoperability problem with which Nathalie was faced in section 2 can be solved by employing the COMM ontology for representing the metadata of all relevant multimedia objects and the presentation itself throughout the whole creation workflow. The student is shielded from details of the multimedia ontology by embedding it in authoring tools and feature analysis web services.

The application of the Winston Churchill face recognizer results in an annotation RDF graph that is depicted in the upper part of Fig. 3 (visualized by an UML object diagram[20]). The decomposition of Fig. 1-A, whose content is represented by id0, into one still region (the bounding box of Churchill's face) is represented by the lighter middle part of the UML diagram. The segment is represented by the image-data instance id1 which plays the still-region-role srr1. It is located by the digital-data instance dd1 which expresses the region-locator-descriptor rld1 (lower part of the diagram). Using the semantic annotation pattern, the face recognizer can annotate the still region by connecting it with the URI `http://en.wikipedia.org/wiki/Winston_Churchill`. An instance of an arbitrary domain ontology concept could also have been used for identifying the resource.

Running the two remaining face recognizers for Roosevelt and Stalin will extend the decomposition further by two still regions, i.e. the image-data instances id2 and id3 as well as the corresponding still-region-roles, spatial-mask-roles and digital-data instances expressing two more region-locator-descriptors (indicated at the right border of Fig. 3). The domain ontologies which provide the instances Roosevelt and Stalin for annotating id2 and id3 with the semantic annotation pattern do not have to be identical to the one that contains

[20] The scheme used in Fig. 3 is instance:Concept, the usual UML notation.

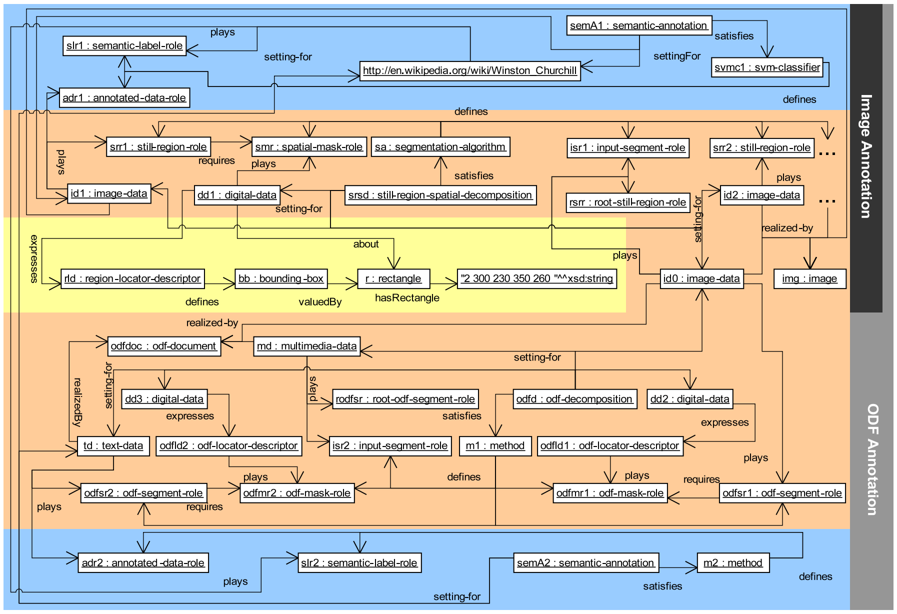


Fig. 3. Annotation of one segment of the Yalta picture and its embedding into an ODF document which contains a text segment that is also about Winston Churchill.

Churchill. If several domain ontologies are used, Nathalie can use the OWL sameAs and equivalentClass constructs to align the three face recognition results to the domain ontology that is best suited for enhancing the automatic generation of the multimedia presentation.

Decomposition of ODF documents is formalized analogously to image segmentation (see Fig. 2-F). Therefore, embedding the image annotation into an ODF document annotation is straightforward. The lower part of Fig. 3 shows the decomposition of a compound ODF document into textual and image content. This decomposition description could result from copying an image from the desktop and pasting it into an ODF editor such as OpenOffice. A plugin of this program could produce COMM metadata of the document in the background while it is produced by the user. The media independent design patterns of COMM allow the implementation of a generic mechanism for inserting metadata of arbitrary media assets into already existing metadata of an ODF document. In the case of Fig. 3, the instance id0 (which represents the whole content of the Yalta image) needs to be connected with three instances of the ODF annotation: *i)* the odf-decomposition instance odfd which is a setting-for all top level segments of the odf-document, *ii)* the odf-segment-role instance odfsr1 which identifies id0 as a part of the whole ODF content md (a multimedia-data instance), *iii)* the instance odfdoc as the image now is also realized-by the odf-document.

Fig. 3 also demonstrates how a domain ontology[21] can be used to define semantically meaningful relations between arbitrary segments. The textual content td as well as the image segment id1 are about Winston Churchill. Consequently, the URI `http://en.wikipedia.org/wiki/Winston_Churchill` is used for annotating both instances using the media independent semantic annotation pattern.

The two segments td and id1 are located within md by two digital-data instances (dd2 and dd3) which express two corresponding odf-locator-descriptor instances. The complete instantiations of the two odf-locator-descriptors are not shown in Fig. 3. The modeling of the region-locator-descriptor, which is completely instantiated in Fig. 3, is shown in Fig. 2-C. The technical details of the odf-locator-descriptor are not presented. However, it is possible to locate segments in ODF documents by storing an XPath which points to the beginning and the end of an ODF segment. Thus, the modeling of the odf-locator-descriptor can be carried out analogously to the region-locator-descriptor.

In order to ease the creation of multimedia annotations with our ontology, we have developed a Java API[22] which provides an MPEG-7 class interface for the construction of meta-data at runtime . Annotations which are generated in memory can be exported to Java based RDF triple stores such as Sesame. For that purpose, the API translates the objects of the MPEG-7 classes into instances of the COMM concepts. The API also facilitates the implementation of multimedia retrieval tools as it is capable of loading RDF annotation graphs (e.g. the complete annotation of an image including the annotation of arbitrary regions) from a store and converting them back to the MPEG-7 class interface. Using this API, the face recognition web service will automatically create the annotation which is depicted in the upper part of Fig. 3 by executing the following code:

```
Image img0 = new Image();
StillRegion isr0 = new StillRegion();
img0.setImage(isr0);
StillRegionSpatialDecomposition srsd1 = new StillRegionSpatialDecomposition();
isr0.addSpatialDecomposition(srsd1);
srsd1.setDescription(new SegmentationAlgorithm());
StillRegion srr1 = new StillRegion();
srsd1.addStillRegion(srr1);
SpatialMask smr1 = new SpatialMask();
srr1.setSpatialMask(smr1);
RegionLocatorDescriptor rld1 = new RegionLocatorDescriptor();
smr1.addSubRegion(rld1);
rld1.setBox(new Rectangle(300, 230, 50, 30));
Semantic s1 = new Semantic();
```

[21] In this example, the domain ontology corresponds to a collection of wikipedia URI's.

[22] The Java API is available at `http://multimedia.semanticweb.org/COMM/api/`.

```
s1.addLabel("http://en.wikipedia.org/wiki/Winston_Churchill");
s1.setDescription(new SVMClassifier());
srr1.addSemantic(s1);
```

7 Conclusion and Future Work

We have presented COMM, an MPEG-7 based multimedia ontology, well-founded and composed of multimedia patterns. It satisfies the requirements, as they are described by the multimedia community itself, for a multimedia ontology framework. The ontology is completely formalized in OWL DL and a stable version is available with its API at: http://multimedia.semanticweb.org/COMM/. It has been used in projects such as K-Space and X-Media.

The ontology already covers the main parts of the standard, and we are confident that the remaining parts can be covered by following our method for extracting more design patterns. Our modeling approach confirms that the ontology offers even more possibilities for multimedia annotation than MPEG-7 since it is interoperable with existing web ontologies. The explicit representation of algorithms in the multimedia patterns describes the multimedia analysis steps, something that is not possible in MPEG-7. The need for providing this kind of annotation is demonstrated in the algorithm use case of the W3C Multimedia Semantics Incubator Group[23]. The intensive use of the D&S reification mechanism causes that RDF annotation graphs, which are generated according to our ontology, are quite large compared to the ones of more straightforwardly designed multimedia ontologies. This presents a challenge for current RDF and OWL stores, but we think it is a challenge worth deep consideration as it is utterly necessary to overcome the isolation of current multimedia annotations and to achieve full interoperability for (nearly) arbitrary multimedia tools and applications.

Acknowledgments

The research leading to this paper was partially supported by the European Commission under contract FP6-027026, Knowledge Space of semantic inference for automatic annotation and retrieval of multimedia content – K-Space, and under contract FP6-26978, X-Media, large scale knowledge sharing and reuse across media.

References

1. Richard Arndt, Raphaël Troncy, Steffen Staab, Lynda Hardman, and Miroslav Vacura. COMM: Designing a Well-Founded Multimedia Ontology for the Web. In 6^{th} *Int. Semantic Web Conference*, 2007.

[23] http://www.w3.org/2005/Incubator/mmsem/XGR-interoperability/

2. S. Blöhdorn, K. Petridis, C. Saathoff, N. Simou, V. Tzouvaras, Y. Avrithis, S. Handschuh, Y. Kompatsiaris, S. Staab, and M. Strintzis. Semantic Annotation of Images and Videos for Multimedia Analysis. In *2^{nd} European Semantic Web Conference*, 2005.
3. Aldo Gangemi, Stefano Borgo, Carola Catenacci, and Jos Lehmann. Task Taxonomies for Knowledge Content. Technical report, Metokis Deliverable 7, 2004.
4. Roberto Garcia and Oscar Celma. Semantic Integration and Retrieval of Multimedia Metadata. In *5^{th} International Workshop on Knowledge Markup and Semantic Annotation*, 2005.
5. Joost Geurts, Jacco van Ossenbruggen, and Lynda Hardman. Requirements for practical multimedia annotation. In *Workshop on Multimedia and the Semantic Web*, 2005.
6. F. Halasz and M. Schwartz. The Dexter Hypertext Reference Model. *Communications of the ACM*, 37(2):30–39, 1994.
7. Jane Hunter. Adding Multimedia to the Semantic Web - Building an MPEG-7 Ontology. In *1^{st} International Semantic Web Working Symposium*, pages 261–281, 2001.
8. Jane Hunter. Combining the CIDOC/CRM and MPEG-7 to Describe Multimedia in Museums. In *6^{th} Museums and the Web Conference*, 2002. `http://www.archimuse.com/mw2002/papers/hunter/hunter.html`.
9. Jane Hunter and Liz Armstrong. A Comparison of Schemas for Video Metadata Representation. In *8^{th} International World Wide Web Conference*, pages 1431–1451, 1999.
10. Antoine Isaac and Raphaël Troncy. Designing and Using an Audio-Visual Description Core Ontology. In *Workshop on Core Ontologies in Ontology Engineering*, 2004.
11. MPEG-21. Part 17: Fragment Identification of MPEG Resources. Standard No. ISO/IEC 21000-17, 2006.
12. MPEG-7. Multimedia Content Description Interface. Standard No. ISO/IEC 15938, 2001.
13. Frank Nack and Adam T. Lindsay. Everything you wanted to know about MPEG-7 (Parts I & II). *IEEE Multimedia*, 6(3-4), 1999.
14. Frank Nack, Jacco van Ossenbruggen, and Lynda Hardman. That Obscure Object of Desire: Multimedia Metadata on the Web (Part II). *IEEE Multimedia*, 12(1), 2005.
15. Bernd Neumann and Ralf Möller. *Cognitive Vision Systems*, chapter On Scene Interpretation with Description Logics, pages 247–275. Springer, 2006.
16. Daniel Oberle, S. Lamparter, S. Grimm, D. Vrandecic, Steffen Staab, and Aldo Gangemi. Towards Ontologies for Formalizing Modularization and Communication in Large Software Systems. *Journal of Applied Ontology*, 1(2):163–202, 2006.
17. Jacco van Ossenbruggen, Frank Nack, and Lynda Hardman. That Obscure Object of Desire: Multimedia Metadata on the Web (Part I). *IEEE Multimedia*, 11(4), 2004.
18. Raphaël Troncy. Integrating Structure and Semantics into Audio-visual Documents. In *2^{nd} International Semantic Web Conference*, pages 566–581, 2003.
19. Raphaël Troncy, Werner Bailer, Michael Hausenblas, Philip Hofmair, and Rudolf Schlatte. Enabling Multimedia Metadata Interoperability by Defining Formal Semantics of MPEG-7 Profiles. In *1^{st} International Conference on Semantics And digital Media Technology*, pages 41–55, 2006.

20. Raphaël Troncy, Óscar Celma, Suzanne Little, Roberto García, and Chrisa Tsinaraki. MPEG-7 based Multimedia Ontologies: Interoperability Support or Interoperability Issue? In *1st International Workshop on Multimedia Annotation and Retrieval enabled by Shared Ontologies*, 2007.
21. Raphaël Troncy, Lynda Hardman, Jacco van Ossenbruggen, and Michael Hausenblas. Identifying Spatial and Temporal Media Fragments on the Web. In *W3C Video on the Web Workshop*, 2007. `http://www.w3.org/2007/08/video/positions/Troncy.pdf`.
22. Chrisa Tsinaraki, Panagiotis Polydoros, Nektarios Moumoutzis, and Stavros Christodoulakis. Integration of OWL ontologies in MPEG-7 and TV-Anytime compliant Semantic Indexing. In *16th International Conference on Advanced Information Systemes Engineering*, 2004.

Index

Author Index