Towards a simplification of COMM-based multimedia annotations

Miroslav VACURA a, Raphaël TRONCY b

a University of Economics, W. Churchill Sq.4, 13067, Prague 3, Czech Republic
b CWI, Kruislaan 413, 1098 SJ Amsterdam, The Netherlands

Abstract. Semantic descriptions of multimedia content are a necessary condition for their effective retrieval and presentation to an end-user. COMM is a core ontology for multimedia annotations, expressed in OWL, that provides a comprehensive model for generating fine-grained descriptions of multimedia assets. Similar to other multimedia ontology models, COMM annotations result in complex RDF graphs that can make the search for multimedia content inefficient and its selection to an end-user inappropriate. On the other hand, the richness of the model allows the description of multiple points of view of an asset, distinguishing, for example, the real world objects from their digital representation. We discuss the various advantages of having a rich formal model for expressing semantic annotations of multimedia content. We then show how the model can be simplified and optimized using logical rules and SPARQL queries and we illustrate the usefulness of this approach in an exploratory environment for searching and browsing multimedia content.

Keywords. COMM, multimedia ontology, multimedia annotations, D&S, OIO, end-user interfaces

1. Introduction

Working with multimedia assets involves their capture, annotation, editing, authoring and/or transfer to other applications for publication and distribution. There is substantial support within the multimedia research community for the collection of machine-processable semantics during established media workflow practices. An essential aspect of these approaches is that a media asset gains value by the inclusion of information (i.e. metadata) about how or when it is created or used, what it represents, and how it is manipulated and organized.

For example, let us imagine that a photographer take a set of photos during an event. S/He wants to publish low resolution versions of these images with appropriate semantic metadata in order to maximise the chance that these photos are found and sold in high resolution to magazines. S/He thus adds the free text description “French midfielder Zinedine Zidane converts a penalty kick for the game’s first score during the semi-final World Cup football match between Portugal and France at Munich’s World Cup Stadium, 05 July 2006.” to the image depicted in Figure 1. Using keywords and terms from controlled vocabularies is already common practice in the image industry. Hence, instead of just tagging the photo, the photographer adds semantic metadata, expressed in RDF, for describing the place and date of the event, the scene depicted, and properties regard-
ing the image data such as its resolution. This task can be partially supported with natural language processing tools, such as OpenCalais\(^1\), that extract named entities from the caption. PhotoRDF\(^2\) provides a minimal RDF schema combining EXIF for representing the camera settings and Dublin Core for the general scenery description. PhotoStuff\(^3\) is a more expressive tool that allows the photographer to fully describe the scene using domain specific ontologies and region-based annotations. These models, however, are specific to the image media, and cannot abstract the multimedia data from one particular representation (i.e. the algorithm and encoding parameters of the content).

The photographer would like to propagate the semantic annotations to all manifestations of the images (thumbnail, full size, etc.) and link them to other descriptions of the event. Multimedia ontologies, such as those based on MPEG-7, provide the richness for such descriptions [6]. They generate, however, complex RDF graphs that are inefficient for search and inappropriate for presentation to an end-user. In this paper, we investigate how to simplify these models while using all their expressivity. First, we show how multimedia content can be described using COMM [1], a core ontology for multimedia (Section 2). Second, we justify the complexity of the resulting annotations showing the added expressivity gained (Section 3). We then present how to simplify these annotations using SPARQL queries and we illustrate the usefulness of this approach in an exploratory environment for searching and browsing multimedia content (Section 4). Finally, we conclude and present future work in Section 5.

2. Using COMM for Describing Multimedia Content

The scenario presented in Section 1 requires the use of a rich and formal ontology model for describing the multimedia content: \(i\) distinguish real world objects from their digital representations that are available in multiple resolutions, \(ii\) make explicit the relationship between the annotations and the media, and the provenance of the annotations and \(iii\) handle fine-grained annotations and the decomposition of the media. The four MPEG-7-based ontologies proposed so far\(^4\) are able to represent this information partially. The

\(^1\)http://www.opencalais.com/
\(^2\)http://www.w3.org/TR/photo-rdf/
\(^3\)http://www.mindswap.org/2003/PhotoStuff/
\(^4\)http://www.w3.org/2005/Incubator/mmsem/XGR-vocabularies/#formal-MPEG-7
resulting annotations are for all of them, however, complex RDF graphs [6]. In addition, COMM also represents the context in which the annotations have been obtained.

Figure 2 shows the RDF graph corresponding to COMM-based annotations of the image depicted in the Figure 1. This description can be obtained using the open-source KAT annotation tool⁵ that makes use of the COMM JAVA API⁶. KAT allows the photographer to use any domain specific ontologies for describing the semantic content of the image. COMM provides additionally the descriptors for decomposing the image into regions, for representing the properties of the image and the low-level descriptors that can be extracted, and for making explicit the link between the annotations and the image. For simplicity, we focus our example on image annotations but both KAT and COMM have been designed for describing audio and video data as well.

COMM-based annotations are based on the DOLCE Description and Situations ontology pattern [3]. The upper-right part of the Figure 2 represents the media abstraction that is currently described. It extends the concept of MediaProfile defined in MPEG-7 [4]. The lower part details the segmentation of the image into three regions. The upper-left part corresponds finally to the semantic annotation of these three segments. We detail in the next section the added-value of using COMM for each part.

---

⁵https://launchpad.net/kat/
⁶http://multimedia.semanticweb.org/COMM/
3. COMM Added-value for Describing Multimedia Content

In this section, we explain the important features of the resulting RDF graph for the three parts mentioned above, namely: the abstraction of the media (3.1), the decomposition of the image (3.2) and the semantic annotation of the scene (3.3).

3.1. Modeling Multimedia Data Properties

The upper-right part of the Figure 2-A describes the media (i.e. the image) using the DOLCE Ontology of Information Objects ontology pattern [2]. The central concept is MediaProfile defined in MPEG-7 that describes the characteristics of the digital data that contains the multimedia content. For example, a digital image can be stored in different resolutions (thumbnail, full size), encoded in different formats (jpeg, tiff, png, bmp) and each compression algorithms may have a particular set of parameters. The MediaProfile descriptor is used for representing these properties. The algorithm used to obtain a particular digital data is specified with $m1:method$ – in fact, it is interpreted as one run of this algorithm for a set of input parameters (not displayed in Figure 2 for simplification).

The Situation is an abstract media annotation represented by the instance $mia1:media-instance-annotation$. One important property of this situation is the identification of the digital data itself (usually a file) provided by the entity $mid1:media-instance-descriptor$. The model handles multiple physical locations of a digital data, represented by several instances of the media-uri property. The same (image) file can therefore be available through file://, http:// or ftp:// URIs. All these identical files are assigned the same unique string represented by the instance $uis1:unique-id-string$. The instance $id1:digital-data$ is then understood as an abstract representation of the digital content of a file regardless its physical location. We may think of this representation as an “input” for further processing the image such as a segmentation as detailed in the next section.

We observe that several other ontology models are able to represent this abstraction. Hence, the FRBR model defines the work, the expression, the manifestation and the item [5], while the VRA model distinguishes the work from the image [7]. Finally, the CRM model has notions of information carrier and information object. COMM contains a list of axioms stating the equivalence of these concepts and can be seen as a bridge for interoperability between rich formal ontology models.

3.2. Modeling the Decomposition of the Image

The lower part of the Figure 2-B describes the image segmentation into three regions. We use again the D&S ontology design pattern for contextualizing the annotation. The Description is in this case the instance $m2:method$ that represents a method of segmentation or a run of a segmentation algorithm that can include some parameters. We have manually drawn three regions within the image using the KAT annotation tool. The Situation is then an abstract segmentation, or in other words, a spatial decomposition represented by the instance $srsd1:still-region-spatial-decomposition$. The two instances on the right side of the graph form the core of the D&S pattern.

The three similar blocks in the middle represent each a region of the image. The abstract digital data of each region is represented by the instance $id2:image-data$ (resp.
Each of these instances play a still region role in the segmentation process. The image data of each region is obtained by applying a mask on the image data of the whole image. The mask role is played by region locator instances (i.e. rld1:region-locator-descriptor) that are further specified by region boundaries of the appropriate polygon (in this case, rectangles). This approach enables to define arbitrary segments of multimedia data using the richness of the MPEG-7 mask descriptor.

3.3. Modeling the Semantic Annotations

The upper-left part of the Figure 2-C describes the semantic annotations for each abstract region resulting from the segmentation process using again the D&S ontology design pattern. The Situation is represented by the instance sa1:semantic-annotation (resp. sa2:semantic-annotation and sa3:semantic-annotation). The Description is the method or algorithm m3:method (resp. m4:method and m5:method) that corresponds to a manual annotation with no parameters in KAT. The design pattern enables to have multiple semantic annotations attached to a segment through a single method, i.e. a user can can precisely described the scene using several concepts and properties from domain ontologies. Several methods can be jointly used for describing a segment. Hence, it is possible to represent the collaborative annotations from several users, or the annotation of a user that completes the result from an automatic annotation algorithm such as a concept detector.

The left part of each semantic annotation graph contains the semantic label used in the annotation. In this example, we use dbpedia URIs for identifying the football players. These URIs are themselves dereferencable and link to further descriptions of the players such as the resume of their career following the Linked Data approach. Although at the first sight, the overall annotation graph is complex, we observe that it is based on the repeated use of the Descriptions and Situations ontology design pattern that provides a powerful approach for representing the context and provenance of the annotations.

4. Towards a Simplification of COMM-based Annotations

While having the complete COMM-based description of multimedia content stored in an RDF repository is useful for querying the context and provenance of different annotations, or for propagating the annotations through the various representations of digital data, it is not always efficient for simpler search tasks or for presentation to an end-user. For example, searching for images depicting the football player Zinedine Zidane uniquely identified by dbpedia:Zidane requires to go though an RDF graph of 8 nodes, but most of this long path is actually fixed by the DOLCE ontology design patterns. We propose to use SPARQL queries for creating additional “shortcuts” in the complex annotation graph and therefore optimize the search and presentation of multimedia content.

For example, we have defined a generic SPARQL query template for retrieving all segments of an image identified by a URL and for each segment, its semantic annotations (Figure 3). The whole image is considered as the root segment for which it is possible to retrieve its annotations in the same way. The decomposed-to property is a relationship...
between the image URI and its segments as they are obtained either manually (using for example the KAT tool) or by a segmentation algorithm. It replaces the longer path that comes from the original COMM description. Similarly, the annotated-by property links directly each abstract representation of an image segment to its multiple semantic descriptions. The definition of these properties provided by a library of SPARQL queries is available at http://nb.vse.cz/~vacuram/comm/pl/.

5. Conclusion

In this paper, we have shown the usefulness of COMM, a very rich formal ontology for describing multimedia content. Annotations can be automatically obtained using appropriate tools such as KAT. The downside of this model, however, is the complexity of the resulting RDF graphs, that make the search for content often inefficient and the presentation of the metadata to an end-user inappropriate. We have defined a number of SPARQL queries as templates for creating shortcuts in the model. Our ultimate goal is to define a complete presentation layer on top of the ontology model, using semantic web query and rule languages.

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".

References