Towards Bridging the Semantic Gap in Multimedia Annotation and Retrieval

Shankar Vembu, Malte Kiesel, Michael Sintek, and Stephan Baumann

Knowledge Management Lab,
German Research Center for Artificial Intelligence,
Erwin-Schrödinger-Straße 57, 67663 Kaiserslautern, Germany
{vembu, kiesel, sintek, baumann}@dfki.uni-kl.de

Abstract. We present a systematic approach to the design of multimedia ontologies based on the MPEG-7 standard and domain-specific vocabularies. The limitations of MPEG-7 in describing the semantics of highly structured domains like sports or medicine has led to an upsurge of interest in adopting an integrated approach to the design of ontologies. We follow suit and use MPEG-7 to model structural and low-level aspects of multimedia documents. The high-level semantics are modeled using a domain-specific ontology designed for soccer games. The integration of these ontologies is achieved by providing appropriate links to the individual ontologies. As a proof-of-concept, we describe a video annotation tool implemented as a plugin for the widely used Protégé ontology editor. The advantage of our methodologies lies in the fact that we use semantic web compliant languages and tools that results in alleviating the interoperability issues currently plaguing the multimedia and the semantic web communities.

1 Introduction

The availability of huge amounts of multimedia documents necessitates a careful design and an efficient implementation of information retrieval systems that facilitate storage, retrieval and browsing of not only textual, but also image, audio and video files. Querying multimedia documents based on their content would not satiate the needs of the user, since low-level features do not encapsulate the high-level semantics of a document. Like in natural language processing, the problems of polysemy and synonymy also arise in multimedia information retrieval and dealing with semantics in an appropriate fashion is indispensable. The need for a high-level representation that captures the true semantics of a document has led to the development of the MPEG-7 standard [1] for describing multimedia documents using machine-consumable metadata descriptors. MPEG-7 was not designed with the semantic web community in mind and has now led to some serious interoperability issues, the foremost one being the use of XML Schema as the definition language for describing the specification. Hunter [2] discusses, at length, the travails of mapping the MPEG-7 specification into a semantic web compliant language like RDF Schema. This epochal paper marked

the beginning of attempts to bridge the chasm between the multimedia and the semantic web communities. In another attempt, Hunter [3] proposes a framework for modeling ontologies that allows for semantic interoperability between MPEG-7 and other domain-specific vocabularies, by expressing, once again, the MPEG-7 descriptors in languages like RDF Schema and DAML+OIL.

The MPEG-7 standard, in addition to providing metadata descriptors for structural and low-level aspects of multimedia documents, also provides a set of tools to model the semantics in narrative worlds like events, objects, agents and places. The reader is referred to [4] for a general description of the semantic part of MPEG-7. The expressiveness of these descriptors in modeling the semantics of highly structured domains like sports or medicine is debatable and therefore integrating domain-specific ontologies with MPEG-7 is essential and has become the focus of research in recent years.

This paper is an attempt to facilitate a seamless integration of multimedia and semantic web vocabularies based on the approaches recently proposed and using a homogenous set of languages and tools. We posit the use of semantic web languages and tools in the design of ontologies, implementation of annotation tools, and semantic querying of multimedia documents. Our approach is to design a multimedia ontology using languages like RDF Schema and OWL. We use MPEG-7 to model low-level and structural aspects of multimedia documents, and domain-specific ontologies to model high-level semantics. Ontologies are modeled using the Protégé [1] editor. We also implemented a Protégé plugin for the semantic annotation of multimedia documents. Our approaches thus present a framework for multimedia annotation using semantic web compliant languages and tools, and circumvents most of the interoperability issues between multimedia and semantic web. In our exposition, we first discuss related work that uses semantic web languages and tools for describing multimedia documents in Section 2. In Section 3, we present our own approaches in designing a multimedia ontology based on MPEG-7 and integrate it with a domain-specific ontology. The domain-specific ontology was designed as part of the ongoing SmartWeb [2] project and a comprehensive description of it is beyond the scope of this paper. We therefore outline only the main concepts. We describe the design and implementation of a video annotation tool in Section 4. We finally point to directions for future work and conclude the paper in Section 5.

2 Related Work

Recent years have seen quite a number of attempts that stress the importance of complementing the MPEG-7 standard with domain-specific ontologies for multimedia annotation [5–11].

In [11], the authors describe a framework for the transparent integration of domain-specific ontologies with audio-visual content standards like MPEG-7 and

[1] http://protege.stanford.edu/
[2] http://www.smartweb-project.org/

TV-Anytime [3]. But the presented approach uses XML Schema as the language for the design of ontologies and makes no attempt to map the existing MPEG-7 standard into a semantic web compliant language.

Tsinaraki et al. [8, 9] describe a framework for extending MPEG-7 and TV-Anytime with domain-specific ontologies. They express the semantic part of MPEG-7 Multimedia Description Schemes (MDS) [12] in OWL and domain-specific ontologies extend this core ontology to fully describe the concepts of application domains. Taking the example of soccer domain, the FootballTeam concept from the domain-specific ontology extends the OrganizationType of the core ontology that was designed based on MPEG-7. An attempt to completely move MPEG-7 to the semantic web world is described by Garcia and Celma [13]. The authors automate the entire conversion of MPEG-7 standard to OWL as envisaged by Hunter [2] using XML Schema to OWL mappings. However, the authors try to map the domain-specific vocabularies to the ones provided by MPEG-7. For example, the concept Artist from a music ontology is mapped to the MPEG-7's concept CreatorType by sub-classing. The scalability of these approaches is questionable, since it tightly couples the domain-specific concepts with the semantic part of MPEG-7 by specialisation. Concept mapping based on semantic relations of equivalence or inclusion in the ontologies may not be feasible for domains with rich semantics.

Troncy [7] asserts the need to infer multimedia documents at both their structural as well as conceptual aspects. MPEG-7/XML Schema is used to express the structural meaning of multimedia documents, and OWL is used to model their semantic aspects using a domain-specific vocabulary. A transformation mechanism from OWL to XML Schema results in XML Schema descriptors of the domain-specific constructs that could, if possible, be linked with existing MPEG-7 types by specialisation. Isaac and Troncy [14] describe a case study in the medical domain and shows that the combination of several ontologies results in better description and retrieval of audio-visual sequences. Bloehdorn et al. [6] describes an ontological framework and a software environment that allows linking of low-level MPEG-7 visual descriptors to concepts in domain-specific ontologies based on a prototype approach.

Our approach to the design of multimedia ontology bears resemblances to the works of Troncy [7], Tsinaraki et al. [8, 9], and Hunter et al. [10] in the sense that we express the MPEG-7 standard in a semantic web compliant language like OWL to resolve the interoperability issues and complement the standard using domain-specific vocabularies. But an important distinction in our approach is that we do not rely on the semantic part of MPEG-7 MDS but only focus on its structural descriptors. The low-level concepts are taken from the visual [15] and audio [16] parts of the standard. The high-level semantics of a domain are entirely captured using domain-specific ontologies. The integration of ontologies is achieved using a simple and straightforward mechanism unlike the approach adopted by Hunter et al. [10] where a more complex metadata integration model like the ABC model [17] is used.

[3] http://www.tv-anytime.org

3 Ontology Design

We first describe an ontology to model events in soccer games. We call this domain-specific ontology as Sport Event ontology. Later on, we present our approaches to designing a multimedia ontology based on the MPEG-7 standard. We postulate, based on our experiences, that the semantic part of MPEG-7 standard is inadequate to model the semantics of domains like soccer. Therefore, the set of structural and low-level MPEG-7 descriptors are complemented with the high-level semantics expressed by the Sport Event ontology. We call this integrated ontology as SmartMedia ontology.

3.1 Sport Event Ontology

Football (i.e., soccer) is an important application domain of the SmartWeb project and we designed the Sport Event ontology to model it. It provides a schema to model entities such as tournaments, matches, match events (like goal shots and fouls) as well as persons, places and standings tables along with results and fixtures. While the ontology provides foundations for modeling generic sport event data, it is most expressive for describing football events.

While it is evident that we cannot model the entire domain of football, the Sport Event ontology provides structure for most of the information found in typical football data sources such as web pages, data bases and books. The Sport Event ontology uses the concepts and properties introduced in SmartSUMO[4] as super-classes of its football domain concepts. Thus, the role FootballPlayer, which relates to a SmartDOLCE:natural-person, automatically can bear information on the football player's birthdate or nicknames without the need to explicitly remodel these properties in the Sport Event ontology.

In order to complement the Sport Event ontology's documentation, a large number of example instances that models facts concerning football world cups is also available. These facts are used to test SmartWeb's ontology-based search components and also to verify modeling decisions.

The main modeling challenges when creating the Sport Event ontology were proper multilingual handling of names of instances, giving structure to the loosely defined terms in the sports domain, providing linguistic information for concepts and relations, allowing extraction components to automatically create new fact instances and even facilitating semi-automatic ontology extensions. Not surprisingly, there have been a lot of technical challenges too: keeping the Sport Event ontology in sync with SmartSUMO was difficult sometimes, adapting components that build on the Sport Event ontology such as the query generation modules was tedious work, and also making sure that the facts supplied by automatic components fit to the ontology schema proved difficult. A discussion of these issues are beyond the scope of this paper.

[4] SmartSUMO is a combination of the DOLCE foundational ontology [18] and the SUMO upper ontology [19] modified for the needs of SmartWeb

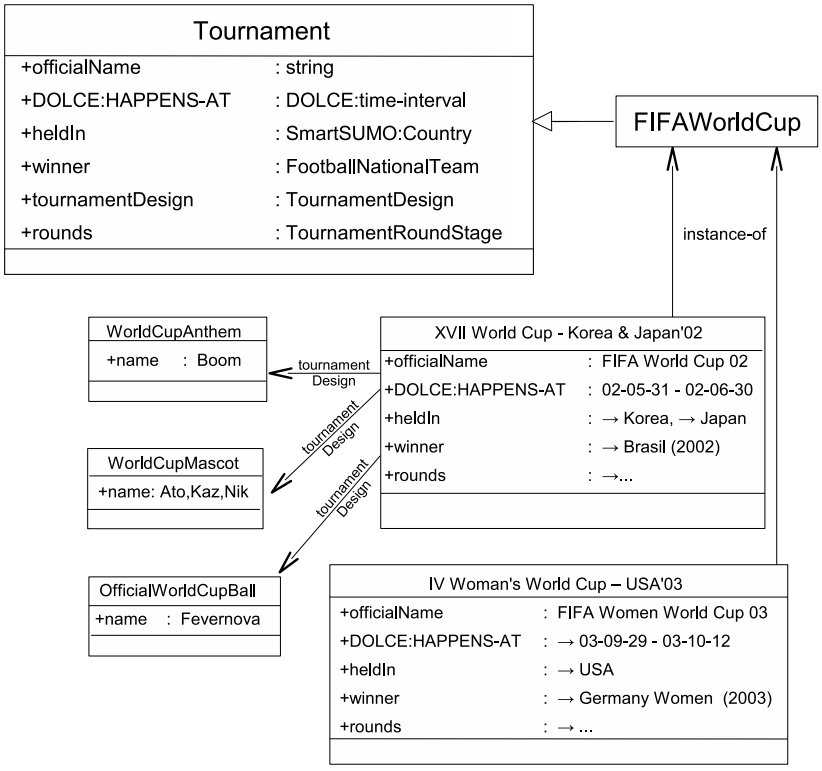

Fig. 1. Example of an FIFAWorldCup instance

In order to give an idea about the size of the Sport Event ontology, there are more than 500 concepts, about 70 relations, about 2000 linguistic annotations, and about 4300 instances, not including relations and annotations inherited from SmartSUMO.

A Sport Event Ontology Tour To describe the essential features of the Sport Event ontology, we start a tour beginning with the concept FIFAWorldCup (sub-concept of FootballWorldCup) shown in Figure 1 along with two of its instances and example instances of the concepts WorldCupAnthem, WorldCupMascot and OfficialWorldCupBall.

In Figure 2, an overview of the concept structure surrounding FIFAWorldCup is shown. Since RDF Schema supports only a single level of instance-of relationships, there can only be concepts and instances (metaclasses notwithstanding), forcing users to draw a sharp and sometimes artificial border as shown in the figure. One could argue that, for example, UEFAChampionsLeague is an instance of EuropeanFootballTournament. However, we chose to model it as a sub-concept of EuropeanFootballTournament, leaving the instance layer for instances of UE-

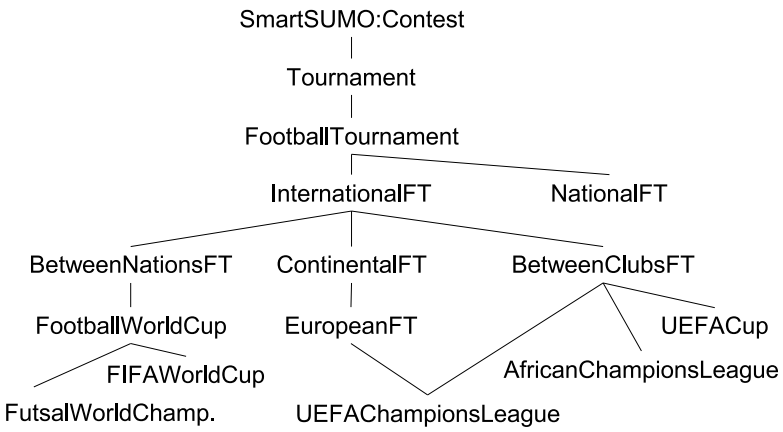

Fig. 2. SmartSUMO:Contest and some of its sub-classes

FAChampionsLeague such as a league taking place in a certain year. Another thing we see in the figure is that multiple inheritance is used since it comes natural with this design decision. If we decided drawing the concept/instance border higher, concepts such as UEFAChampionsLeague would have been modeled as instances which has two major drawbacks: first, this instance would have to be an instance of two concepts (EuropeanFT and BetweenClubsFT), which only few tools support, and second, the relationship between UEFAChampionsLeague and its instances would have to use an arbitrary relation without taking advantage of RDF Schema's built-in semantics.

There are a number of other important concepts, most notably the Match and the Team concepts and their sub-concepts modeling individual matches, teams and clubs. An instance of a Match can contain information on the stadium and country it was held in, participants (both players and other people such as referees and trainers), events that have occurred in the match, and other information such as the number of spectators.

Modeling teams is a challenge in the football domain since Team is quite an ambiguous concept ranging from football clubs and squads to the team playing in an actual match. In order to reflect this, the ontology includes a Team concept which is supposed to be as fuzzy as "Team" is naturally, but there are also subclasses that are far less ambiguous and should be used if hard data is available (i.e. for facts extracted by hand or by mapping databases, and not by automatic natural language processing).

3.2 Multimedia Ontology

Our multimedia ontology is based on the MPEG-7 standard and we use Protégé for modeling it, thereby allowing us to export the concepts to languages like RDF Schema and OWL. The discrepancies that arise during such a mapping can be

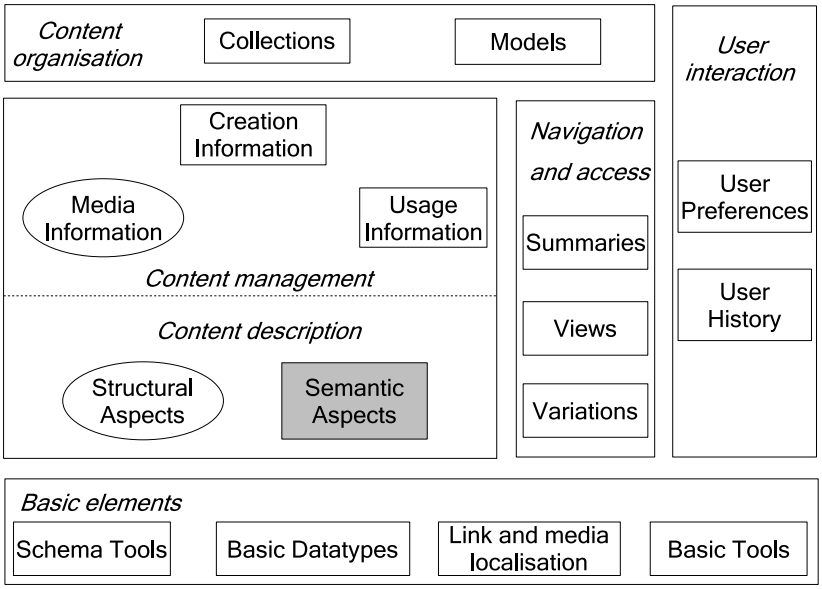

Fig. 3. Overview of MPEG-7 MDSs (adaptation of Figure 5 in [2]). Entities enclosed in ellipses are the DSs modeled in our multimedia ontology. The entity in the grey-coloured rectangle is modeled using our domain-specific ontology

resolved, to a large extent, by following the guidelines laid out by Hunter [2] and Garcia and Celma [13]. As a first step, we model only those MPEG-7 concepts that fit well to our project at hand. The MPEG-7 MDS specifies various description tools or metadata structures for describing and annotating audio-visual content. The MDSs are organised into the following six categories [2]: Basic Elements, Content Description, Content Management, Content Organisation, Navigation and Access, and User Interaction. We are interested in concepts that provide storage features of multimedia documents like format, encoding and location. These concepts are found in the Media Information DS. We focus on structural aspects like spatial, temporal, spatio-temporal concepts, and also on certain low-level features [15, 16] like colour, shape, texture, timbre and melody for images, video and audio files. The semantic aspects are modeled using the Sport Event ontology. Figure 3 provides an overview of the MPEG-7 MDSs. The description schemes modeled in our multimedia ontology are shown in ellipses.

We describe the taxonomical details of our multimedia ontology. The top-level multimedia content entity is called MultimediaContent. The concepts of Audio, Video, Image and Text are sub-classes of this top-level entity. As laid out in the MPEG-7 specification, these concepts are modeled in such a way so as to describe the intrinsic recursiveness of multimedia in general. For example, an audio file can be considered to be a set of segments or snippets at different temporal locations, which could in turn be recursively divided. In the case of videos, the recursion also extends in the dimension of space giving rise to a complex set of

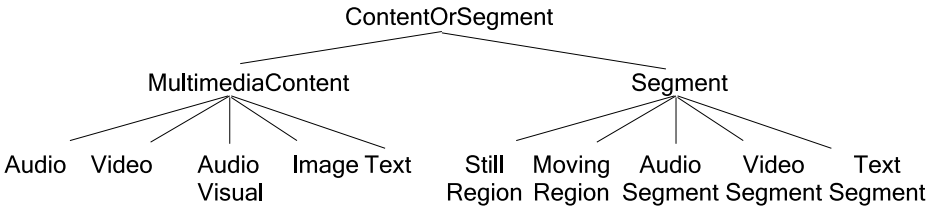

Fig. 4. MPEG-7 taxonomy

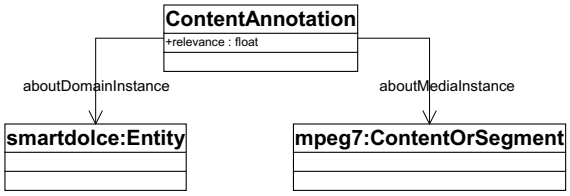

Fig. 5. Top-level concept of the SmartMedia ontology

temporal, spatial and spatio–temporal segments. The MPEG-7 concepts namely MultimediaContent and Segment together make this kind of recursive modeling possible. We desgined a class called ContentOrSegment, which is an abstraction of the high-level entities MultimediaContent and Segment of MPEG-7 as shown in Figure 4, and is also is the top-level concept of our multimedia ontology.

3.3 SmartMedia Ontology

The SmartMedia ontology integrates the multimedia ontology and the Sport Event ontology described in the earlier sections. We designed a class called ContentAnnotation shown in Figure 5 with concepts Entity from SmartDOLCE ontology and ContentOrSegment from MPEG-7 ontology as its properties. The property *relevance* indicates the strength of relationship between the domain instance and the multimedia instance. For example, it takes a value of 1.0 for an image showing a football player X (completely and exclusively) and a domain instance representing X. By linking concepts from MPEG-7 and domain specific ontologies, the concept of ContentAnnotation forms the basis to bridge the so-called *semantic gap* in multimedia and semantic web. Moreover, this integrated ontology can be exported in its entirety, to formats like RDF Schema and OWL using Protégé.

4 Protégé Multimedia Plugin

Protégé provides a rich set of functionalities for the creation, visualisation and manipulation of ontologies in various representation formats like RDF Schema and OWL. Furthermore, its extensible, plug-and-play architecture makes it an

enticing platform to implement our annotation tool. We envisage that such a plugin would allow users to leverage the features provided by the editor for modeling complex ontologies and at the same time efficiently annotate multimedia documents by integrating domain-specific ontologies with MPEG-7 ontology following the guideliness described in the previous section.

We outline the technical details and elucidate the functionalities provided by the multimedia plugin. The plugin was implemented using Java Media Framework API [5] and open-source APIs like OmniVidea FOBS [6]. Figure 6 shows a screenshot of the plugin. Ontologies can be imported and displayed using the class browser, a standard feature of Protégé shown to the left of the screenshot. The user is allowed to create media instances on the fly, open media documents and drag-and-drop these instances at specific temporal locations of the document. A drag-and-drop action would automatically populate the properties of the instances. For example, let's consider a concrete scenario where a user would like to annotate soccer videos. He is interested in annotating a particular temporal location in the video where a goal was shot. He selects an appropriate concept representing the semantics of *shot-on-goal* from the Sport Event ontology and performs a drag-and-drop action on the media document. This would immediately trigger the creation of an instance for the top-level concept of the SmartMedia ontology namely, ContentAnnotation having the domain-specific instance (shot-on-goal) and a multimedia instance as its properties. A certain level of automation can be achieved by populating properties of the relevant multimedia concepts like format and encoding of the document, and the temporal point at which the annotation was performed. The current version of the plugin does not support any advanced signal processing techniques, but a straightforward extension would be to facilitate ways to automatically extract metadata descriptors using approaches developed in the image processing techniques for video documents and the music information retrieval community for music documents.

The user is also provided with a list of his annotations displayed to the right of the screenshot. The selected annotation is displayed in the instance editor provided by Protégé. The user is thus allowed to manually modify his annotations using the instance editor. The built-in RDF Schema support of Protégé allows the user to export his annotations into formats like RDF. An inference engine that supports querying RDF documents using languages like SPARQL or RDQL can be used to retrieve the annotated multimedia documents.

5 Conclusions and Future Work

In this paper, we have described a systematic approach for the annotation of multimedia documents by relying on a homogeneous set of languages and tools used in the semantic web community. We believe that mapping the MPEG-7 specification to a semantic web compliant language, leveraging its low-level

[5] http://java.sun.com/products/java-media/jmf/index.jsp
[6] http://fobs.sourceforge.net/index.html

Fig. 6. Protégé multimedia plugin

and structural metadata descriptors, and integrating it with domain-specific ontologies to model the high-level semantic aspects is indeed a step in the right direction to bridge the semantic gap in multimedia annotation and retrieval. We use the phrase *semantic gap* to highlight the fact that content-based retrieval of documents in itself is not sufficient to satisfy the end-user and that high-level semantics captured using domain-specific ontologies play an important role. Whilst inferring semantic descriptors from multimedia documents manually is trivial, automating this process is still an active area of research that draws attention from various fields like machine learning, audio and image processing.

At this juncture, we would like to point out to few directions that are worth considering in the future. It is needless to mention the pervasive use of search engines like Google that uses a query-by-text approach to the retrieval of documents. Owing to the fact that the common user is very well acquainted to querying text-based documents using key words or phrases, it seems natural that he would like to do the same for retrieving multimedia documents. Recent developments in machine learning has led to the design of systems that automatically annotate multimedia documents using keywords. The reader is referred to [20–23] for an overview of the state-of-the-art approaches for the automatic semantic annotation of images and videos. It would definitely be beneficial to take

advantage of these developments and integrate them with existing multimedia annotation systems used in semantic web research.

Looking beyond the borders of the semantic web, it is obvious that knowledge discovery is a further topic of interest when dealing with multimedia content. We are interested in using the presented approaches to annotate a collection of music video clips. This corpus consists of 160 clips of the most famous directors in this field. The domain ontology is rich by means of nested semantic relations about the bands, singers, actors and general themes involved. On the structural and low level, we would like to apply standard image and video processing techniques to automatically perform the annotation with basic features. Finally, we aim at the discovery of typical aesthetical concepts of specific directors or musical genres using machine learning algorithms.

Acknowledgement

This research was funded by the German Federal Ministry for Education and Research under grant number 01IMD01A, *SmartWeb* (http://www.smartweb-project.org). We would like to thank the reviewers for their valuable comments that led to improvements of an earlier version of this paper.

References

1. Chang, S.F., Sikora, T., Puri, A.: Overview of the MPEG-7 standard. IEEE Transactions on Circuits and Systems for Video Technology **11**(6) (2001) 688–695
2. Hunter, J.: Adding multimedia to the semantic web - Building an MPEG-7 ontology. In: Proceedings of the International Semantic Web Working Symposium, Stanford University, California, USA (2001)
3. Hunter, J.: Enhancing the semantic interoperability of multimedia through a core ontology. IEEE Transactions on Circuits and Systems for Video Technology **13**(1) (2003) 49–58
4. Benitez, A.B., Rising, H., Jörgensen, C., Leonardi, R., Bugatti, A., Hasida, K., Mehrotra, R., Tekalp, A.M., Ekin, A., Walker, T.: Semantics of multimedia in MPEG-7. In: Proceedings of the IEEE International Conference on Image Processing, Rochester, New York, USA (2002)
5. Bloehdorn, S., Simou, N., Tzouvaras, V., Petridis, K., Handschuh, S., Avrithis, Y., Kompatsiaris, I., Staab, S., Strintzis, M.G.: Knowledge representation for semantic multimedia content analysis and reasoning. In: Proceedings of the European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology, London, U.K. (2004)
6. Bloehdorn, S., Petridis, K., Saathoff, C., Simou, N., Tzouvaras, V., Avrithis, Y., Handschuh, S., Kompatsiaris, Y., Staab, S., Strintzis, M.G.: Semantic annotation of images and videos for multimedia analysis. In: Proceedings of the Second European Semantic Web Conference, Heraklion, Crete, Greece (2005)
7. Troncy, R.: Integrating structure and semantics into audio-visual documents. In: Proceedings of the Second International Semantic Web Conference, Florida, USA (2003)

8. Tsinaraki, C., Polydoros, P., Christodoulakis, S.: Integration of OWL ontologies in MPEG-7 and TV-Anytime compliant semantic indexing. In: Proceedings of the Sixteenth International Conference on Advanced Information Systems Engineering, Riga, Latvia (2004)
9. Tsinaraki, C., Polydoros, P., Christodoulakis, S.: Interoperability support for ontology-based video retrieval applications. In: Proceedings of the Third International Conference on Image and Video Retrieval, Dublin, Ireland (2004)
10. Hunter, J., Drennan, J., Little, S.: Realizing the hydrogen economy through semantic web technologies. IEEE Intelligent Systems - special eScience issue (2004)
11. Tsinaraki, C., Polydoros, P., Kazasis, F., Christodoulakis, S.: Ontology-based semantic indexing for MPEG-7 and TVAnytime audiovisual content. Special issue of Multimedia Tools and Applications Journal on Video Segmentation for Semantic Annotation and Transcoding **26** (2005) 299–325
12. ISO/IEC 15938-5: FCD Information technology - Multimedia content description interface - Part 5: Multimedia description schemes (2003)
13. Garcia, R., Celma, O.: Semantic integration and retrieval of multimedia metadata. In: Proceedings of the Fifth International Workshop on Knowledge Markup and Semantic Annotation at the Fourth International Semantic Web Conference, Galway, Ireland (2005)
14. Isaac, A., Troncy, R.: Using several ontologies for describing AV documents : A case study in the medical domain. In: Proceedings of the Multimedia and the Semantic Web Workshop at the Second European Semantic Web Conference. (2005)
15. ISO/IEC 15938-3: FCD Information technology - Multimedia content description interface - Part 3: Visual (2001)
16. ISO/IEC 15938-4: FCD Information technology - Multimedia content description interface - Part 4: Audio (2001)
17. C. Lagoze, J.H.: The abc ontology and model (v3.0). Journal of Digital Information **1**(2) (2001)
18. Gangemi, A., Guarino, N., Masolo, C., Oltramari, A., Schneider, L.: Sweetening ontologies with DOLCE. In: Proceedings of the Thirteenth International Conference on Knowledge Engineering and Knowledge Management, Siguenza, Spain (2002)
19. Pease, A., Niles, I., Li, J.: Origins of the IEEE standard upper ontology. In: Working Notes of the AAAI-2002 Workshop on Ontologies and the Semantic Web, Edmonton, Canada (2002)
20. Barnard, K., Duygulu, P., Forsyth, D., de Freitas, N., Blei, D., Jordan, M.: Matching words and pictures. Journal of Machine Learning Research **3** (2003) 1107–1135
21. Duygulu, P., Barnard, K., de Freitas, N., Forsyth, D.: Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In: Proceedings of the Seventh European Conference on Computer Vision, Copenhagen, Denmark (2002)
22. Iyengar, G., Duygulu, P., Feng, S., Ircing, P., Khudanpur, S.P., Klakow, D., Krause, M.R., Manmatha, R., Nock, H.J., Petkova, D., Pytlik, B., Virga, P.: Joint visual text modeling for automatic retrieval of multimedia documents. In: Proceedings of the ACM Multimedia, New York, NY, USA (2004)
23. Lavrenko, V., Feng, S., Manmatha, R.: Statistical models for automatic video annotation and retrieval. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Montreal, Quebec, Canada (2004)