A lightweight web video model with content and context descriptions for integration with linked data

Conference or Workshop Item

A Lightweight Web Video Model with Content and Context Descriptions for Integration with Linked Data

Smitashree Choudhury
Digital Enterprise Research Institute
National University of Ireland, Galway
Dangan, Galway, Ireland
+ 353 91 495053
smitashree.choudhury@deri.org

John G. Breslin
School of Engineering and Informatics
National University of Ireland, Galway
University Road, Galway, Ireland
+353 91 492622
john.breslin@nuigalway.ie

Stefan Decker
Digital Enterprise Research Institute
National University of Ireland, Galway
Dangan, Galway, Ireland
+ 353 91 495011
stefan.decker@deri.org

ABSTRACT
The rapid increase of video data on the Web has warranted an urgent need for effective representation, management and retrieval of web videos. Recently, many studies have been carried out for ontological representation of videos, either using domain dependent or generic schemas such as MPEG-7, MPEG-4, and COMM. In spite of their extensive coverage and sound theoretical grounding, they are yet to be widely used by users. Two main possible reasons are the complexities involved and a lack of tool support. We propose a lightweight video content model for content-context description and integration. The uniqueness of the model is that it tries to model the emerging social context to describe and interpret the video. Our approach is grounded on exploiting easily extractable evolving contextual metadata and on the availability of existing data on the Web. This enables representational homogeneity and a firm basis for information integration among semantically-enabled data sources. The model uses many existing schemas to describe various ontology classes and shows the scope of interlinking with the Linked Data cloud.

Categories and Subject Descriptors
H.3.5 [**Online Information Services**]: Web-based services

General Terms
Management, Design, Experimentation, Standardization.

Keywords
Multimedia, ontology, Semantic Web, linked data, user-generated content, annotation.

1. INTRODUCTION
The rapid increase in the amount of available videos on the Web has resulted in demands for more effective and efficient approaches for video search and retrieval. Representation and semantic annotation of multimedia documents and content have been identified as an important step towards the efficient management and retrieval of multimedia. In order to achieve semantic analysis and retrieval of multimedia, content has to be described in machine understandable formats with ontological support [18]. Recently, the W3C has proposed that video be made into a first-class object on the Web. Numerous studies to understand visual media have been able to reduce the "semantic gap" (between representation formats and user cognition), but these are still not viable at web scale. Core researchers in computer vision and multimedia retrieval have recently shifted their focus slightly to include knowledge-based approaches including ontology-supported methods to address the problem of the semantic gap. Efforts have been made to model non-textual data at various levels of abstractions. Different vocabularies have been suggested for representing audio, video and images so that they can be efficiently organized, managed and retrieved. There are two different approaches. The first is *shallow modeling*, which treats the video as a document or as a part of a document, and a vocabulary is created to represent these document properties. Examples of such vocabularies are Dublin Core and Media RSS[1], where a video is represented mainly with its editorial information such as "creator", "date of creation", "duration", etc. The second approach, *deep modeling*, is where efforts are made to describe and represent the content information at the signal level (pixels, audio, etc.). In a minimalist view, generic image annotations can be extended for video frames using vocabularies such as Digital Media[2] or Image Region using RDF [17] representations. However, video representation cannot be fully captured through image descriptions. The inherent temporal dimension has to be taken into consideration. Studies have been carried out to represent pixel information at various levels of segments such as video frame, image region, signal segment, etc. Many complex and comprehensive studies have been performed with details of video contents and their representation scope. Among the latter, MPEG-7 [9] is the most comprehensive content description framework for audio and video data. In spite of its wide scope and comprehensively-detailed specification, it has some problems which have prevented it from being widely accepted. Due to complex data types and a lack of formalization, the same document may be described in various ways by different agents giving rise to interoperability problems. As a result, with the evolution of Semantic Web technologies, researchers have tried to partially or fully transform the MPEG-7 framework into a Semantic Web framework.

In the present study, we have designed a lightweight and user-centric model to jump start the metadata creation process with minimal user input. In previous approaches, video has been assumed to be an individual digital object or document which contains various data and metadata elements that need to be exploited and represented. To our understanding, video on the

[1] http://search.yahoo.com/mrss/

[2] http://www.mindswap.org/2005/owl/digital-media

Web cannot solely be considered as a document or as a part of a document. A video object on the Web always appears in some context in general and more recently in various social contexts. It is shared, distributed and discussed in various social groups on media sharing sites or social networking sites. Even professional agencies such as production houses and news channels now create and upload their videos to such sites for improved viewer coverage.

We will consider two simple examples in our practical use cases. Firstly, let us imagine a researcher looking for perspective differences amongst news channels who are reporting on specific issues. For example, he or she wants to collect all the videos about post-election violence in Iran in 2009 uploaded by the two news channels "CNN" and "Al Jazeera". Can they get the desired results without much manual intervention? The query involves many facets, "the video has to be about post-poll protests", "it has to be in a certain time frame, i.e. 2009", and moreover, "the results should enlist those videos uploaded by the above news channels". On the Web, he or she will have to go to different video sharing sites, manually browse the videos for the desired topics, and subsequently filter some video listings using different constraints and facets before getting the final list.

It is also highly imprecise to formulate a query through simple keyword search or tag-based search. In our second use case, let us imagine we have a new research student working on a Semantic Web project who wants to find all videos where Tim Berners-Lee is speaking about the Linking Open Data project. The problem is that most of the videos are from quite long presentations, and he or she only needs those segments where the topic of interest appears. With present frameworks, getting a video segment out of a document is not functional on the Web. Any retrieval will end up with the entire video as the result, and the student has to manually watch the entire video to seek the appropriate cue point.

Fragment identifiers, i.e. localized identifiers for media segments both at the spatial and temporal level, are a fundamental requirement for querying and retrieving multimedia documents and segments. We have followed W3C guidelines (as in their working draft document [23]), using both temporal and spatial dimensions to identify the segments.

The rest of this paper is organized as follows: Section 2 presents a short review of related studies. Section 3 presents the requirements of a multimedia ontology, followed by our contribution. Section 4 gives a detailed description of the content model. Section 5 presents the use cases for the model. The final section outlines our conclusions.

2. RELATED STUDIES

This section describes some of the studies related to our present study. The 'Image Annotation on the Semantic Web' document[3] from the W3C Multimedia Semantics Incubator Group references various vocabularies, applications and use cases that can be used for image annotation tasks. More work is ongoing to define a set of mappings between these various models and on how to efficiently combine different ontologies for annotating multimedia content in general. This is one of the current tasks of the W3C Media Annotations Working Group (MAWG) [23], as defined in their document on "web video"[4]. As defined by the charter of the group, its goal 'is to provide an ontology designed to facilitate cross-community data integration of information related to media objects in the Web, such as video, audio and images'. A first draft of this ontology was published on 18 June 2009[5]. For the sake of interoperability, it has defined mappings between its core properties and other existing media models such as Media RSS, EXIF, ID3, MPEG7, etc. The Media Object ontology from MAWG is still in progress, and describes mostly editorial properties at the document level (we may reuse some of these while establishing mappings to other properties).

Hunter et al. [4] described work to convert MPEG-7 modules into RDFS and later integrated this into the ABC ontology. Garcia and Celma [3] produced the first complete MPEG-7 ontology, automatically generated using a generic mapping from XSD to OWL. Simou [20] proposed an OWL-DL Visual Descriptor Ontology (VDO) based on the visual part of MPEG-7 and this was used for image and video analysis. Troncy et al. [16] proposed a core multimedia ontology to add semantics to MPEG-7 for content description.

Another approach for modeling video content [15] is domain specific, making it difficult to generalise across domains. Jaimes et al. [8] proposed extending a linguistic ontology for multimedia by means of semi-automatic learning from domain-related videos. The Large Scale Concept Ontology for Multimedia [11] was designed for conceptual indexing of news videos. Researchers identified more than 800 semantic concepts in the taxonomy. Researchers in [1] describe video content based on physical objects and their spatial relationships, but ignore the spatio-temporal aspects of video. Petkovic and Jonker propose in [10] a model for events in video sequences and consider four information layers (from low-level to high-level layers): pixels, characteristics, objects, and events. Thus, it is possible to detect specific events by defining the states and the interactions between suitable objects.

All of the above studies and many others describe media either at the document level or at the content level (or both) in great detail, but the absence of evolving social context descriptions and its possible contribution to content understanding has motivated us in our present work.

3. REQUIREMENTS OF MULTIMEDIA ONTOLOGY

The requirements for our ontology across domains are mostly uniform, but there are some media-centric requirements such as separation of concern. Below are some the major ontological requirements that had to be fulfilled when designing our model.

3.1 Interoperation and Integration

Semantic interoperability across schemas and their concepts is one of the prime requirements for any ontology model. It should be possible to map to the concepts of existing ontologies such as the W3C Media Object ontology, SIOC, FOAF, Dublin

[3] http://www.w3.org/TR/swbp-image-annotation/

[4] http://www.w3.org/2008/WebVideo/Annotations/

[5] http://www.w3.org/TR/2009/WD-mediaont-10-20090618/

Core, VDO, etc., either by means of "*rdfs:seeAlso*" or "*owl:sameAs*" predicates. Interoperability harnesses greater integration and provides support for an in-depth analysis and searching of data across various data domains.

3.2 Ease of Use

An ontology that is expected to be used by people other than domain experts and ontology engineers has to follow the principles of KISS (Keep It Small and Simple). No matter how comprehensive and widely-modeled ontology is, the framework is of little value if it cannot be used by users with a minimum of effort and a small learning curve. Simplicity will increase its usage and popularity, and this leads to more applications and tool support.

3.3 Modularity and Extensibility

Clear and independent ontology modules allow for ease of maintenance and extensibility for the framework when and where needed. In the proposed model, we have used several separate and clearly distinguished modules such as an events module, a time module, etc., and we have linked to the Visual Descriptor Ontology (VDO) for content representation.

3.4 Representation of content structure

For multimedia reasoning, the multimedia object should be represented at different levels of structural granularity such as shot, frame, image region, audio segment, etc. The spatial and temporal relations between various structural components have to be explicit in order to use heuristics and reasoning services.

3.5 Separation of Content and Domain Knowledge

A multimedia model has to maintain a strict separation between its own content feature description and the depicted semantic domain concepts. The need for such a requirement comes from the idea that media processing is multimodal in nature and no single mode will suffice for the entire spectrum of metadata. Content processing is often required for establishing the mapping between low-level signals and high-level semantic concepts through some induction algorithms.

3.6 Reasoning Support

"A little bit of semantics" goes a long way, so basic relations such as subsumption and transitivity may be all that is required for most reasoning in ontologies. However, a multimedia ontology also needs to support the notion of uncertainty and fuzzy reasoning.

4. OUR APPROACH

Our contribution to the paper is in three areas:

1. Social context modeling
2. Integrating events in video descriptions
3. Interlinking to the Linked Data cloud

4.1 Context Contribution

Our proposed model presumes that video is a social object and it can therefore best be understood in terms of its context of use and interaction. On the Web, context includes information of the user who created the video, his or her interests, tagging information, usage statistics such as the number of viewers, the number of ratings received, the number of times a video has been bookmarked by users, geolocation information (where the video has been recorded), its inclusion in topical groups, as well any comments by other users. Explicit user contexts may not be directly available, but many cameras are available that incorporate various sensors to capture data such as location, motion, temperature, weather, etc., and all of these can be included in the process of metadata extraction to understand the media better. Social context is dynamic in nature and evolves over a certain period of time. This socially-generated data can assist in searching and in personalizing the media.

4.2 Event Modeling

Videos are mostly event centric. The original purpose for making a video is to capture certain ongoing events. The representation and recognition of events in video is important for content-based browsing and retrieval. Events may be a simple event or a complex one with a combination of simple events. The present event module prescribes a simple representation of an event, but a more specialized version of an event model will be integrated at a later stage for a finer representation of sequences.

4.3 Resource Interlinking

As well as describing video data using our ontology, we wish to go one step further by semantically enriching the metadata following Linked Data principles. Since our proposed model is meant for web videos, our aim is to integrate the web videos into the Linked Open Data cloud that aims to create a "web of data" instead of a web of documents. As part of this initiative, many open data sources (e.g. Wordnet[6], DBLP, Wikipedia[7], GeoNames, MusicBrainz) with billions of facts have been interlinked and mapped (see Figure 1).

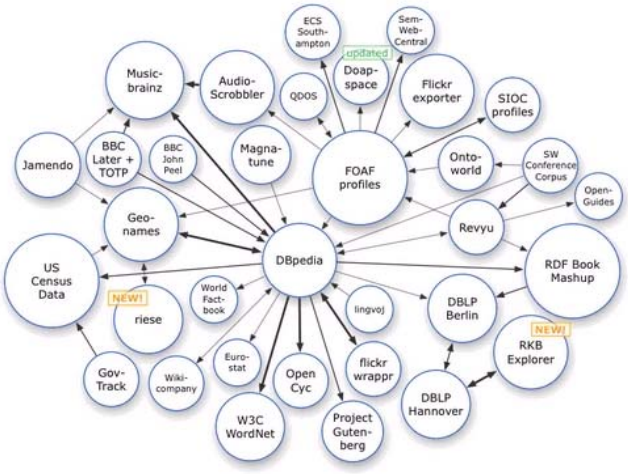

Figure 1. The Linking Open Data dataset cloud[8].

The basic premises of data interlinking is to have a URI for every entity on the Web (objects, people, event, concepts) which can be dereferenceable on the Web. Dereferencing URIs can make more related information about the resource available. Thanks to this effort, lots of RDF data is now available on the Web and it can be used in various applications, from advanced data visualisations and querying systems to complex mashups.

[6] http://wordnet.princeton.edu/

[7] http://www.wikipedia.org/

[8] http://richard.cyganiak.de/2007/10/lod/

More importantly, this data is all linked together, which means that one can easily navigate from one information source to another through the linked graph. A user agent can use these resources to crawl and extract more information, and can even perform reasoning based on the underlying ontological properties. Our model explores all possibilities of data interlinking to various sources such as DBpedia [2], FOAF[9], geospatial data, etc. Location information can be mapped to the formalized knowledge available in GeoNames. Genre and tagging information can be mapped to various domain concepts in DBpedia. People information such as creators, musicians or directors can be linked to their FOAF profiles.

5. WEB VIDEO FRAMEWORK

The proposed framework is based on three clusters of information spaces (as shown in Figure 2 below), where each contributes towards a more flexible and robust media understanding and interpretation: (1) the editorial, production and technical feature space; (2) content features description; and (3) the social and contextual feature space. The uniqueness of the proposed model is that it has attempted to model the evolving social context of the media to provide a better understanding of the media. We will discuss existing vocabularies such as Time, VDO, FOAF, Dublin Core[10], SIOC[11] and other ontologies used in the framework before going on to the core framework model.

Time: Due to the temporal nature of video all the temporal descriptions are modeled as a time object. We have used the OWL Time ontology to describe the temporal attributes. The Time ontology formalises different temporal notions in a Semantic Web framework by providing web identifiers for a number of concepts. Each temporal instance is expressed as a *time:TimeInterval* which has *time:hasBeginning* and *time:hasEnd* attributes. More details about the Time ontology are available in [22].

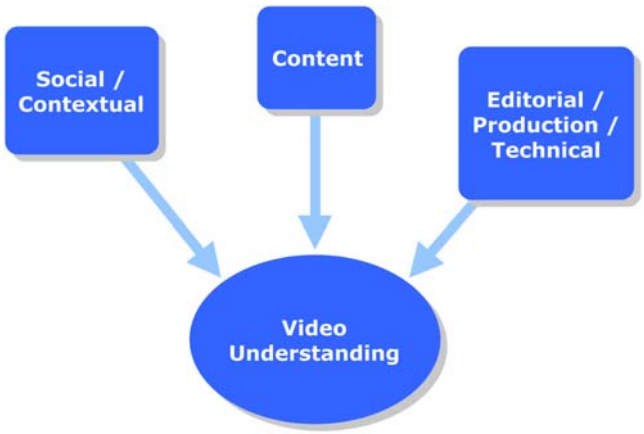

Figure 2. Contributing information spaces.

Event: Video, time and events are inseparable and have to be described in the same breadth. The model describes an Event class as a major content description. Events cannot be described in simple textual descriptions and are not always perceivable. It is a cognitive construct based on some spatio-temporal configuration patterns of objects and their interactions within a time frame. For example, there is no event called "pole vault" directly observable in an image or video but we can construct a "pole vault" event from an image based on the presence of some visual objects such as a person, bar, pole, etc. and their spatial relationships such as "near", "above", etc. The event is characterised by properties such as (Figure 3) *hasParticipant(Event, Agent), hasLocation(Event, Location), hasSubEvent(Event, Event), hasTime(Event, Time), coOccurredWith (Event, Event).*

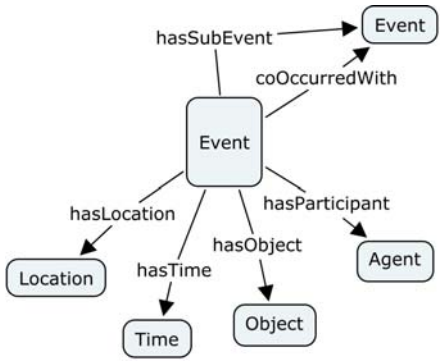

Figure 3. Event class model.

The "*hasLocation*" predicate of an Event object has two sub properties "*hasPhysicalLocation*" and "*hasTimeLineLocation*". Physical location relates the event instance to a *geo:SpatialThing* instance at the same time. Timeline location relates to the temporal location over the video timeline which is described in terms of time interval.

5.1 Core Model

In this section we describe the rest of the content model and its relationships to other existing vocabularies. The types of descriptions required can be clustered into three categories (Figure 2). However, we will note that the focus of description in this paper is the content and context parts of the model, not the editorial or document level properties which are mostly reused from other ontologies. Since a prime focus of this work is to exploit the evolving social context as a metadata contributor, we will also define this aspect in more detail.

5.1.1 Editorial/Production/Technical

This cluster includes the basic descriptions of a media document in terms of its creator, creation and other document-related descriptions such as title, description, date of creation, duration, copyright, director, frame rate, size, actor, genre, etc.

Dublin Core: DC is a framework used to describe Web documents including multimedia items. We have used many editorial attributes from the Dublin Core vocabulary such as "*dc:title*", "*dc:description*", "*dc:date*", "*dc:subject*" to describe various media attributes.

FOAF: The Friend-of-Friend vocabulary is one of the most widely-used Semantic Web vocabularies for describing people and their networks, and is an obvious choice for describing the class "*Person*". *Person* is described as a subclass of "*foaf:Agent*", "*foaf:primaryInterest*" is used to describe the interests of the person and "*foaf:knows*" is used to describe social links between people. A participatory agent of an event can be described as a

[9] http://xmlns.com/foaf/0.1/

[10] http://dublincore.org/

[11] http://sioc-project.org/

foaf:Agent, and similarly a "video has creator" statement can be described as a relation between a "*Video*" and a *foaf:Agent*.

5.1.2 Content Description

An ontology is defined as the representation of concepts, properties and their relationships expressed in linguistic terms for textual data processing. In order to support video annotation and content retrieval, the traditional ontological paradigm should be extended and should include perceptual elements at the signal level such as visual and audio descriptors [6]. The model consists of the following main classes as described in Figure 4.

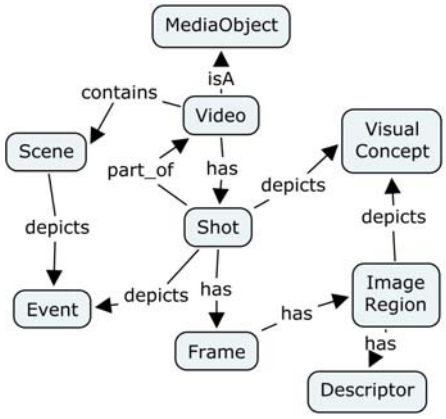

Figure 4. Content model.

MediaObject is a root class which specialises into "Video", "Audio", "Image", "Text" and "Segment".

Segment is a subclass of the MediaObject class and is the result of a temporal or spatial decomposition of the document. Shot , Scene, Frame and ImageRegions are subclasses of Segment.

Descriptor is a superclass of visual, aural, texture, shape, motion and other low-level descriptors. VDO has described visual descriptors conforming to MPEG-7, and we can follow these for visual descriptions of the image and image regions.

Shot is an uninterrupted image sequence of a camera recording and is hence treated as an atomic unit for video analysis. It is a temporal decomposition of video with non-uniform length. It is a subclass of Segment and is linked to a video by a "part_of" relationship. A Shot can be automatically created using various algorithms. Each Shot can be annotated at multiple levels such as audio, visual and textual content.

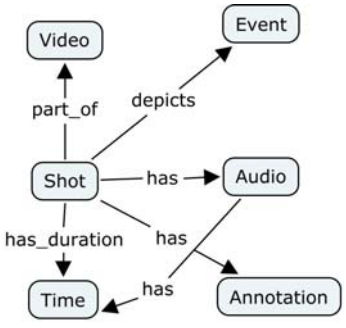

Figure 5. The Shot class model.

Frame is also a subclass of Segment and holds a "part_of" relationship with Shot. One or more Frames are automatically selected to represent a Shot depending on various criteria. A Frame can also be described with an ImageRegion which may be the entire Frame or a small part of it.

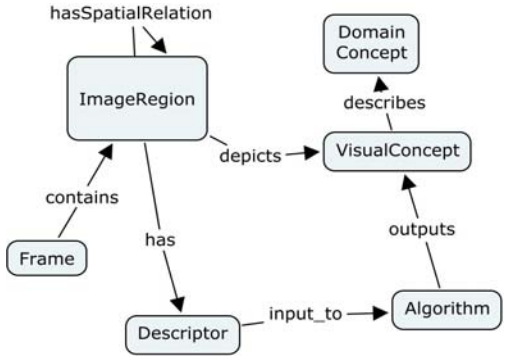

Figure 6. ImageRegion class description.

ImageRegion (r0) is the smallest structural decomposition of the still image. The entire video Frame (f0) or any segment of the frame may be considered as an instance of ImageRegion. An ImageRegion segment depicts a VisualConcept (vc0). Visual Descriptors (d0) (i.e. features from pixel data) are attached to r0. The code snippet below shows the relation between an ImageRegion and a semantic domain concept via a VisualConcept.

```
r0 a ImageRegion
  .partOf(r0,f0)
  .depicts(r0,vc0)
  .ov:describes(vc0,domainConcept)
```

Annotation is a unique user-generated object attached to a media segment (Shot, Frame, ImageRegion). The Annotation object is described with the following attributes:

```
ov:createdBy(ov:Annotation,sioc:User),
ov:startTime(ov:Annotation,Time),
ov:endTime(ov:Annotation,Time).
```

VisualConcept is the representation of a real-world semantic concept which can be described with the attribute "*describes*" to connect to the domain concept. A VisualConcept object can also be described with perceptual attributes such as "*hasDominantColour*", "*hasShape*" and "*coOccuredWith*".

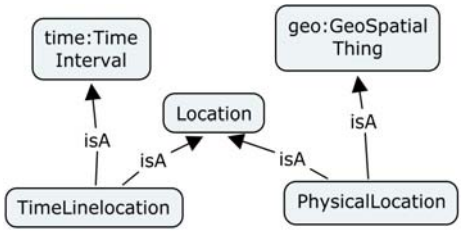

Figure 7. Location class description.

Location values in a multimedia context can be of two different types, *PhysicalLocation* which is an instance of geolocation with geo-coordinates as attributes, and the *TimeLineLocation*. *TimeLineLocation* is a time point within the media stream.

5.1.3 Contextual and Social

Multimedia data is highly contextual in nature. In multimedia systems, being ignorant of context while accessing and trying to understand media is a naïve approach, specifically in the case of web videos. The basic contexts we can immediately utilise are the time of the media recording and the geolocation of the recording. As well as the above two contexts, there is one more category of contextual data which may have a greater contribution towards media understanding: this is the social context, i.e. where the video is distributed, discussed, accessed and shared.

A video (once uploaded to the Web) becomes a shared object amongst its viewers: they are subject to review, feedback, and can be bookmarked and interlinked with other information objects such as websites, blogs, or presentations by various users. These kinds of interactions enrich the video with collective intelligence. Similar videos are grouped under a thematic concept or category. Communities can be organized as a type of special interest group where users share videos with a community of interest. Videos are also subscribed to by various users with shared interests. People talk about videos, their contents and their production values as well. Viewership and ratings reveals the popularity of a video. Viewers can recommend and reuse videos for various purposes. We have captured the above dynamics in our next module called the social module, linking to terms from the well-established online community vocabulary SIOC.

SIOC: Semantically-Interlinked Online Communities is a framework used to describe online communities such as wikis, blogs, forums, mailing lists and other user-generated content in a machine-readable format. Metadata representing social context can be leveraged by applications for information discovery and personalized advertisements, media browsing and sharing, and context-sensitive device management (to name a few). Below we will discuss some of the main classes from the social module.

Group is a community created by users based on their shared interests. The videos part of a group shares some common topics.

UserTag is a concept for representing and describing the semantic content in the video. It can be mapped to a skos:Concept and is linked using the properties sioc:topic or dc:subject. Tags are explicit user intentions (as expressed by free-text keywords) used to identify and understand the content of a resource. The prime objective of creating a UserTag object is to make it more connected and integrated with a discoverable URI. For example, the following code snippet describes a UserTag associated with a video:

```
.a usertag:UserTag
  .usertag:tagText “Coral reef”
  .usertag:createdBy(User)
  .usertag:taggedTo(Video)
  .rdfs:seeAlso
  ”http://dbpedia.org/page/Coral_reef”
```

Response (either text or video) is a feedback item or a comment on a media item or a part of a media item added by various user accounts. A Response can be attached to an entire video or to a segment of a video with time stamps. It has two subclasses: “VideoResponse” and “TextResponse”. TextResponse is a sub-class of *sioc:Post*. A Response can also be described in terms of user sentiments or opinions, e.g. “*hasOpinion*”, whose value may be “positive/negative/neutral”.

User is subclass of *sioc:User*. It is the virtual presence of a *foaf:Agent* (Person, Organization, Group). *User* belongs to a *foaf:Person* and as per the SIOC specification, one person may have multiple user accounts on different sites.

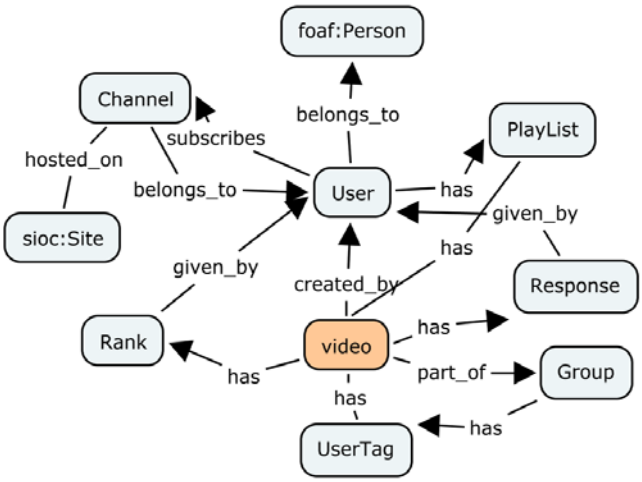

Figure 8. Context model.

6. MODEL USE CASE

This section describes the usability of the model in terms of annotation and retrieval. The complex queries posed by users (as described in the introductory section) can be achieved with the use of the model to describe the video content and its contexts. The model (implemented as a Java API) will generate an RDF graph for a video.

6.1 Use Case 1

Query: Find all videos from 2009 of post-election protests in Iran as uploaded by CNN and Al Jazeera.

Below is a snippet of a possible annotation describing the video and its semantic events.

```
<http://deri.org/sw/video/1.0/Video/IrX5hCX>
rdf:type ov:Video
  .ov:category “News”
  .dc:subject “election”
  .sioc:has_creator <http://youtube.com/user/cnn>
  .ov:depictsEvent Event01
  .ov:hasLocation “Tehran”
  .ov:hasTime “June, 2009”
  .rdfs:label “Protest”
```

Figure 8. RDF snippet of a video event annotation.

SPARQL: W3C have recommended SPARQL[12] as the Semantic Web query language for accessing RDF-based data for complex queries. The code snippet below is the example of the SPARQL query to answer use case 1.

```
PREFIXES…

SELECT DISTINCT ?title
WHERE {
?s rdf:type ov:Video .
?s dc:title ?title .
?s dc:date ?date .
FILTER (?date > 2009)
?s ov:hasCreator ?name .
FILTER regex ( ?name,'CNN','i') && FILTER regex
(?name,'aljazeera','i')
} ORDER BY DESC(?date)}
```

6.2 Use Case 2

Query: Find all the video shots of "Tim Berners-Lee speaking on Linked Data".

Below is a snippet of the video shot annotation describing the shot and its transcript along with a time stamp. The textual description of transcript mentions that Linked Data is being discussed in that particular segment.

```
<ov:video id="01">
<ov:shot id="video01shot03">
  <ov:frame>
  <ov:depicts>
  http://dbpedia.org/resource/Tim_Berners-Lee
  </ov:depicts>
  </ov:frame>

<time:hasBeginning>0:0:05:51</time:hasBeginning>
  <time:hasEnd>0:0:06:32</time:hasEnd>
  <ov:audio>
  <ov:annotation> So I want us now to think about
not just two pieces of data being connected, or
six like he did, but I want to think about a
world where everybody has put data on the web and
so virtually everything you can imagine is on the
web. and then calling that linked data. The
technology is linked data, and it's extremely
simple.</ov:annotation>
    </ov:audio>
  </ov:shot>
</ov:video>
```

Figure 9. Video shot annotation.

The following SPARQL query will extract the segment as a result.

```
PREFIXES…

SELECT DISTINCT ?shot
WHERE
{
?shot rdf:type Shot .
?shot ov:depicts ?people
?People foaf:name "Tim Berners-Lee" .
?shot ov:hasAnnotation ?anno .
?anno ov:annotext "linked data" .
}
```

6.3 Use Case 3

Query: Find all the users who bookmarked videos tagged with 'topicA'.

```
PREFIXES…
select ?username ?title
where
  {
    ?entity rdf:type ov:Video .
    ?entity dc:subject 'topicA' .
    ?entity dc:title ?title .
    ?entity ov:bookmarked_by ?user .
    ?user sioc:username ?username .
 }
```

7. REQUIREMENTS SATISFACTION

We will discuss here whether the requirements mentioned above are satisfied with the proposed model.

Interoperability among existing vocabularies is a core requirement for wide acceptability of any ontology. The ontology has been formalized with OWL-DL which makes it available for interoperation with other widely-used schemas. This model has reused extensively-used classes from other vocabularies wherever necessary and sometimes adds redundancies to cover a wider community, for example, a user-generated tag can be described in various ways: using *dc:subject* (Dublin Core) and as a *sioc:topic* (SIOC).

Ease of Use: The proposed model mainly focuses on maximizing the automatic extraction of metadata either from the content module or from context module. Most of the social context metadata related to a video are produced by users on the Web over time.

Representation of content structure: Our proposed model describes the content structure of the media at different abstraction levels, both temporally as well as spatially. Shots and Frames are temporally segmented, and ImageRegion describes content in spatial dimensions with proper fragment identification.

Modularity and Extensibility: Our proposed model is an aggregation of different modules such as Time, Location, Event, Descriptor, and UserTag. Most of the entities are described as a basic class with scope for further extension. The model can be extended with media types such as 3D models, graphics, and sketches as specialized classes of media object.

Separation of Concern: Content representation and separation of concern is achieved through predicates defined between spatial or temporal segments and their depicted semantic concepts.

Multimedia Reasoning: An ontological model should describe the reasoning scope. Multimedia reasoning includes reasoning at both the content and domain levels. Neuman et al. [12] described that an aggregated composition of parts constrained with spatial and temporal relations will facilitate concept reasoning such as object configuration, occurrences, events and scene interpretation. Our proposed model provides

[12] http://www.w3.org/TR/rdf-sparql-query/

ample scope for the extraction and description of such spatial and temporal segments which can be used for reasoning.

1. *Content-to-Concept:* Lower-level content descriptions of segments (whether visual descriptions or aural descriptions) are the primary input to algorithms used to detect the depicted concept in the segment.
2. Since the model describes both structural and content descriptions, the *part_of* relationship helps in reasoning by providing a means of aggregation for different parts of the content, for example, visual concepts present in several shots during a certain time range can be aggregated together to produce a semantic description of the scene which makes use of many shots.
3. Spatial and geometrical attributes of the ImageRegion class help us to reason about the spatial positioning of different objects in the frame at a point in time. Concept (Sky) *isAbove* Concept (Sea) can be extracted from the ImageRegion descriptions of a frame.
4. Temporal attributes of segments will help in spatio-temporal reasoning.

8. CONCLUSIONS

We have developed a lightweight video content model for web videos satisfying the ontological requirements as laid out by the multimedia community [4]. The most unique approach of the model is its inclusion of emerging social context as part of the video metadata. The ontology is formalised in OWL-DL and is available through a Java API.

In future work, a more robust evaluation framework using the model will be developed. We will also develop a tool for supporting the annotation and retrieval of video content.

9. ACKNOWLEDGEMENTS

The work presented in this paper has been funded in part by Science Foundation Ireland under Grant No.SFI/08/CE/I1380 (Lion-2). Our thanks to Dr. Alexandre Passant for his useful insights regarding this work.

10. REFERENCES

[1] S. Beretti, Alberto Del Bimbo, Enrico Vicario: Efficient Matching and Indexing of Graph Models in Content-Based Retrieval. IEEE Transactions on pattern Analysis and Machine Intelligence 23(10): 1089-1105, 2001.

[2] Dbpedia: A Nucleus for a Web of Open Data. In the Proceedings of 6th International Semantic Web conference (ISWC,2007).

[3] Garcia, R., Celma, O.: Semantic Integration and Retrieval of Multimedia Metadata. In: 5th Int. Workshop on Knowledge Markup and Semantic Annotation. (2005)

[4] Geurts, J., Ossenbruggen, J.v., Hardman, L.: Requirements for practical multimedia annotation. In: Workshop on Multimedia and the Semantic Web. (2005)

[5] Hunter, J.: Adding Multimedia to the Semantic Web - Building an MPEG-7 Ontology. In: 1st Int. Semantic Web Working Symposium. (2001) 261-281

[6] H. Ghosh, S. Chaudhury, K. Kashyap, and B. Maiti. Ontology Specification and Integration for Multimedia Applications, chapter 9. Springer, 2006.

[7] Huang, W, Eze, E., Webster D. (2006) "Integrating Semantics of Multi- Media Resources and Processes in e-Learning", ACM/Springer Journal on Multimedia systems, Special Issue on Educational Multimedia Systems.

[8] Jaimes, A., and Smith, J. Semi-automatic, data-driven construction of multimedia ontologies. In *Proceedings of IEEE Int'l Conference on Multimedia & Expo*, 2003.

[9] MPEG-7: Multimedia Content Description Interface. ISO/IEC 15938 (2001)

[10] M. Petkovic and W. Jonker, "A framework for video modeling", in IASTED International Conference on Applied Informatics, Innsbruck, Autriche, Février 2000.

[11] Naphade, M.: Large-scale concept ontology for multimedia. IEEE MultiMedia, 13(3):86–91, (2006)

[12] Neumann, B., M oller, R.: On Scene Interpretation with Description Logics. In:Cognitive Vision Systems. Springer (2006) 247-275.

[13] OWL, Web Ontology Language Reference Version 1.0, W3C Recommendation (2004),http://www.w3.org/TR/owl-ref/

[14] Rattenbury, T., Good, N., Naaman, M.: Towards automatic extraction of event and place semantics from Flickr tags. In Proc. of SIGIR, pages 103-110(2007)

[15] Reidsma, D. Kuper, ,J. , Declerck, T. , Saggion, H. and Cunningham,H. Cross document ontology based information extraction for multimedia retrieval. In Supplementary proc. of the ICCS03, Dresden, July 2003.

[16] Troncy,R., Richard,A., Staab,s.,Hardman,L.& vacura,M. COMM: Designing a Well-Founded Multimedia Ontology for the Web. In 6th International Semantic Web Conference (ISWC'2007), Busan, Korea, November 11-15, 2007.

[17] RDF, Resource Description Framework Primer, W3C Recommendation (2004), http://www.w3.org/TR/rdf-primer/

[18] S. Beretti, Alberto Del Bimbo, Enrico Vicario: Efficient Matching and Indexing of Graph Models in Content-Based Retrieval. IEEE Transactions on pattern Analysis and Machine Intelligence 23(10): 1089-1105, 2001.

[19] Staab, S., and Studer, R. *Handbook on Ontologies*. International Handbooks on Information Systems, Springer Verlag, Heidelberg, 2004.

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

[21] http://www.w3.org/2000/01/rdf-schema#

[22] http://www.w3.org/TR/owl-time/

[23] http://www.w3.org/2008/WebVideo/Fragments/WD-media-fragments-spec/