A Survey of Semantic Image and Video Annotation Tools

S. Dasiopoulou, E. Giannakidou, G. Litos, P. Malasioti, and I. Kompatsiaris

Multimedia Knowledge Laboratory, Informatics and Telematics Institute,
Centre for Research and Technology Hellas
{dasiop,igiannak,litos,xenia,ikom}@iti.gr

Abstract. The availability of semantically annotated image and video assets constitutes a critical prerequisite for the realisation of intelligent knowledge management services pertaining to realistic user needs. Given the extend of the challenges involved in the automatic extraction of such descriptions, manually created metadata play a significant role, further strengthened by their deployment in training and evaluation tasks related to the automatic extraction of content descriptions. The different views taken by the two main approaches towards semantic content description, namely the Semantic Web and MPEG-7, as well as the traits particular to multimedia content due to the multiplicity of information levels involved, have resulted in a variety of image and video annotation tools, adopting varying description aspects. Aiming to provide a common framework of reference and furthermore to highlight open issues, especially with respect to the coverage and the interoperability of the produced metadata, in this chapter we present an overview of the state of the art in image and video annotation tools.

1 Introduction

Accessing multimedia content in correspondence with the meaning pertained to a user, constitutes the core challenge in multimedia research, commonly referred to as the *semantic gap* [1]. The current state of the art in automatic content analysis and understanding supports in many cases the successful detection of semantic concepts, such as persons, buildings, natural scenes vs manmade scenes, etc. at a satisfactory level of accuracy; however, the attained performance remains highly variable when considering general domains, or when increasing, even slightly, the number of supported concepts [2–4]. As a consequence, the manual generation of content descriptions holds an important role towards the realisation of intelligent content management services. This significance is further strengthened by the need for manually constructed descriptions in automatic content analysis both for evaluation as well as for training purposes, when learning based on pre-annotated examples is used.

The availability of semantic descriptions though is not adequate per se for the effective management of multimedia content. Fundamental to information sharing, exchange and reuse, is the interoperability of the descriptions at both syntactic and semantic levels, i.e. regarding the valid structuring of the descriptions and the endowed meaning respectively. Besides the general prerequisite for interoperability, additional requirements arise from the multiple levels at which multimedia content can be represented including structural and low-level features information. Further description levels induce from more generic aspects such as authoring & access control, navigation, and user history & preferences. The strong relation of structural and low-level feature information to the tasks involved in the automatic analysis of visual content, as well as to retrieval services, such as transcoding, content-based search, etc., brings these two dimensions to the foreground, along with the subject matter descriptions.

Two initiatives prevail the efforts towards machine processable semantic content metadata, the Semantic Web activity[1] of the W3C and ISO's Multimedia Content Description Interface[2] (MPEG-7) [5, 6], delineating corresponding approaches with respect to multimedia semantic annotation [7, 8]. Through a layered architecture of successively increased expressivity, the Semantic Web (SW) advocates formal semantics and reasoning through logically grounded meaning. The respective rule and ontology languages embody the general mechanisms for capturing, representing and reasoning with semantics. They do not capture application specific knowledge. In contrast, MPEG-7 addresses specifically the description of audiovisual content and comprises not only the representation language, in the form of the Description Definition Language (DDL), but also specific, media and domain, definitions; thus from a SW perspective, MPEG-7 serves the twofold role of a representation language and a domain specific ontology.

Overcoming the syntactic and semantic interoperability issues between MPEG-7 and the SW has been the subject of very active research in the current decade, highly motivated by the complementary aspects characterising the two aforementioned metadata initiatives: media specific, yet not formal, semantics on one hand, and general mechanisms for logically grounded semantics on the other hand. A number of so called *multimedia ontologies* [9–13] issued in an attempt to add formal semantics to MPEG-7 descriptions and thereby enable linking with existing ontologies and the semantic management of existing MPEG-7 metadata repositories. Furthermore, initiatives such the W3C Multimedia Annotation on the Semantic Web Taskforce[3], the W3C Multimedia Semantics Incubator Group[4] and the Common Multimedia Ontology Framework[5], have been established to address the technologies, advantages and open issues related to the creation, storage, manipulation and processing of multimedia semantic metadata.

In this chapter, bearing in mind the significance of manual image and video annotation in combination with the different possibilities afforded by the SW and MPEG-7 initiatives, we present a detailed overview of the most well known

[1] http://www.w3.org/2001/sw/

[2] http://www.chiariglione.org/mpeg/

[3] http://www.w3.org/2001/sw/BestPractices/MM/

[4] http://www.w3.org/2005/Incubator/mmsem/

[5] http://www.acemedia.org/aceMedia/reference/multimedia_ontology/index.html

manual annotation tools, addressing both functionality aspects, such as coverage & granularity of annotations, as well as interoperability concerns with respect to the supported annotation vocabularies and representation languages. Interoperability though does not address solely the harmonisation between the SW and MPEG-7 initiatives; a significant number of tools, specially regarding video annotation, follow customised approaches, aggravating the challenges. As such, this survey serves a twofold role; it provides a common framework for reference and comparison purposes, while highlighting issues pertaining to the communication, sharing and reuse of the produced metadata.

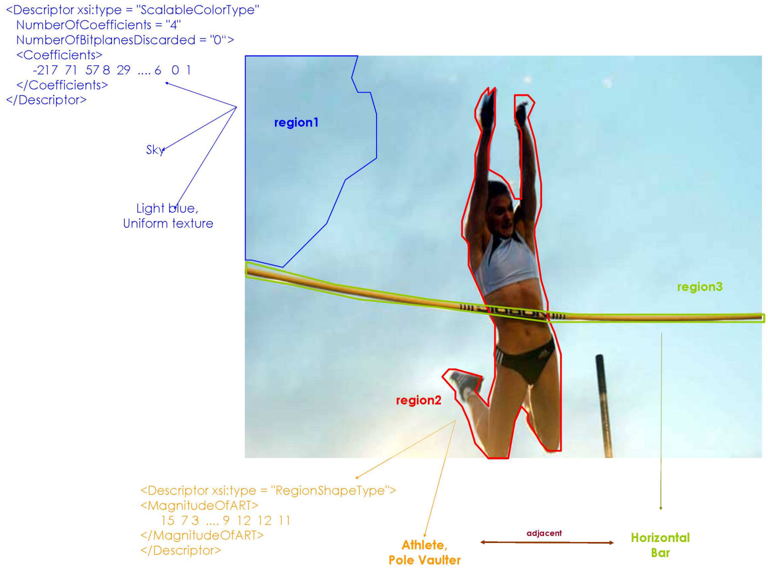
The rest of the chapter is organised as follows. Section 2 describes the criteria along which the assessment and comparison of the examined annotation tools is performed. Sections 3 and 4 discuss the individual image and video tools respectively, while Section 5 concludes the paper, summarising the resulting observations and open issues.

2 Semantic Image and Video Annotation

Image and video assets constitute extremely rich information sources, ubiquitous in a wide variety of diverse applications and tasks related to information management, both for personal and professional purposes. Inevitably, the value of the endowed information amounts to the effectiveness and efficiency at which it can be accessed and managed. This is where semantic annotation comes in, as it designates the schemes for capturing the information related to the content.

As already indicated, two crucial requirements featuring content annotation are the interoperability of the created metadata and the ability to automatically process them. The former encompasses the capacity to share and reuse annotations, and by consequence determines the level of seamless content utilisation and the benefits issued from the annotations made available; the latter is vital to the realisation of intelligent content management services. Towards their accomplishment, the existence of commonly agreed vocabularies and syntax, and respectively of commonly agreed semantics and interpretation mechanisms, are essential elements.

Within the context of visual content, these general prerequisites incur more specific conditions issuing from the particular traits of image and video assets. Visual content semantics, as multimedia semantics in general, comes into a multilayered, intertwined fashion [14, 15]. It encompasses, amongst others, thematic descriptions addressing the subject matter depicted (scene categorisation, objects, events, etc.), media descriptions referring to low-level features and related information such as the algorithms used for their extraction, respective parameters, etc., as well as structural descriptions addressing the decomposition of content into constituent segments and the spatiotemporal configuration of these segments. As in this chapter semantic annotation is investigated mostly with respect to content retrieval and analysis tasks, aspects addressing concerns related to authoring, access and privacy, and so forth, are only shallowly treated.



```
<Descriptor xsi:type = "ScalableColorType"
  NumberOfCoefficients = "4"
  NumberOfBitplanesDiscarded = "0">
  <Coefficients>
     -217  71  57 8  29  .... 6   0  1
  </Coefficients>
</Descriptor>
```

```
<Descriptor xsi:type = "RegionShapeType">
<MagnitudeOfART>
    15  7 3  .... 9  12  12  11
</MagnitudeOfART>
</Descriptor>
```

Fig. 1. Multi-layer image semantics.

Figure 1 shows such an example, illustrating subject matter descriptions such as "Sky" and "Pole Vaulter, Athlete", structural descriptions such as the three identified regions, the spatial configuration between two of them (i.e. region2 above region3), and the ScalableColour and RegionsShape descriptor values extracted for two regions. The different layers correspond to different annotation dimensions and serve different purposes, further differentiated by the individual application context. For example, for a search and retrieval service regarding a device of limited resources (e.g. PDA, mobile phone), content management becomes more effective if specific temporal parts of video can be returned to a query rather than the whole video asset, leaving the user with the cumbersome task of browsing through it, till reaching the relative parts and assessing if they satisfy her query.

The aforementioned considerations intertwine, establishing a number of dimensions and corresponding criteria along which image and video annotation can be characterised. As such, interoperability, explicit semantics in terms of liability to automated processing, and reuse, apply both to all types of description dimensions and to their interlinking, and not only to subject matter descriptions, as is the common case for textual content resources.

In the following, we describe the criteria along which we overview the different annotation tools in order to assess them with respect to the aforementioned considerations. Criteria addressing concerns of similar nature have been grouped together, resulting in three categories.

2.1 Input & Output

This category includes criteria regarding the way the tool interacts in terms of requested / supported input and the output produced.

- Annotation Vocabulary. Refers to whether the annotation is performed according to a predefined set of terms (e.g. lexicon / thesaurus, taxonomy, ontology) or if it is provided by the user in the form of keywords and free text. In the case of controlled vocabulary, we differentiate the case where the user has to explicitly provide it (e.g. as when uploading a specific ontology) or whether it is provided by the tool as a built-in; the formalisms supported for the representation of the vocabulary constitute a further attribute. We note that annotation vocabularies may refer not only to subject matter descriptions, but as well to media and structural descriptions. Naturally, the more formal and well-defined the semantics of the annotation vocabulary, the more opportunities for achieving interoperable and machine understandable annotations.

- Metadata Format. Considers the representation format in which the produced annotations are expressed. Naturally, the output format is strongly related to the supported annotation vocabularies. As will be shown in the sequel though, where the individual tools are described, there is not necessarily a strict correspondence (e.g. a tool may use an RDFS[6] or OWL[7] ontology as the subject matter vocabulary, and yet output annotations in RDF[8]). The format is equally significant to the annotation vocabulary as with respect to the annotations interoperability and sharing.

- Content Type. Refers to the supported image/video formats, e.g. jpg, png, mpeg, etc.

2.2 Annotation Level

This category addresses attributes of the annotations per se. Naturally, the types of information addressed by the descriptions issue from the intended context of usage. Subject matter annotations, i.e. thematic descriptions with respect to the depicted objects and events, are indispensable for any application scenario addressing content-based retrieval at the level of meaning conveyed. Such retrieval may address concept-based queries or queries involving relations between concepts, entailing respective annotation specifications. Structural information is crucial for services where it is important to know the exact content parts associated with specific thematic descriptions, as for example in the case of semantic transcoding or enhanced retrieval and presentation, where the parts of interest can be indicated in an elaborated manner. Analogously, annotations intended for

[6] http://www.w3.org/TR/rdf-schema/

[7] http://www.w3.org/TR/owl-features/

[8] http://www.w3.org/RDF/

training purposes need to include low-level features descriptions and moreover to provide support for their linking with domain notions. Similarly, administrative descriptions may or may not be of significance. To capture the aforementioned considerations, the following criteria have been used.

- Metadata Type. Refers to the annotation dimension. For the purposes of this overview, we identify the following types:
  - content descriptive metadata addressing subject matter information,
  - structural metadata describing spatial, temporal and spatioteporal decomposition aspects
  - media metadata referring to low-level features, and
  - administrative, covering descriptions regarding the creation date of the annotation, the annotation creator, etc.

- Granularity. Specifies whether the annotation describes the content assets as a whole or whether it refers to specific parts of it.
  - For image assets, annotation may refer to the whole image, usually termed as scene or global level annotation, or it may refer to specific spatial segments, for which case the terms region-based, local and segment-based annotation are commonly used
  - For video assets, annotation may refer to the entire video, temporal segments (shots), frames (temporal segments with zero duration), regions within frames, or even to moving regions, i.e. a region followed for a sequence of frames. It worths noting that due to the more complex structural patterns applicable for video, many tools besides the annotation functionality provide corresponding visualisation functionalities through the use of timelines. Thereby, the associations of subject matter annotations with respect to the video structure can be easily inspected.

- Localisation. This criterion relates to the supported granularity, and refers to the way in which a part of interest is localised within a content asset. We discriminate two cases with respect to whether localisation is performed automatically (through some segmentation or shot detection algorithm embedded in the tool) or whether manual drawing services are provided.

- Annotation expressivity. Refers to the level of expressivity supported with respect to the annotation vocabulary. For example, in the case an ontology is used for subject matter descriptions, some tools may support only concept based annotation, while others enable to create annotations representing relations among concepts as well.

2.3 Miscellaneous

This category summarises additional criteria that do not fall under the previous dimensions. The considered aspects relate mostly to attributes of the tool itself rather than of the annotation process. As such, and given the scope of this chapter, in the description of the individual tools that follows in the two subsequent Sections, these criteria are treated very briefly.

– Application Type: Specifies whether the tool constitutes a web-based or a stand-alone application.
– Licence: Specifies the kind of licence condition under which the tool operates, e.g. open source, etc.
– Collaboration: Specifies whether the tool supports concurrent annotations (referring to the same media object) by multiple users or not.

3 Tools for Semantic Image Annotation

In this Section we describe prominent semantic image annotation tools with respect to the dimensions and criteria outlined in Section 2. As will be illustrated in the following, Semantic Web technologies have permeated to a considerable degree the representation of metadata, with the majority of tools supporting ontology-based subject matter descriptions, while a considerable share of them adopts ontological representation for structural annotations as well. In order to provide a relative ranking with respect to SW compatibility, we order the tools according to the extend to which the produced annotations bear formal semantics.

3.1 KAT

The K-Space Annotation Tool[9] (KAT), developed within the K-Space[10] project, implements an ontology-based framework for the semantic annotation of images. Figure 2 depicts a screenshot using the KAT 0.2.1 release to annotate the pole vaulter and pole regions in an image depicting a pole vault attempt.

KAT's annotation framework [16] is based on the Core Ontology of MultiMedia (COMM) [13]. COMM extends the *Descriptions & Situations (D&S)* and *Ontology of Information Objects (OIO)* design patterns of DOLCE [17, 18], while incorporating re-engineered definitions of MPEG-7 description tools[19, 20]. As such, COMM models the various annotation levels and their linking (e.g. of descriptive and structural annotations), while providing MPEG-7 based structural and media descriptions of formal semantics.

KAT currently supports descriptive and structural annotations. A user loaded ontology provides the vocabulary and semantics for the subject matter descriptions. The latter are strictly concept based, i.e. considering the aforementioned annotation example it is not possible to annotate the pole as being next to the pole vaulter, and may refer to the entire image or to specific regions of it. The localisation of image regions is performed manually, using either of the rectangle and polygon drawing tools. COMM provides the definitions for the structural

[9] htpps://launchpad.net/kat
[10] http://www.k-space.eu/

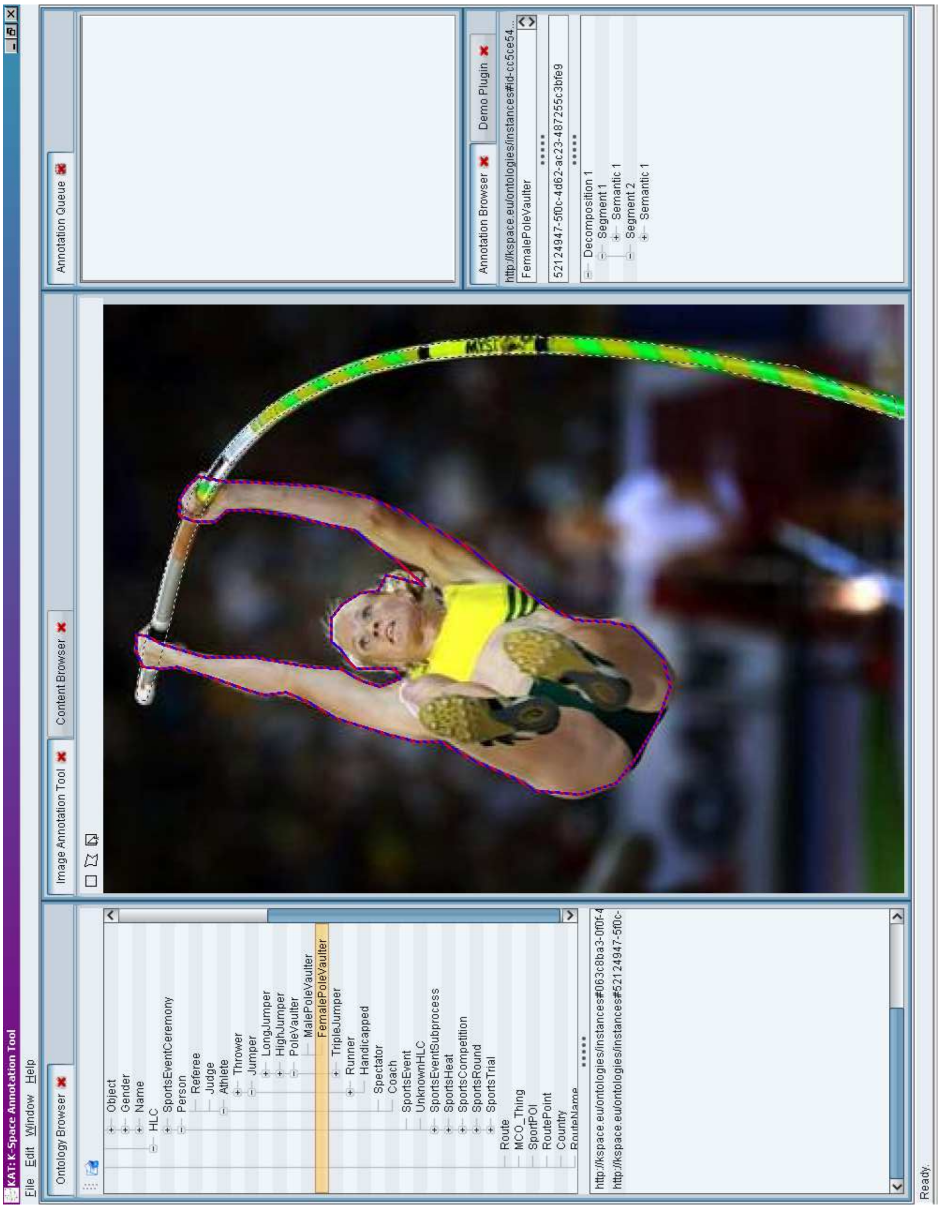


Fig. 2. Example image annotation using KAT.

and localisation semantics, leaving them hidden to the user. The supported input ontology languages include RDFS and OWL, and the produced annotations are in OWL.

It should be noted that the COMM based annotation framework implemented by KAT is media independent, i.e. additional content types can be supported as long as respective media management functionalities (e.g. video player) are included. Furthermore, the COMM based annotation scheme renders quite straightforward the extension of the annotation dimensions supported by KAT. For example, COMM provides means to represent low-level features and additionally to associate them with the corresponding extraction algorithm and its parameters. Thus, assuming the availability of descriptor extraction capability, KAT could support media annotations as well.

3.2 PhotoStuff

PhotoStuff[11], developed by the Mindswap group[12], is an ontology-based image annotation tool that supports the generation of semantic image descriptions with respect to the employed ontologies. Figure 3 illustrates a screenshot of PhotoStuff 3.33 Beta, used during this overview; following the previous example, two regions have been annotated: the one depicting the female pole vaulter localised using a rectangle and the one depicting the pole, for whose localisation a polygon has been used.

PhotoStuff [21] addresses primarily two types of metadata, namely descriptive and structural. Regarding descriptive annotations, the user may load one or multiple domain-specific ontologies from the web or from the local hard drive, while with respect to structural annotations, two internal, hidden to the user, ontologies are used: the Digital-Media[13] ontology and the Technical[14] one. The two ontologies model the different multimedia content and multimedia segments types in accordance with the MPEG-7 specifications. Furthermore, they provide a simple schema for linking content instances (or parts of it) with the depicted domain-specific instances and its respective low-level descriptors. Specifically, the *depicts* property of FOAF[15] and its inverse, i.e. *depiction*, are used to link a media instance to the depicted content and vice versa, while the properties *descriptor* and *visualDescriptor* provide connection with low-level descriptors. However, nor the representation neither the extraction of such descriptors is addressed.

It is worth noticing that the modeling of content structure reminds a simplified version of well known multimedia ontologies, including Hunter's [9], the acemedia Multimedia Content[16] ontology and the Rhizomik ontology [11]. Specifically, only part of the content and segment class hierarchy has been retained,

[11] http://www.mindswap.org/2003/PhotoStuff/
[12] http://www.mindswap.org/
[13] http://www.w3.org/2004/02/image-regions♯
[14] http://www.mindswap.org/ glapizco/technical.owl♯
[15] http://xmlns.com/foaf/0.1/
[16] http://www.acemedia.org/aceMedia/results/ontologies.html

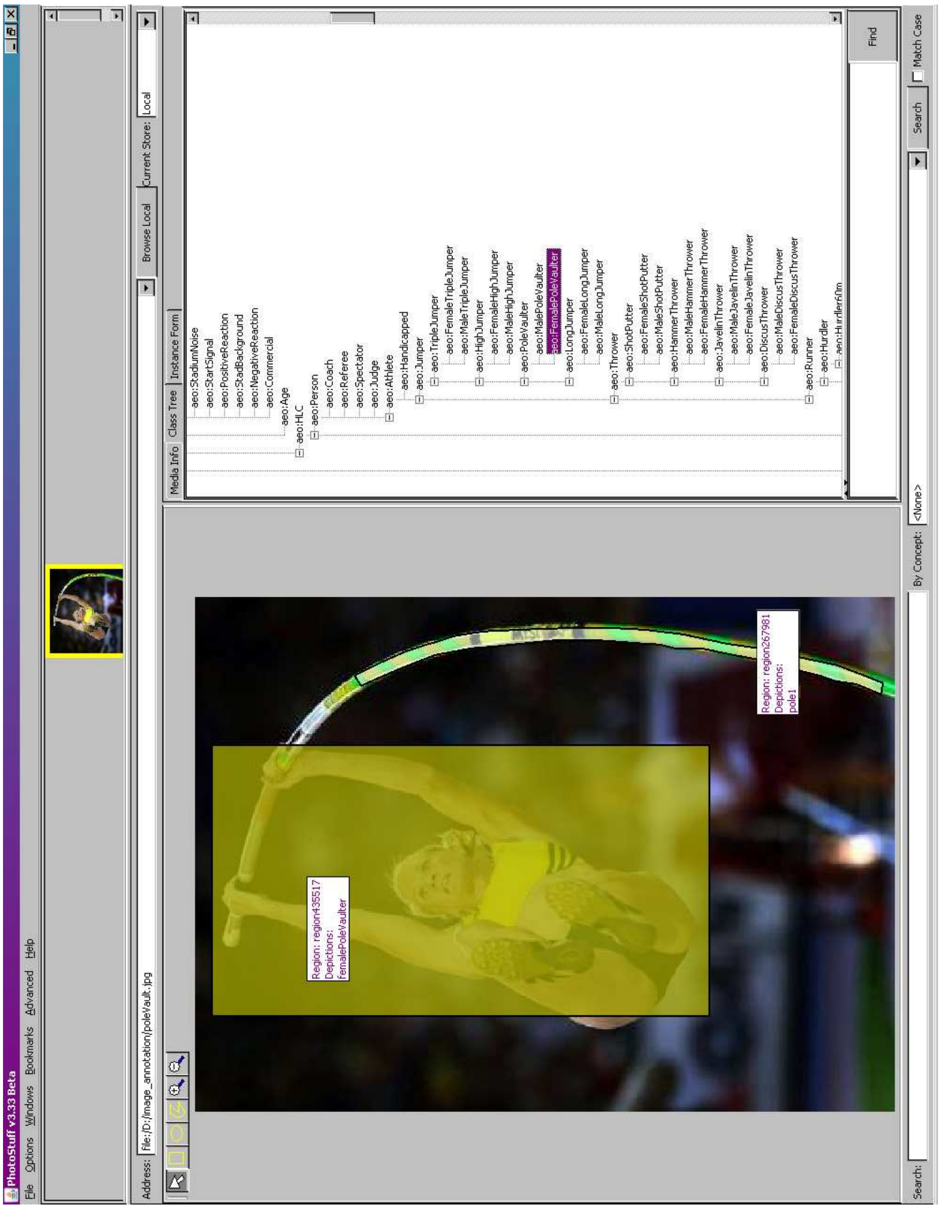


Fig. 3. Example image annotation using PhotoStuff.

in combination with a minimal set of decomposition and localisation properties, such as the properties *regionOf*, *startFrame* and *coords*.

As aforementioned, additional types of metadata can be addressed as long as an appropriate ontology is loaded. For example, authoring metadata can be generated if the Dublin Core[17] element set is used in addition to the domain-specific ontologies. The supported ontology languages are OWL and RDF/RDFS, while the generated annotations are expressed in RDF. Annotations can be attached to either the entire image or to specific regions, using one of the available drawing tools, that is circle, rectangle, and polygon (as an approximation to free hand drawing). Notably, annotations may refer not only to concept instantiations, but also to relations between concept instances already identified in an image. As additional functionalities, PhotoStuff allows keyword-based search through the generated semantic annotations, editing of previously created annotations, as well as parsing and translation of embedded media metadata such as EXIF[18] and IPTC[19].

3.3 AktiveMedia

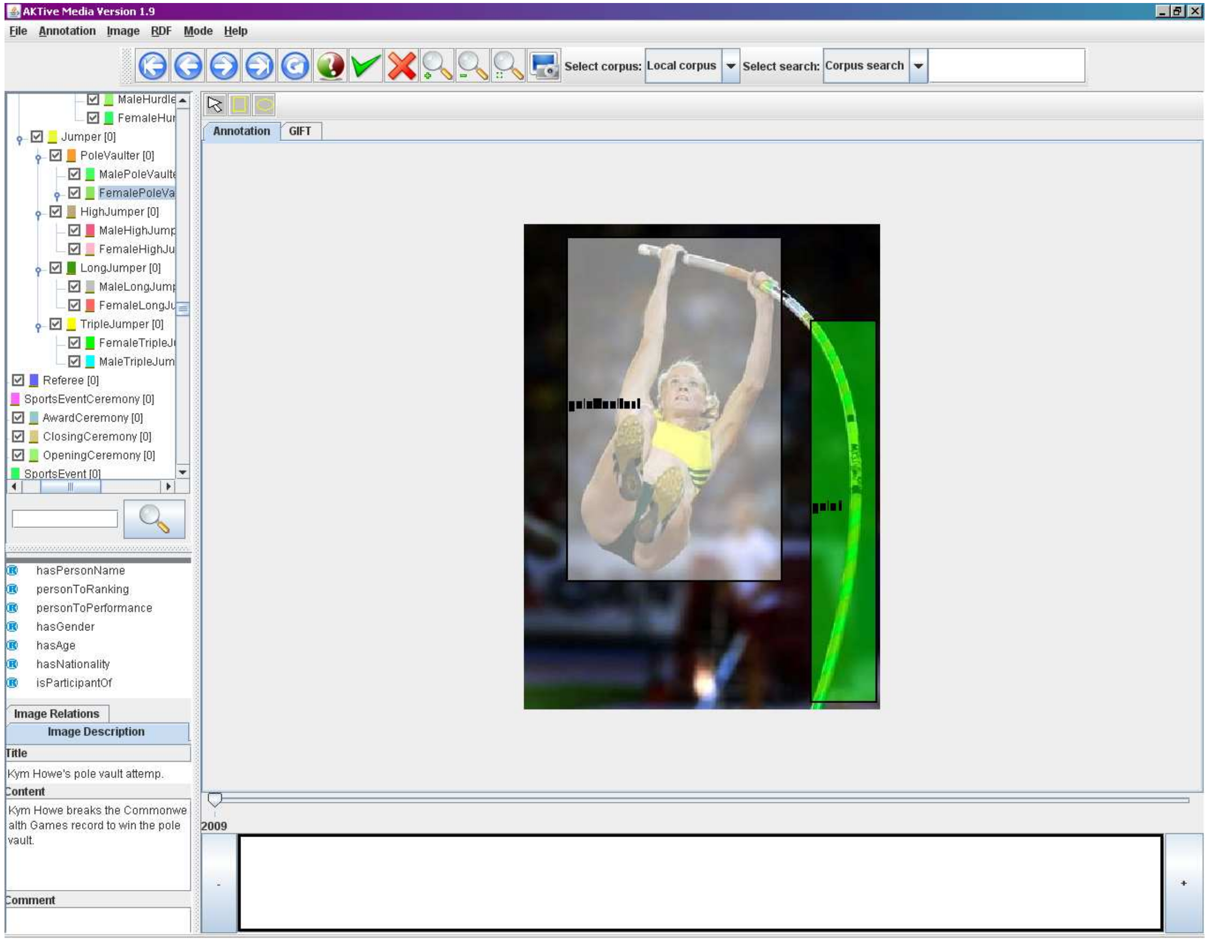
AktiveMedia[20], developed within AKT[21] and X-Media[22] projects, is an ontology-based cross-media annotation system addressing text and image assets. Figure 4 illustrates a screenshot of the image annotation mode for the AktiveMedia 1.9 release, for the previously considered pole vault annotation example.

In image annotation mode, AktiveMedia supports descriptive metadata with respect to user selected ontologies, stored in the local hard drive [22]. Multiple ontologies can be employed in the annotation of a single image; unlike PhotoStuff though, a single ontology is displayed each time in the ontology browser. AktiveMedia provides also localisation metadata through a simple built-in schema that defines corresponding properties for the representation of coordinates, as well as the linking of media-specific to domain-specific instances through a *hasAnnotation* property.

Annotations can refer to image or region level. To describe an entire image, AktiveMedia provides three free text fields, namely title, content and comment. Utilising the text mode, the respective user entered descriptions can be subsequently annotated with respect to an ontology. Region based annotations are associated to either rectangular or circular regions of the image, and are directly associated with a domain-specific concept.

The supported ontology languages include RDFS and OWL, as well as older semantic web languages such as DAML and DAML-ONT; RDF is used for the representation of the generated annotations. Contrary to Photostuff which uses

[17] http://dublincore.org/documents/dces/
[18] http://www.digicamsoft.com/exif22/exif22/
[19] http://www.iptc.org/
[20] http://www.dcs.shef.ac.uk/ ajay/html/cresearch.html
[21] http://www.aktors.org/akt/
[22] http://www.x-media-project.org/



Select corpus: Local corpus

Select search: Corpus search

Title

Kym Howe's pole vault attemp.

Content

Kym Howe breaks the Commonwealth Games record to win the pole vault.

Comment

Fig. 4. Example image annotation using AktiveMedia.

URIs to identify the class to which an instance belongs, AktiveMedia explicitly models the ontology to which the descriptive annotations refer through a *usesOntology* property, and nests correspondingly the values of *hasConcept* and *hasAnnotationText*, i.e. the class and corresponding instance names. As such, the semantics of generated RDF metadata, i.e. the annotation semantics as it entails from the respective ontology definitions, are not direct but require additional processing to retrieve and to reason over.

An interesting feature of AktiveMedia, though not directly related to the task of image annotation, is its ability to learn during textual annotation mode, so that suggestions can be subsequently made to the user, thus realising semi-automatic text annotation. Such facility can prove beneficial when considering the free text and keyword annotations that a user may enter when annotating an image as a whole.

3.4 M-OntoMat-Annotizer

M-Ontomat-Annotizer[23], developed within the aceMedia[24] project, enables the ontology-based representation of associations between domain specific concepts and their respective low-level visual descriptors. Figure 5 illustrates a screenshot of the latest release, namely v0.60, where in the context of the pole vault annotation example, selected descriptors have been extracted and associated to the female pole vaulter and pole instances.

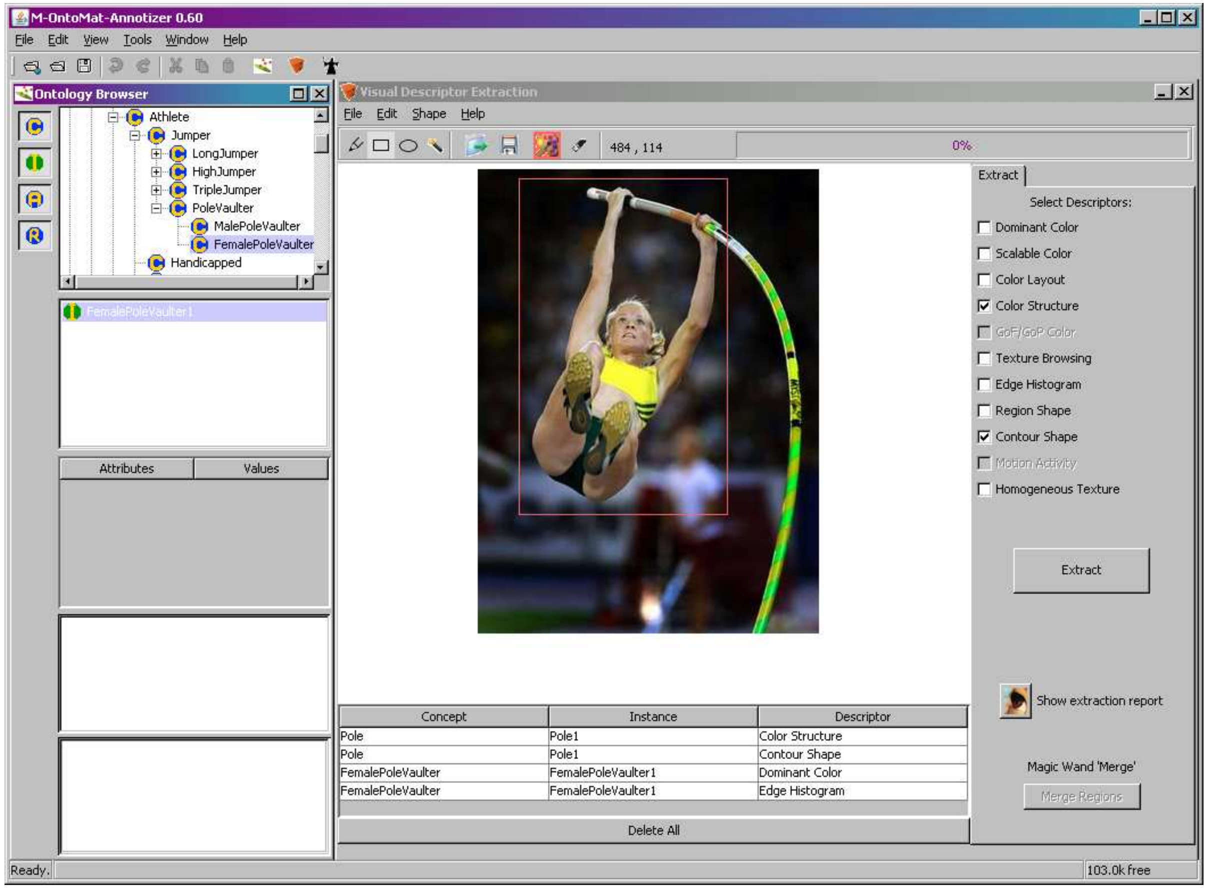
In order to formalise the linking of domain concepts with visual descriptors, M-Ontomat-Annotizer [23] employs the Visual Annotation Ontology (VAO) and the Visual Descriptor Ontology (VDO) [24], both hidden to the user. The VAO serves as a meta-ontology allowing to model domain specific instances as prototype instances and to link them to respective descriptor instances through the *hasDescriptor* property. The VDO[25] models in RDFS the core MPEG-7 visual descriptors (i.e. colour, texture, shape, motion, and localisation)[20]. As in the previous cases, the domain specific instances are in accordance with the domain ontology loaded by the user.

The domain specific instances, and by analogy the extracted descriptor instances, may refer to a specific region or to the entire image. For the identification of a specific region the user may either make use of the automatic segmentation functionality provided by the M-Ontomat-Annotizer or use one of the manually drawing tools, namely the predefined shapes (rectangle and ellipse), free hand and magic wand. To further facilitate the identification of the intended image parts, region merging is also supported. Thereby under-segmentation phenomena can be alleviated, while the annotation of compound objects becomes significantly faster (e.g. merging a face and body region to create a person annotation).

[23] http://www.acemedia.org/aceMedia/results/software/m-ontomat-annotizer.html

[24] http://www.acemedia.org/aceMedia

[25] http://www.acemedia.org/aceMedia/files/software/m-ontomat/acemedia-visual-descriptor-ontology-v09.rdfs



Select Descriptors:

- ☐ Dominant Color
- ☐ Scalable Color
- ☐ Color Layout
- ☑ Color Structure
- ☐ GoF/GoP Color
- ☐ Texture Browsing
- ☐ Edge Histogram
- ☐ Region Shape
- ☑ Contour Shape
- ☐ Motion Activity
- ☐ Homogeneous Texture

Concept	Instance	Descriptor
Pole	Pole1	Color Structure
Pole	Pole1	Contour Shape
FemalePoleVaulter	FemalePoleVaulter1	Dominant Color
FemalePoleVaulter	FemalePoleVaulter1	Edge Histogram

Fig. 5. Example image annotation using M-Ontomat-Annotizer.

The supported input ontology languages are RDFS and DAML, while the generated annotations are in RDFS. It should be noted that compared to PhotoStuff which provides a corresponding *hasDescriptor* property, M-Ontomat-Annotizer provides in addition both the means to extract descriptors and an ontology to formally represent them. However, it lacks structural descriptions, i.e. explicit representation of spatial decomposition instances and direct descriptive annotations. In a following release within the K-Space project, M-Ontomat 2.0[26] provides support for descriptive and structural annotations in the typical semantic search and retrieval sense.

3.5 Caliph

Caliph[27] is an MPEG-7 based image annotation tool that supports all types of MPEG-7 metadata among which descriptive, structural, authoring and low-level visual descriptor annotations. In combination with Emir, they support content-based retrieval of images using MPEG-7 descriptions. Figure 6 illustrates two screenshots corresponding to the generic image information and the semantic (descriptive) annotation tabs.

Contrary to the aforementioned tools, Caliph allows descriptive annotations only at image level [25]. The descriptions may be either in the form of free text or structured, in accordance to the SemanticBase description tools provided by MPEG-7 (i.e. Agents, Events, Time, Place and Object annotations [26]). The so called semantic tab (illustrated at the right part of Figure 6) allows for the latter, offering a graph based interface. A subset of the relations specified in MPEG-7 are available; it is not clear though how to extend them, while additional issues emerge to users unfamiliar with MPEG-7 tools with respect to which relations and how should be used.

3.6 SWAD

SWAD[28] is an RDF-based image annotation tool that was developed within the SWAD-Europe project[29]. The latter ran from May 2002 to October 2004 and aimed to support the Semantic Web initiative in Europe through targeted research, demonstrations and outreach activities. Although the SWAD tool [27] has not been maintained since, we chose to provide a very brief description here for the purpose of illustrating image annotation in the Semantic Web as envisaged and realised by that time, as a reference and comparison point for the various image annotation tools that have been developed afterwards.

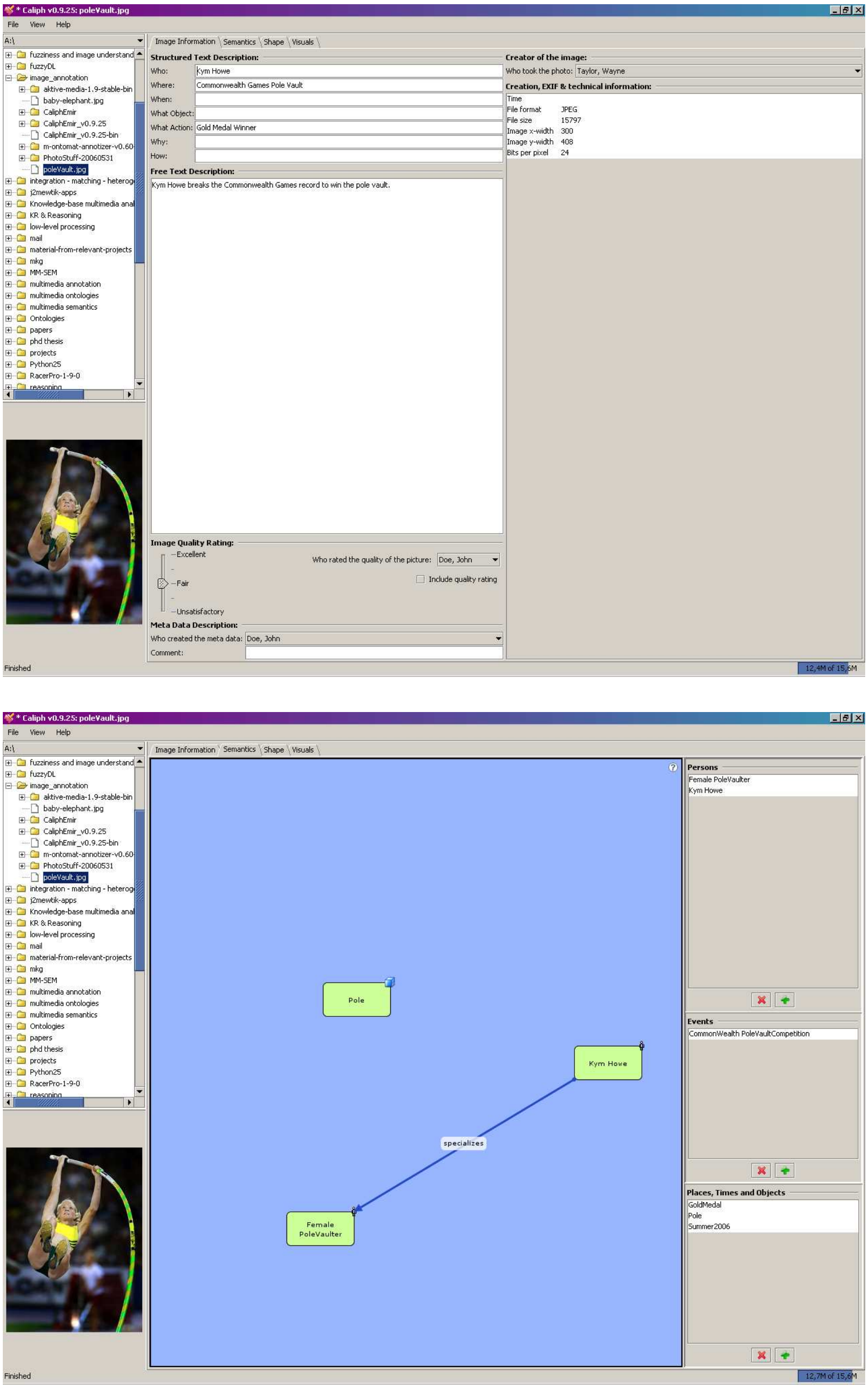
Figure 7 illustrates a screenshot of SWAD's web-based interface. Different tabs allow to insert descriptions regarding who or what is depicted in the image (person, object, event), when and where it was taken, and additional creator and

[26] http://mklab.iti.gr/m-onto2

[27] http://www.semanticmetadata.net/features/

[28] http://swordfish.rdfweb.org/discovery/2004/03/w3photo/annotate.html♯

[29] http://www.w3.org/2001/sw/Europe/



- fuzziness and image understand
- fuzzyDL
- image_annotation
  - aktive-media-1.9-stable-bin
  - baby-elephant.jpg
  - CaliphEmir
  - CaliphEmir_v0.9.25
  - CaliphEmir_v0.9.25-bin
  - m-ontomat-annotizer-v0.60
  - PhotoStuff-20060531
  - poleVault.jpg
- integration - matching - heterog
- j2mewtik-apps
- Knowledge-base multimedia anal
- KR & Reasoning
- low-level processing
- mail
- material-from-relevant-projects
- mkg
- MM-SEM
- multimedia annotation
- multimedia ontologies
- multimedia semantics
- Ontologies
- papers
- phd thesis
- projects
- Python25
- RacerPro-1-9-0
- reasoning

Structured Text Description:

Who: Kym Howe

Where: Commonwealth Games Pole Vault

When:

What Object:

What Action: Gold Medal Winner

Why:

How:

Free Text Description:

Kym Howe breaks the Commonwealth Games record to win the pole vault.

Creator of the image:

Who took the photo: Taylor, Wayne

Creation, EXIF & technical information:

Time	
File format	JPEG
File size	15797
Image x-width	300
Image y-width	408
Bits per pixel	24

Image Quality Rating:

Who rated the quality of the picture: Doe, John

□ Include quality rating

Meta Data Description:

Who created the meta data: Doe, John

Comment:

Finished

12,4M of 15,6M

Persons

Female PoleVaulter

Kym Howe

Events

CommonWealth PoleVaultCompetition

Places, Times and Objects

GoldMedal

Pole

Summer2006

Finished

12,7M of 15,6M

Fig. 6. Example image annotation using Caliph; generic (image information) and (semantic) descriptive annotation tabs.

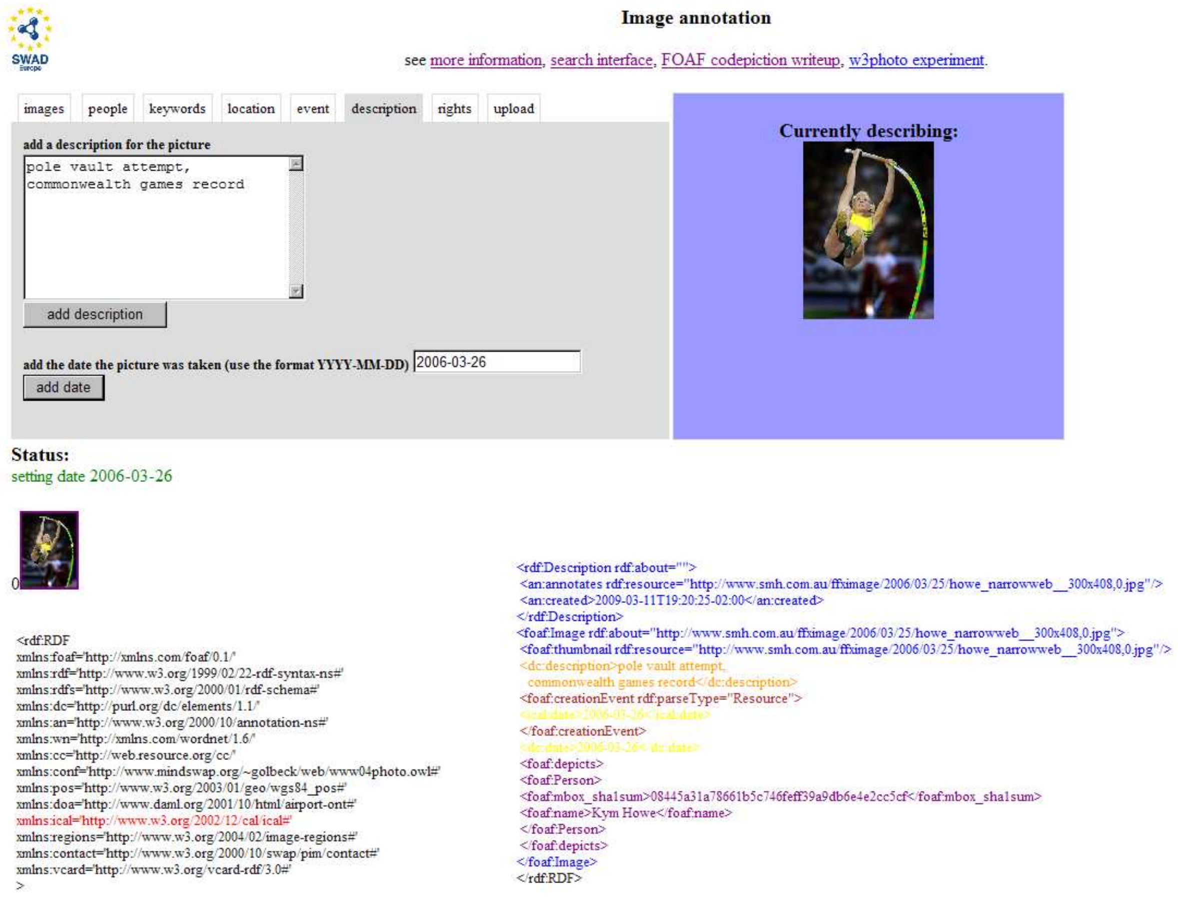


Image annotation

see more information, search interface, FOAF codepiction writeup, w3photo experiment.

add a description for the picture

```
pole vault attempt,
commonwealth games record
```

add the date the picture was taken (use the format YYYY-MM-DD) 2006-03-26

Status:
setting date 2006-03-26

0

```
<rdf:RDF
xmlns:foaf='http://xmlns.com/foaf/0.1/'
xmlns:rdf='http://www.w3.org/1999/02/22-rdf-syntax-ns#'
xmlns:rdfs='http://www.w3.org/2000/01/rdf-schema#'
xmlns:dc='http://purl.org/dc/elements/1.1/'
xmlns:an='http://www.w3.org/2000/10/annotation-ns#'
xmlns:wn='http://xmlns.com/wordnet/1.6/'
xmlns:cc='http://web.resource.org/cc/'
xmlns:conf='http://www.mindswap.org/~golbeck/web/www04photo.owl#'
xmlns:pos='http://www.w3.org/2003/01/geo/wgs84_pos#'
xmlns:doa='http://www.daml.org/2001/10/html/airport-ont#'
xmlns:ical='http://www.w3.org/2002/12/cal/ical#'
xmlns:regions='http://www.w3.org/2004/02/image-regions#'
xmlns:contact='http://www.w3.org/2000/10/swap/pim/contact#'
xmlns:vcard='http://www.w3.org/vcard-rdf/3.0#'
>
```

```
<rdf:Description rdf:about="">
 <an:annotates rdf:resource="http://www.smh.com.au/ffximage/2006/03/25/howe_narrowweb__300x408,0.jpg"/>
 <an:created>2009-03-11T19:20:25-02:00</an:created>
</rdf:Description>
<foaf:Image rdf:about="http://www.smh.com.au/ffximage/2006/03/25/howe_narrowweb__300x408,0.jpg">
 <foaf:thumbnail rdf:resource="http://www.smh.com.au/ffximage/2006/03/25/howe_narrowweb__300x408,0.jpg"/>
 <dc:description>pole vault attempt,
  commonwealth games record</dc:description>
 <foaf:creationEvent rdf:parseType="Resource">
 <ical:date>2006-03-26</ical:date>
 </foaf:creationEvent>
 <dc:date>2006-03-26</dc:date>
 <foaf:depicts>
 <foaf:Person>
 <foaf:mbox_sha1sum>08445a31a78661b5c746feff39a9db6e4e2cc5cf</foaf:mbox_sha1sum>
 <foaf:name>Kym Howe</foaf:name>
 </foaf:Person>
 </foaf:depicts>
</foaf:Image>
</rdf:RDF>
```

Fig. 7. Example image annotation using the SWAD annotation tool.

licensing information as described in the respective SWAD deliverable[30]. When entering a keyword description, the respective Wordnet[31] hierarchy is shown to the user, assisting her in determining the appropriateness of the keyword and in selecting descriptions of further accuracy. The number of RDF vocabularies the tool utilises is quite impressive, including FOAF, the Dublin Core element set, RDFiCalendar[32] as well as an experimental by the time namespace for WordNet, the latter in an attempt towards explicit subject matter semantics.

3.7 LabelMe

LabelMe[33] is a database and web-based image annotation tool, aiming to contribute in the creation of large annotated image databases for evaluation and training purposes [28]. It contains all images from the MIT CSAIL[34] database, in addition to a large number of user uploaded images. Figure 8 depicts a screenshot using LabelMe to annotate the pole vaulter and pole objects of the example image.

LabelMe [28] supports descriptive metadata addressing in principle region-based annotation. For each image, randomly selected from the database or user uploaded, the user my annotate as many objects as desired in order to further enrich already annotated images or provide new ones. There is no functionality for adopting a controlled vocabulary; instead each user may enter as many words as she considers appropriately in order to precisely describe the annotated object. For the localisation of regions, a manual drawing facility is provided. Specifically, the user defines a polygon enclosing the annotated object through a set of control points. Defining a polygon that equals the entire image allows for scene level annotations; we note though, that such behaviour rather diverges from the intended goal, i.e. the construction of a large, rich and open data set of annotated objects.

The resulting annotations are stored in XML format, with the choice of XML based on portability and extensibility concerns. A proprietary schema is followed, including attributes such as *filename*, *folder*, and *object* that allow to represent information regarding the image and its location, and the annotation itself. Additional elements under the *object* attribute, allow to represent the various words ascribed to the annotated object, the coordinates of the polygon, the date the annotation was created, and whether it has been verified by the use or not.

Summing up, LabelMe addresses image annotation from a rather different perspective than the rest of the tools. Its focus on requirements related to object recognition research, rather than image search and retrieval, entails different notions regarding the utilisation, sharing and purpose of annotation. In a way,

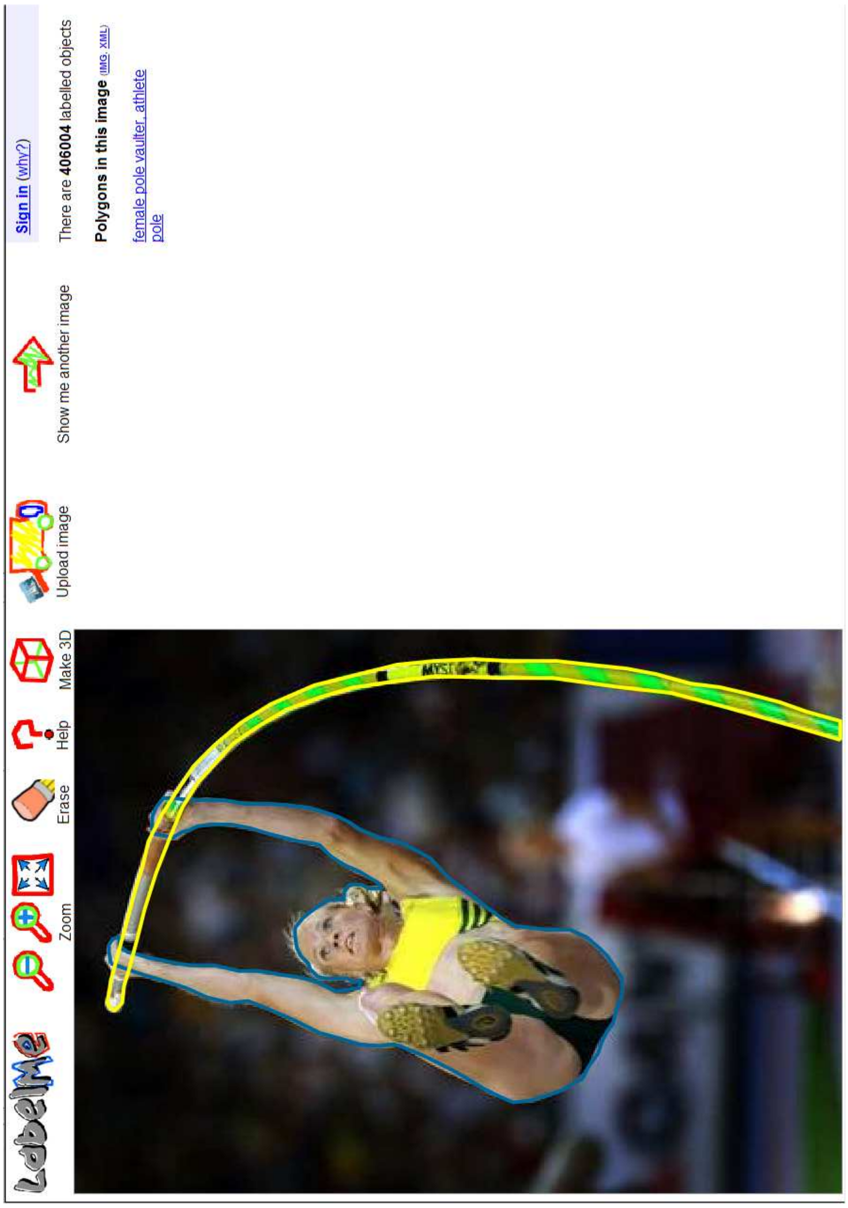
[30] http://www.w3.org/2001/sw/Europe/reports/report_semweb_access_tools/♯WN

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

[32] http://www.w3.org/2002/12/cal/

[33] http://labelme.csail.mit.edu/

[34] http://web.mit.edu/torralba/www/database.html



There are **406004** labelled objects

Polygons in this image (IMG, XML)

female pole vaulter, athlete

pole

Fig. 8. Example image annotation using LabelMe.

it is closer to M-Ontomat-Annotizer, but lacking formal domain specific as well as low-level descriptors representation; in addition the extraction of descriptors and their linking with domain concepts is left up to the algorithms using the annotations to implement object recognition.

3.8 Application-specific Image Annotation Tools

Apart from the afore described semantic image annotation tools, a variety of application-specific tools are available. Some of them relate to Web 2.0 applications addressing tagging and sharing of content among social groups, while others focus on particular application domains, such as medical imaging, that impose additional specifications pertaining to the individual application context. Aspiring to specific usages, these tools induce different perspectives and specifications on the annotation process. In the following, we briefly go through some representative examples.

iPad (image Physician Annotation Device) supports clinicians in the semantic annotation of radiological images [29]. Using the provided drawing facilities the user selects the regions of interest and attaches to them descriptions referring to anatomical, pathological and imaging observations. Utilising radiology specific ontologies, iPad enhances the annotation procedure by suggesting more specific terms and by identifying incomplete descriptions and subsequently prompting for missing parts in the description (e.g. “enlarged” is flagged as incomplete while “enlarged liver” is acceptable). The created annotations are stored in XML based on a proprietary schema, which can be subsequently transformed into different standard formats such as DICOM[35] and HL7[36] in order to support seamless and effective interchange of medical data across heterogenous systems. Furthermore, aspiring to enhance interoperability with Semantic Web technologies, translation to OWL is also provided.

FotoTagger[37] builds on the paradigm of the popular Web 2.0 application of Flickr[38]. It comes both as a Web-based and a standalone application, allowing users to attach tags to specific image regions with the purpose of enhancing content management in terms of accessing and sharing it. It supports descriptive and structural metadata, where region localisation is performed through a rectangle drawing facility. The produced descriptions are in RDF/XML following a proprietary schema[39] that models the label constituting the tag, its position (the label constitutes a rectangle region in itself), and the position of the rectangle that encloses the annotated region in the form of the top left point coordinates and width and height information. Furthermore, general information about the image is included such as image size, number of regions annotated, etc. Oriented towards Web 2.0, FotoTagger places significant focus on social aspects pertaining to content management, allowing among others to publish tagged images to

[35] http://www.rsna.org/Technology/DICOM/

[36] http://www.hl7.org/

[37] http://www.fototagger.com/

[38] http://www.flickr.com/

[39] http://www.cogitum.com/fototagger/

blogs and to upload/download tagged images to/from Flickr, while maintaining both FotoTagger's and Flickr's descriptions.

Given the general purpose scope of the current survey, elaborating into the various application specific tools and the particular annotation aspects they introduce falls beyond the intended scope. It is worth noting though that as the corresponding literature shows, interoperability, even when not necessarily in conformance with the SW notion, constitutes a major concern.

3.9 Discussion

The aforementioned overview reveals that the utilisation of Semantic Web languages for the representation, interchange and processing of image metadata has permeated semantic image annotation. This is particularly evident for subject matter descriptions, where from the examined tools only Caliph and LabelMe follow a different approach. Caliph though is more oriented towards content-based annotation and retrieval in the "traditional" multimedia community sense, and thus adopts the MPEG-7 perspective. The choice of a standard representation shows the importance placed on creating content descriptions that can be easily exchanged and reused across heterogenous applications, and works like [10, 11, 30] provide bridges between MPEG-7 metadata and the Semantic Web and existing ontologies. The case is different with LabelMe, where the tool serves a very specific purpose that of creating a large object annotated database, and does not address retrieval tasks. Even in this case though, one can speculate that adopting a more formal vocabulary the descriptions added by users could be better exploited.

The representation of structural and localisation information appears to be also wide established, illustrating that there is a considerable need to attach descriptions to specific content parts. It is interesting that in all tools supporting such kind of description, an ontology has been used (Caliph is the exception following the MPEG-7 decomposition schemes), which is hidden from the user. Thus unlike subject matter descriptions, where a user can choose which vocabulary to use (in the form of a domain ontology, a lexicon or user provided keywords), structural descriptions are tool specific. The different ontologies used by the tools reflect the undergoing efforts towards making structural semantics explicit and the variations witnessed due to the loose semantics of the corresponding MPEG-7 definitions on which these ontologies are based on [31, 12]. Media related information on the other hand in terms of low-level descriptors can be represented in a rather straightforward manner, practically eliminating interoperability issues. The choice of whether or not to include support for media related annotations depends on whether the tool aims to contribute to analysis tasks as well.

Summing up, the choice of a tool depends primarily on the intended context of usage, which provides the specifications regarding the annotation dimensions supported, and subsequently on the desired formality of annotations, again related to a large extend to the application context. Thus for semantic retrieval purposes, where semantic refers to the SW perspective, KAT, PhotoStuff, SWAD

Tool	**Input & Output**		**Annotation level**			
	Metadata Format	*Annotation Vocabulary*	*Metadata Type*	*Granularity*	*Localisation*	*Expressivity*
KAT	OWL	U: domain ontology (RDFS/OWL) T: COMM	descriptive, structural	image, region-based	rectangle, polygon	concepts
PhotoStuff	OWL	U: domain ontology (OWL) T: Digital Media, Technical ontologies	descriptive, structural, administrative	image, region-based	rectangle, circle, polygon	concepts, relations
AktiveMedia	RDF	U: domain ontology (RDFS/DAML/OWL), free text T: customised structural schema	descriptive	image, region-based	rectangle, circle	concepts
M-Ontomat Annotizer	RDF	U: domain ontology (RDFS/DAML) T: VAO, VDO	descriptive, media	image, region-based	rectangle, eclipse, polygon, free hand	concepts
Caliph	MPEG-7/XML custom XML	U: free text, keywords T: MPEG-7	descriptive, structural, media, administrative	image	N/A	concepts relations
SWAD	RDF	U: free text, keywords T: Dublin Core, FOAF, WordNet	descriptive, administrative	image	N/A	concepts relations
LabelMe	custom XML	U: free text, keywords	descriptive	image region-based	polygon	concepts

Table 1. Image annotation tools summarisation. In the *A*nnotation Vocabulary field, “U” denotes user-entered vocabularies, while “T” refers to vocabularies embedded within the tool, and thus hidden to the user.

and AkiveMedia would be the more appropriate choices. In cases that domain semantics need to be associated with low-level representations a tool like M-Ontomat-Annotizer or KAT should be selected. Finally, when adopting a strict MPEG-7 perspective is required, then a tool like Caliph should be preferred. We note the difference between MPEG-7 metadata, i.e. XML descriptions according to the respective Description Schemes, and MPEG-7 compliant metadata that can be as well in RDFS or OWL. Table 1, summarises the comparative study of the examined image annotation tools with respect to the *Input & Output* and *Annotation Level* criteria described in Section 2. Regarding the miscellaneous criteria (see Section 2.3), as illustrated in the individual tools descriptions, none provides supports for collaborative annotation. Web-based and stand-alone are equally popular choices, and all tools are freely available for non-commercial use[40].

4 Tools for Semantic Video Annotation

The increase in the amount of video data deployed and used in today's applications not only caused video to draw increased attention as a content type, but also introduced new challenges in terms of effective content management. Image annotation approaches as described in the previous section can be employed for the description of static scenes found in a video stream; however, in order to capture and describe the information issuing from the temporal dimension featuring a video object, additional requirements emerge.

In the following we survey typical video annotation tools, highlighting their features with respect to the criteria delineated in Section 2. In addition to tools that constitute active research activities, we also examine representative video annotation systems that despite no longer maintained, are still accessible and functional; however, tools that are neither maintained nor accessible have not been considered. In the latter category fall tools such as VIDETO[41], Ricoh Movie Tool[42], or LogCreator[43]. It is interesting to note that the majority of these tools followed MPEG-7 for the representation of annotations. As described in the sequel, this favourable disposition is still evident, differentiating video annotation tools from image ones, where the Semantic Web technologies have been more pervasive.

4.1 VIA

The Video and Image Annotation[44] (VIA) tool has been developed by the MK-Lab[45] within the BOEMIE[46] project. A snapshot of the interface of the tool,

[40] In many cases, the source code is available for research purposes

[41] http://www.zgdv.de/zgdv/zgdv/departments/zr4/Produkte/videto/

[42] http://www.ricoh.co.jp/src/multimedia/MovieTool/

[43] http://project.eia-fr.ch/coala/demos/demosFrameset.html

[44] http://mklab.iti.gr/project/via

[45] http://mklab.iti.gr

[46] http://www.boemie.org

during a shot annotation of a video file is shown in Figure 9. The shot records a pole vaulter holding a pole and sprinting at the jump point.

VIA supports descriptive, structural and media metadata of image and video assets. Descriptive annotation is performed with respect to a user loaded OWL ontology, while free text descriptions can also be added. Administrative metadata follow a customised schema internal to the tool, including information about the creator of the annotations, the date of the annotation creation, etc. A customised XML schema is also used for the representation of structural information, allowing for example to nest a video segment as part of a video and to define its start and end frame / time interval. The produced metadata can be exported either in XML or as in a more human readable format in textual format.

Regarding image (and by consequence frame) annotation, the granularity levels supported include the entire image and specific still regions. The localisation of regions is performed either semi-automatically, providing the user a segmented image and allowing her to correct it by region merging, or manually, using one of the drawing functionalities provided, i.e. free hand, polygon, circle, rectangle. In the case of image annotation, the tool supports additionally the extraction of MPEG-7 visual descriptors per each annotated region, based on MPEG-7 XM [32], so the annotation outcome can be used as a training set for semantics extraction algorithms.

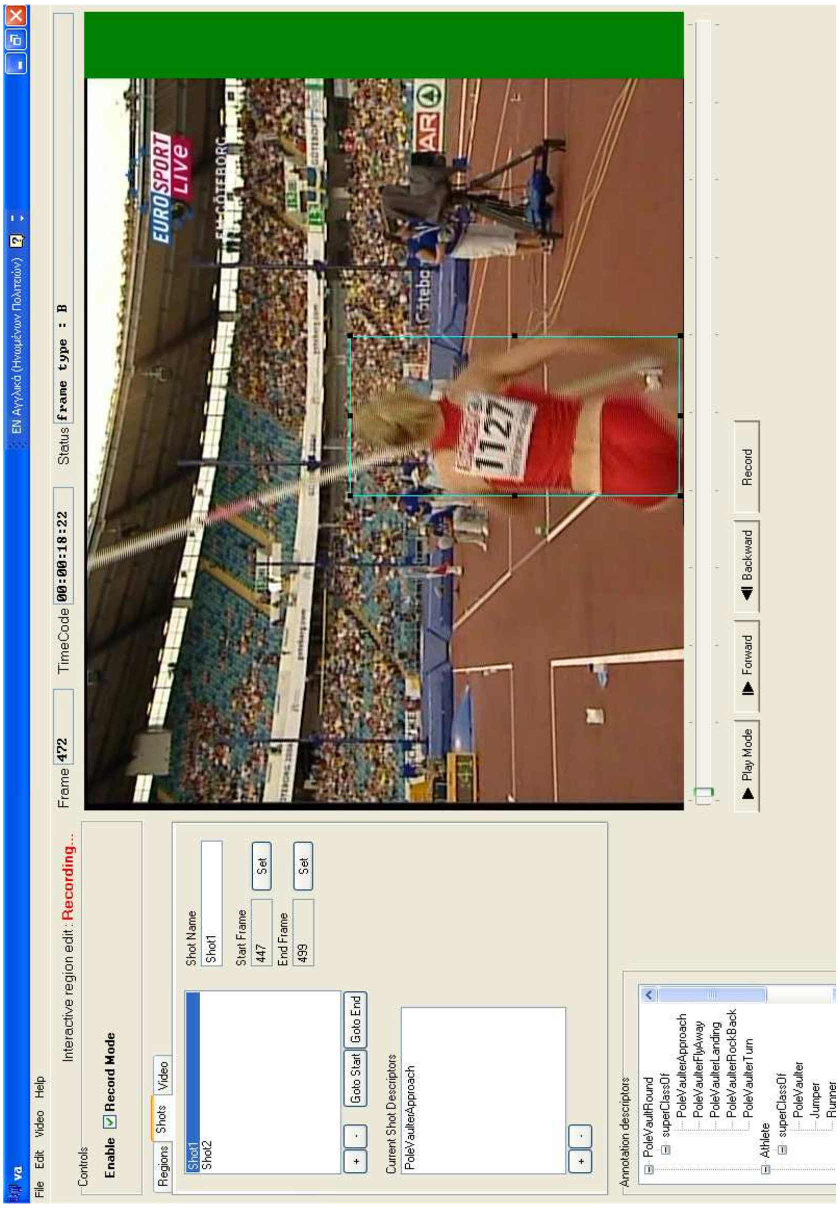
Regarding video annotation, the supported annotation granularity may refer respectively either to the entire video, video segments, moving regions, frames or even still regions within a frame. The annotation can be performed in real time, on MPEG-1 and MPEG-2 videos, using an interface consisting of three panels. The first one is concerned with region annotation, in which the user selects rectangular areas of the video content and subsequently adds corresponding annotations. The other two panels are used for annotation at shot and video level respectively. Shot boundaries are defined manually, by selecting its start and end frames. An important feature about region annotation is that the user can drag the selected region whereas at the same time the video is playing, so as to follow the movement of the desired region.

The annotations performed with VIA can be saved as annotation projects, so that the original video, the imported ontologies, and the annotations can be retrieved and updated at a later time. VIA is publicly available.

4.2 VideoAnnEx

The IBM VideoAnnEx[47] annotation tool addresses video annotation with MPEG-7 metadata. Although the project within which VideoAnnEx was developed has finished and the tool is no longer maintained, VideoAnnEx is accessible and provides an illustrative case of content annotation in accordance to the MPEG-7 initiative. A screenshot of the annotation interface of the tool is shown in Figure 10.

[47] http://www.research.ibm.com/VideoAnnEx/index.html



Interactive region edit: Recording...

Controls

Enable ☑ Record Mode

Shot1
Shot2

Shot Name: Shot1

Start Frame: 447

End Frame: 499

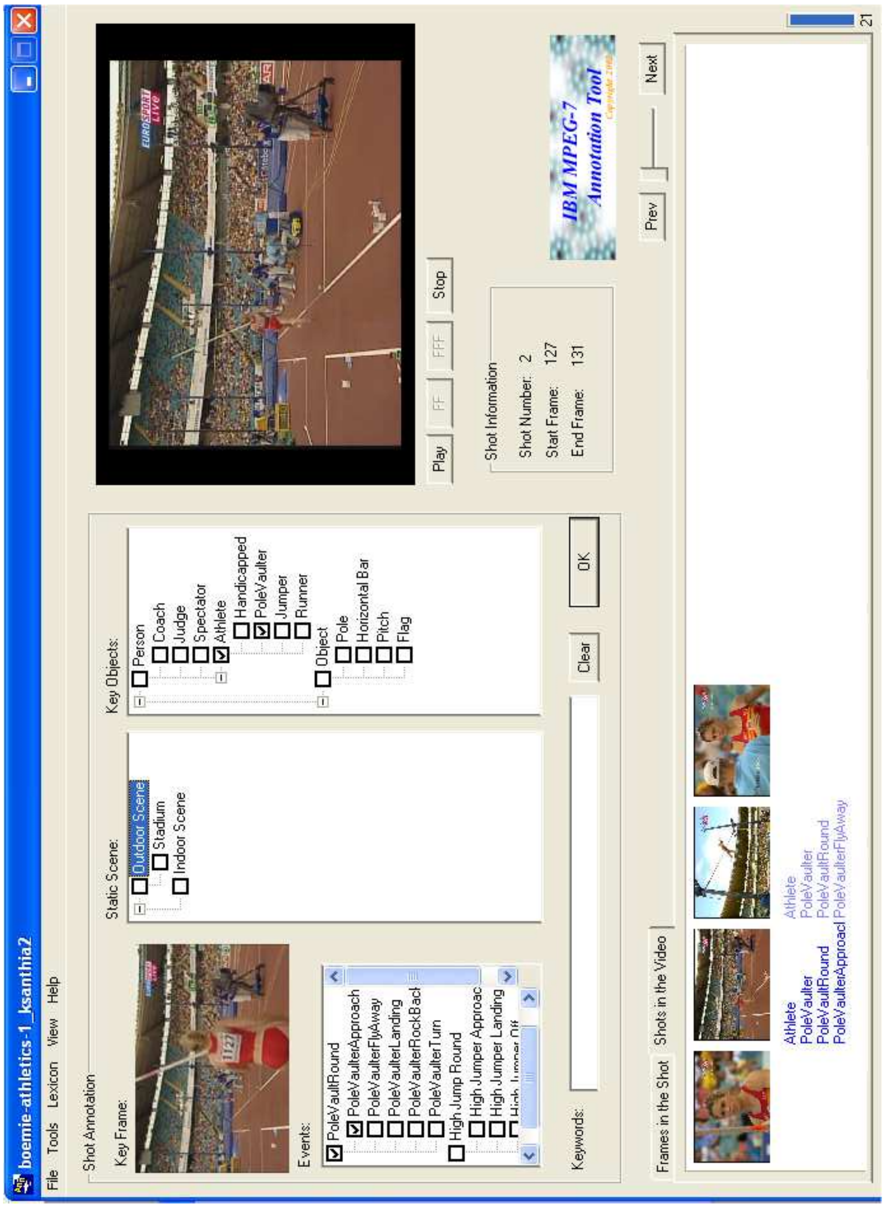
Current Shot Descriptors: PoleVaulterApproach

Frame 472 TimeCode 00:00:18:22 Status frame type : B

Annotation descriptors

```
⊟ PoleVaultRound
  ⊟ superClassOf
     PoleVaulterApproach
     PoleVaulterFlyAway
     PoleVaulterLanding
     PoleVaulterRockBack
     PoleVaulterTurn
⊟ Athlete
  ⊟ superClassOf
     PoleVaulter
     Jumper
     Runner
```

Fig. 9. Example video annotation using VIA.



Static Scene:
- □ Outdoor Scene
  - □ Stadium
- □ Indoor Scene

Key Objects:
- □ Person
  - □ Coach
  - □ Judge
  - □ Spectator
  - ☑ Athlete
    - □ Handicapped
    - ☑ PoleVaulter
    - □ Jumper
    - □ Runner
- □ Object
  - □ Pole
  - □ Horizontal Bar
  - □ Pitch
  - □ Flag

Events:
- ☑ PoleVaultRound
  - ☑ PoleVaulterApproach
  - □ PoleVaulterFlyAway
  - □ PoleVaulterLanding
  - □ PoleVaulterRockBack
  - □ PoleVaulterTurn
- □ High Jump Round
  - □ High Jumper Approac
  - □ High Jumper Landing
  - □ High Jumper Dif

Shot Information
Shot Number: 2
Start Frame: 127
End Frame: 131

Athlete
PoleVaulter
PoleVaultRound
PoleVaulterApproacl

Athlete
PoleVaulter
PoleVaultRound
PoleVaulterFlyAway

Fig. 10. Example video annotation using VideoAnnEx.

VideoAnnex supports descriptive, structural and administrative annotations according to the respective MPEG-7 Description Schemes. Descriptive metadata may refer at the entire video, at specific video segments (shots), or even at still regions within keyframes. The tool supports default subject matter lexicons in XML format, and additionally allows the user to create and load her own XML lexicon, design a concept hierarchy through the interface menu commands, or insert free text descriptions.

As illustrated in Figure 10, the VideoAnnEx annotation interface consists of four components. On the upper right-hand corner of the tool is the Video Playback window with shot information. It allows standard VCR operations (such as play, pause, etc.) and loads video files in MPEG-1 or MPEG-2 format. On the upper left-hand corner of the interface is the Shot Annotation panel with a key frame image display. The tool supports either automatic shot detection or loading of customised video segmentation lists. In the space between the two display windows, the concept hierarchy of the loaded XML lexicon is displayed.

On the bottom part of the tool, two views are available of the annotation preview: one contains the I-frames of a shot and the keyframes of each shot in the video, respectively. The user may see under the keyframe of each shot, the annotation this shot has received, up to this point. A fourth component, not shown in Figure 10, is the region annotation pop-up window for specifying annotated regions using a rectangle. After the text annotations are identified on the shot annotation window, each description can be associated with a corresponding rectangular region on the selected key frame of that shot.

It worths noticing an extra feature this tool offers, which is annotation learning. This utility assist the annotator in finding similar shots and labeling them with the same descriptions. VideoAnnEx runs on Windows platforms and can be used under the IBM terms of use[48].

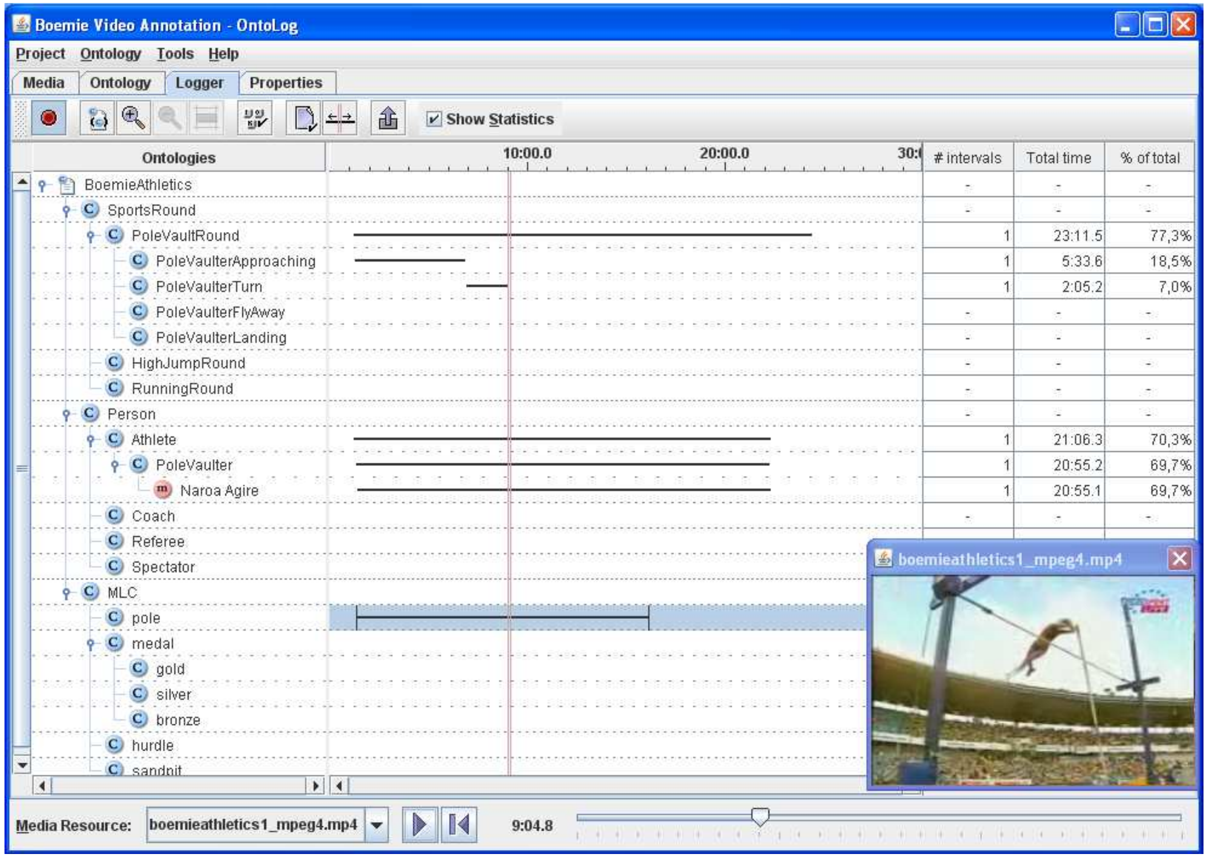
4.3 Ontolog

Ontolog[49] is a tool for annotating video and audio sources using structured sets of terms/concepts. It is a java application, designed and developed as part of a Ph.D. thesis in the Norwegian University of Science and Technology. Though not maintained the past four years, the source code is available upon request. A screenshot of a video annotation process is shown in Figure 11.

Ontolog addresses various types of metadata, including descriptive, structural and administrative. Descriptive annotations are inserted according to one or more RDFS ontologies, imported or created by the user. The user can further enrich the subject matter descriptions by introducing additional properties. For the representation of administrative metadata, Ontolog provides by default two ontologies, namely the Dublin Core Element Set and the Dublin Core Qualified Element Set. Structural descriptions referring to video segments are created in correspondence with user-defined intervals, following the simplified structure

[48] http://www.ibm.com/legal/

[49] http://www.idi.ntnu.no/ heggland/ontolog/



Boemie Video Annotation - OntoLog
Project Ontology Tools Help
Media Ontology Logger Properties
☑ Show Statistics
10:00.0
20:00.0
30:0
boemieathletics1_mpeg4.mp4
Media Resource:
9:04.8

Ontologies	# intervals	Total time	% of total
BoemieAthletics	-	-	-
SportsRound	-	-	-
PoleVaultRound	1	23:11.5	77,3%
PoleVaulterApproaching	1	5:33.6	18,5%
PoleVaulterTurn	1	2:05.2	7,0%
PoleVaulterFlyAway	-	-	-
PoleVaulterLanding	-	-	-
HighJumpRound	-	-	-
RunningRound	-	-	-
Person	-	-	-
Athlete	1	21:06.3	70,3%
PoleVaulter	1	20:55.2	69,7%
Naroa Agire	1	20:55.1	69,7%
Coach	-	-	-
Referee			
Spectator			
MLC			
pole			
medal			
gold			
silver			
bronze			
hurdle			
sandpit			

Media Resource: boemieathletics1_mpeg4.mp4

Fig. 11. Example video annotation using Ontolog.

representation defined in the Ontolog Schema[50] ontology. The produced annotations are in RDF.

Ontolog's interface consists of four components: a Media Panel, an Ontology Editor, a Logging Panel and a Property Editor. The media panel handles the video assets that are contained in an annotation project. For media loading either Quicktime (for Java) or the JMF framework can be used (and the corresponding media formats). The Ontology Editor provides mechanisms for the definition of concept hierarchies; properties defining relations between concepts can be specified in the Property Editor. Each property may optionally specify what kind of concept it may be applied to (domain) and what kind of values it may take (range).

OntoLog's logging interface is shown in Figure 11. The left panel contains the ontologies the user is working with. The right panel displays a horizontal timeline with the annotation intervals corresponding to each concept in the ontology (referred to as "annotation strata", in the context of this tool). Each stratum consists of a series of interval lines along the time axis, indicating the positions of the media resource where the concept is present. The strata corresponding to collapsed concepts (concepts with subconcepts that are not currently displayed in the tree) are shown as of lines of varying thickness. This is because they represent an aggregation of the strata beneath them in the hierarchy. The time intervals are specified manually, i.e. automatic or semiautomatic temporal segmentation is not supported.

An extra feature the tool offers involves the extraction of simple statistics, such as the length of the intervals per concept/instance, the percentage of this length with regard to the total length of the media resource, etc. In addition, the resulting set of annotation intervals (i.e. strata) serves as a visual index to the media file, with dynamic level of detail due to the tree-based, aggregating visualisation technique. Moreover, the logging panel provides a SMIL export function. This produces a SMIL file [33], specifying a "virtual edit" of the selected media resource, namely a concatenation of the intervals related to the currently selected concept. For instance, a user may create a SMIL version of the "Olympics 2008" video with just the parts with running events. Concluding, Ontolog is accompanied with the Ontolog Crawler software[51], which implements many search queries and facilitates the task of retrieval.

4.4 Advene

Advene[52] (Annotate Digital Video, Exchange on the NEt) is an ongoing project in the LIRIS[53] laboratory at University Claude Bernard Lyon. Advene addresses a twofold goal, namely to provide an annotation model for sharing descriptions about digital video documents, and to serve as an authoring tool for visualising

[50] http://www.idi.ntnu.no/ heggland/ontolog/ontolog-schema#

[51] http://folk.ntnu.no/heggland/ontolog-crawler/login.php

[52] http://liris.cnrs.fr/advene/

[53] http://liris.cnrs.fr/

and accessing hypervideos, i.e. videos augmented with annotations. A screenshot of the interface of the tool during a video annotation is shown in Figure 12.

Annotation in Advene is performed according to user-created schemas which group together descriptions of related annotation dimensions (i.e. subject matter, administrative, etc.). Schemas including concept level descriptions are referred as annotation types, while schemas defining relations between concepts, comprise the relation types. Each annotation type defines in addition a content type for its annotations, in the form of a MIME type (text/plain,text/XML, image/jpeg, audio/wav, etc.). If the type is text/XML, it can be further constrained by a structured description (e.g. using DTD). Analogously, a relation type defines a content type for its instances. In addition, it specifies the number of participating annotations and their respective types. The generated annotations may contain descriptive, administrative and structural information and may pertain to the entire video or to temporal segments of it. The output is stored in XML format.

Advene uses the VLC video player[54] that supports various audio and video formats, such as MPEG-1, MPEG-2, MPEG-4, DivX, mp3, ogg, and so on, as well as DVDs, VCDs, and various streaming protocols. The tool offers the ability to dynamically control the video player based on the annotations, as well as to define dynamic visualisation means (views). Moreover, it allows multiple ad-hoc views of annotations (e.g. timeline, tree-view, transcription, etc) and the annotations' content may be displayed as SVG caption on the video. The annotations along with the views may be shared in packages independently from the audiovisual material, through an embedded web server which dynamically generates XHTML[55] documents, using data taken from the annotations.

The main focus of Advene is not so much to support the annotation task itself, but rather to offer visualisation means and the functionalities afore described, so as to facilitate the management of readily available annotation metadata. This accounts for the variety of annotation formats that the tool supports, among which TXT files where each line contains the start time, the end time and the contents of the annotation separated by tabs, SRT[56] subtitle files, XI[57] XML files, EAF[58] files produced with ELAN, PRAAT[59] files, CMML[60] files, Anvil files, MPEG-7 files containing only free text annotations, AnnotationGraph[61], Shotdetect and IRI files[62]. Advene is distributed under the GPL conditions and runs on Linux, Windows and MacOS platforms.

[54] http://www.videolan.org/vlc/

[55] http://www.w3.org/TR/xhtml1/

[56] http://www.matroska.org/technical/specs/subtitles/srt.html

[57] http://www.ananas.org/xi/index.html

[58] http://www.let.kun.nl/sign-lang/echo/ELAN/ELAN_intro.html

[59] http://www.fon.hum.uva.nl/praat/

[60] http://www.anodex.net/

[61] http://sourceforge.net/projects/agtk

[62] http://www.iri.centrepompidou.fr/

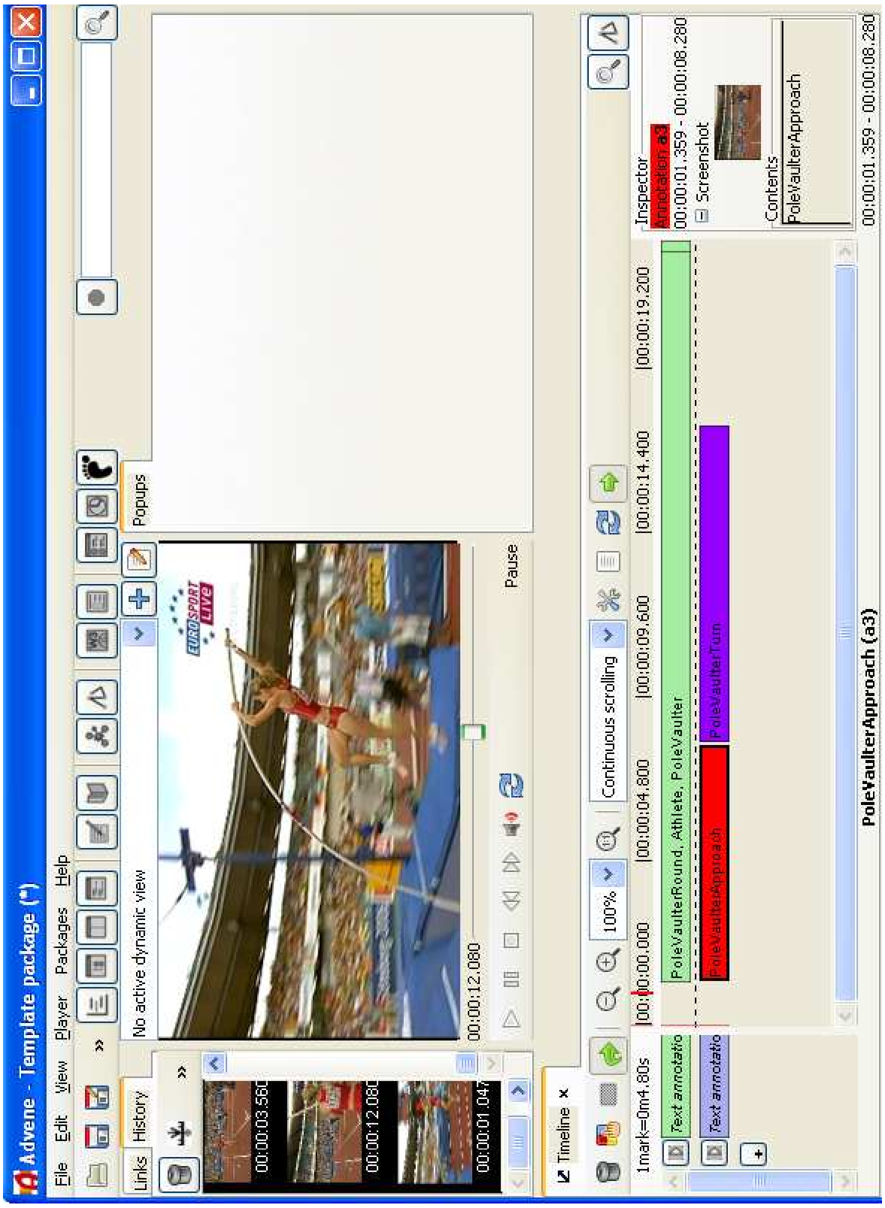


Fig. 12. Example video annotation using Advene.

4.5 Elan

Elan[63], developed at the Max Planck Institute for Psycholinguistics[64], is an annotation tool designated primarily for linguistic purposes, involving issues related to analysis of language, sign language and gestures in audio and video resources. A screenshot showing a video annotation along with the user interface of Elan is shown in Figure 13.

The tool addresses exclusively descriptive annotations, where an annotation may be a sentence, word or gloss, and in general any description of a feature observed in the media file. The user may also create and use her own vocabularies, containing frequently used terms, so that she avoids repetitive typing of the same term. The produced metadata is in XML format and refer either to the entire video or to temporal segments of it.

Annotations, in Elan, can be created on multiple layers, called tiers which can be hierarchically interconnected, so that annotations in a referring tier are linked to annotations on a referred tier. This feature pertains to the linguistic design and multi-language support of the tool, so that different tiers correspond to different translations. However, it can also be used so as to simulate a structural description of the content (parent tiers describe video objects and children tiers describe segments of the former) or, in general, produce annotations containing meta information about other annotations.

In the upper left part of the interface of Elan is the media player. The kind and number of supported video formats depend upon the media framework the user has installed. There are three supported media players, that is Windows Media Player, QuickTime and JMF. Below the player window, there are the media control buttons. Apart from the standard VCR operations, the tool supports browsing based on frames and on user-assigned annotations. The lower part of the interface includes the timeline viewer. There are multiple timelines, one for each particular tier. The timeline viewer displays the tiers and their annotations, whereby each annotation corresponds to a specific time interval. With regard to the localisation of the video content, the user has to manually select the intervals, she wants to annotate.

Further, the tool offers keyword-based and regular expression based search functionalities that facilitate the task of retrieval, as well as it supports a variety of import/export functions with formats, such as Shoebox/Toolbox[65], CHAT[66], Transcriber[67], Praat[68], SMIL[33], etc. Elan is distributed under the GPL conditions and runs on Windows, MacOS and Linux platforms.

[63] http://www.lat-mpi.eu/tools/elan/

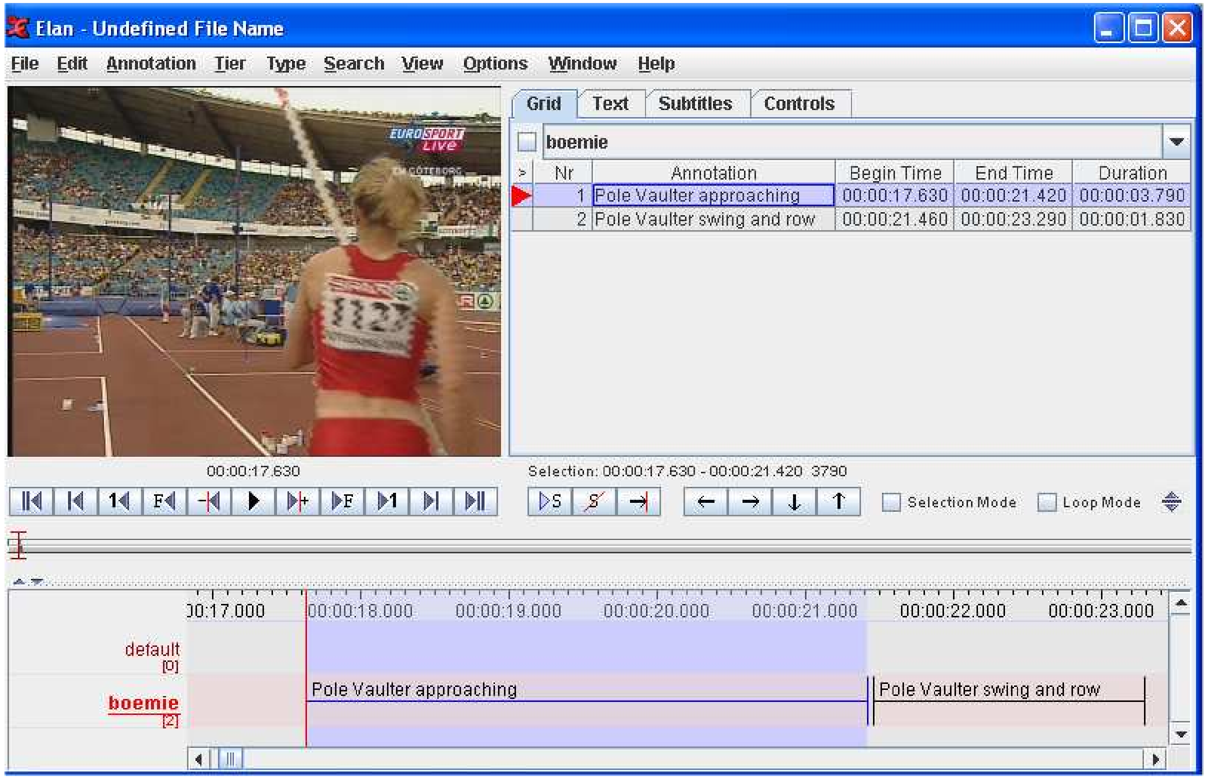
[64] http://www.mpi.nl

[65] http://www.sil.org/computing/catalog/show_software.asp

[66] http://childes.psy.cmu.edu/

[67] http://trans.sourceforge.net/en/history.php

[68] http://www.fon.hum.uva.nl/praat/



□ boemie

Nr	Annotation	Begin Time	End Time	Duration
1	Pole Vaulter approaching	00:00:17.630	00:00:21.420	00:00:03.790
2	Pole Vaulter swing and row	00:00:21.460	00:00:23.290	00:00:01.830

Selection: 00:00:17.630 - 00:00:21.420 3790

Pole Vaulter approaching

Pole Vaulter swing and row

Fig. 13. Example video annotation using Elan.

4.6 Anvil

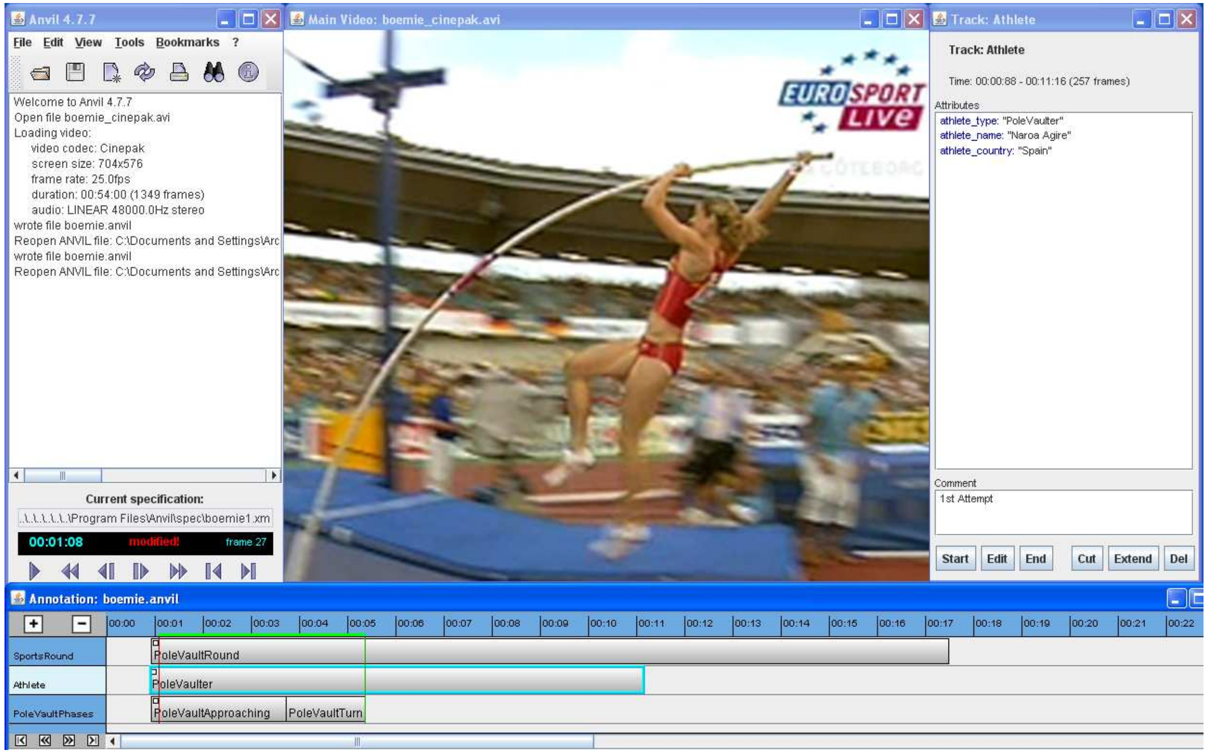
Anvil[69] is a tool that supports audiovisual content annotation, but which was primarily designed for linguistic purposes, in the same vein as the previously described tool. It was developed as part of a Ph.D. thesis at the Graduate College for Cognitive Science[70] and the German Research Center for Artificial Intelligence (DFKI[71]). A screenshot showing a video annotation along with the user interface of Anvil v4.7.7 is shown in Figure 14.

Anvil [34] supports descriptive, structural and administrative annotations of video or audio objects that refer to the entire assets or to temporal segments of them. User-defined XML schema specification files provide the definition of the vocabulary used in the annotation procedure. The output is an XML file containing administrative information in its head segment, while its body includes the descriptive metadata along with structural information regarding the temporal localisation of the possible video segments. Recently, Anvil has been extended to support spatiotemporal annotation as well by allowing annotations to be attached to specific points [35]; interpolation functionalities and arbitrary shapes constitute future extensions.

The tool uses hierarchical user-defined layers, in exactly the same way as described in the previous tool. Its interface consists of the media player window, the annotation board and the metadata window. The player loads files in AVI and MOV format and supports standard video controls, including frame-by-frame stepping. The annotation board contains except for the standard timeline, a waveform timeline, a pitch/intensity timeline and timelines for each described concept. The latter timelines follow the hierarchy of the concept definition in the XML file and may be collapsed or not for better viewing. As in most described tools, also in Anvil, the user has to manually define the temporal segments that wants to annotate.

Anvil can import data from the phonetic tools PRAAT[72] and XWaves which perform speech transcriptions. Moreover, it can export data to SPSS[73] and Statistica[74] for statistical analysis of the annotated data. As in more tools described in this Section, Anvil offers functionalities that allow search in the annotations, facilitating, thus, the retrieval task. It also allows the creation of bookmarks that correspond to the favorite annotations of each user. Anvil is written in Java, runs on Windows, Macintosh and Unix (Solaris/Linux) platforms and it is publicly available upon request.

[69] http://www.anvil-software.de/
[70] http://www.ps.uni-sb.de/gk/kog/cognition.html
[71] http://www.dfki.de/web
[72] http://www.fon.hum.uva.nl/praat/
[73] http://www.spss.com/statistics/
[74] http://www.statsoft.com/products/products.htm



```
Welcome to Anvil 4.7.7
Open file boemie_cinepak.avi
Loading video:
    video codec: Cinepak
    screen size: 704x576
    frame rate: 25.0fps
    duration: 00:54:00 (1349 frames)
    audio: LINEAR 48000.0Hz stereo
wrote file boemie.anvil
Reopen ANVIL file: C:\Documents and Settings\Arc
wrote file boemie.anvil
Reopen ANVIL file: C:\Documents and Settings\Arc
```

Current specification:
..\..\..\..\..\Program Files\Anvil\spec\boemie1.xm

00:01:08 modified! frame 27

Track: Athlete

Time: 00:00:88 - 00:11:16 (257 frames)

Attributes
athlete_type: "PoleVaulter"
athlete_name: "Naroa Agire"
athlete_country: "Spain"

Comment
1st Attempt

Fig. 14. Example video annotation using Anvil.

4.7 Semantic Video Annotation Suite

The Semantic Video Annotation Suite[75] (SVAS), developed by Joanneum research Institute of Information Systems & Information Management[76], targets the creation of MPEG-7 video annotations. Figure 15 illustrates a screenshot of the 1.5 release.

SVAS [36] encompasses two tools: the Media Analyzer, which extracts automatically structural information regarding shots and key-frames, and the Semantic Video Annotation Tool (SVAT), which allows to edit the structural metadata obtained through the Media Analyzer and to add administrative and descriptive metadata, in accordance with MPEG-7. The administrative metadata include information about the creator, the production date, the video title, shooting and camera details, and so forth.

The descriptive annotations correspond to the MPEG-7 semantic description tools deriving from the SemanticBase DS allowing to capture subject matter descriptions regarding persons, places, events, objects, and so forth, and may refer either to shot (video segment) or region level. Regarding the latter, the localisation of specific regions in a key frame (or any other frame) can be performed either manually using the provided bounding box and polygon drawing facilities, or by deploying automatic image segmentation. Once the location of an object of interest is determined, SVAT provides an automatic matching service in order to detect similar objects throughout the entire video. The detection results are displayed in a separate key-frame view, where for each of the computed key frames the detected object is highlighted. The user can partially enhance the results of this matching service by removing irrelevant key-frames; however more elaborate enhancement such as editing of the detected region's boundaries or of its location is not supported. The annotations entered for a specific region can be copied by one mouse click to all matching objects within the video, thus reducing massively the manual annotation time required. All views, including the shot view tree structure, can be exported to a CSV file and the metadata is saved in an MPEG-7 XML file. SVAS is publicly available.

4.8 Application-specific Video Annotation Tools

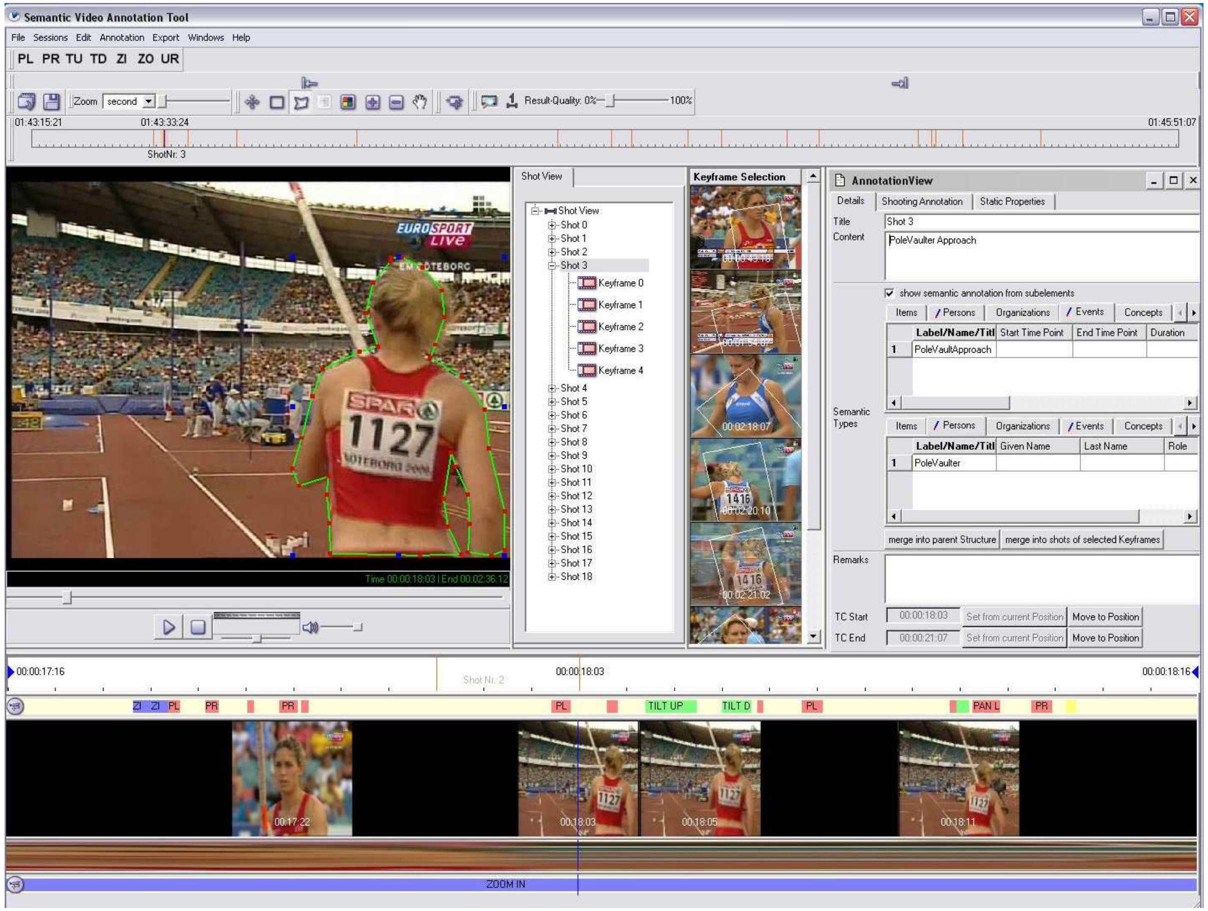
Apart from the afore described semantic video annotation tools, a number of additional annotation systems have been proposed that aspiring to specific application contexts induce different perspectives on the annotation process. To keep the survey comprehensive, in the following we examine briefly some representative examples.

Vannotea[77] is a tool for collaborative indexing, browsing, annotation and discussion of video content [37], developed by the University of Queensland. Contrary to the afore described annotation tools, Vannotea's primary focus consists

[75] http://www.joanneum.at/en/fb2/iis/products-solutions-services/semantic-video-annotation.html

[76] http://www.joanneum.at/en/jr.html

[77] http://www.itee.uq.edu.au/ eresearch/projects/vannotea/index.html



- Shot View
  - Shot 0
  - Shot 1
  - Shot 2
  - Shot 3
    - Keyframe 0
    - Keyframe 1
    - Keyframe 2
    - Keyframe 3
    - Keyframe 4
  - Shot 4
  - Shot 5
  - Shot 6
  - Shot 7
  - Shot 8
  - Shot 9
  - Shot 10
  - Shot 11
  - Shot 12
  - Shot 13
  - Shot 14
  - Shot 15
  - Shot 16
  - Shot 17
  - Shot 18

Title: Shot 3

Content: PoleVaulter Approach

☑ show semantic annotation from subelements

	Label/Name/Titl	Start Time Point	End Time Point	Duration
1	PoleVaultApproach			

	Label/Name/Titl	Given Name	Last Name	Role
1	PoleVaulter			

TC Start 00:00:18:03

TC End 00:00:21:07

Fig. 15. Example video annotation using SVAT.

in providing support for collaborative, real-time, synchronous video conferencing services. Interoperability concerns, in conjunction with the requirements for simple and flexible annotations, led to the adoption of an XML-based description schemes. Building on a simplified translation of the respective MPEG-7 and Dublin Core descriptions, Vannotea metadata can be easily transformed into the corresponding standardised representations through the use of XSLT. It is worth noticing that Vannotea builds on the Annotea initiative, a W3C activity aiming to advance the sharing of metadata on the Web. Advocating W3C standards, Annotea adopts RDF based annotation schemes and XPointer[78] for locating the annotations within the annotated resource.

ProjectPad[79] is a web-based system for collaborative media annotation and management tailored to distributed teaching and learning applications. Similarly to Vannotea, ProjectPad focused on providing synchronous interaction in terms of creation and editing of digital media collections and learning object metadata, for the purpose of supporting thematic content organisation, search and retrieval services. Annotations can be attached to the entire video (audio) asset or to specific temporal segments (spatial segments correspondingly in the case of images). Content is identified via Uniform Resource Identifiers[80] (URIs), while for the representation and storage of metadata both XML and RDF are supported.

The Video Performance Evaluation Resource Kits Ground Truth[81] (ViPER-GT) tool has been developed by the Language And Media Processing (LAMP) lab, at the University of Maryland, with the aim to assist in the evaluation of approaches addressing automatic semantic video analysis. ViPER-GT enables the creation and editing of frame-by-frame annotations at scene and object level, providing a number of predefined shape drawing facilities for the localisation of objects. To speed up the process of annotation, the automatic propagation of descriptions is supported. Specifically, by choosing to copy a description from one frame to another, the description is assigned to all frames in between as well. In case of object level descriptions, subsequent editing allows to adjust the exact position at each frame. Object level descriptions can be also propagated through dragging while the video is playing. ViPER-GT uses a simple proprietary XML-based format, which for the case of descriptive annotations can be edited by the user so as to include additional attributes.

For a more detailed list and pointers to additional tools, the reader is referred to the Tools&Resources[82] report of the W3C Multimedia Semantics Incubator Group.

[78] http://www.w3.org/XML/Linking

[79] http://dewey.at.northwestern.edu/ppad2/

[80] http://www.isi.edu/in-notes/rfc2396.txt

[81] http://viper-toolkit.sourceforge.net/

[82] http://www.w3.org/2005/Incubator/mmsem/wiki/Tools_and_Resources

4.9 Discussion

As illustrated in the aforementioned descriptions, video annotation tools make a rather poor utilisation of Semantic Web technologies and formal meaning, XML being the most common choice for the capturing and representation of the produced annotations. The use of MPEG-7 based descriptions, may constitute a solution towards standardised video descriptions, yet raises serious issues with respect to the automatic processing of annotations, especially the descriptive ones, at a semantic level. The localisation of temporal segments is performed mostly manually, indicating the issues involved in automatically identifying the time interval corresponding to the semantic notion addressed by the annotation; only Advene, SVAT and VideoAnnex perform automatic shot detection. Furthermore, VideoAnnex, VIA and SVAT are the only ones that offer selection and annotation of spatial regions on frames of the video, as well. Anvil has recently presented a new annotation mechanisms called spatiotemporal coding aiming to support point and region annotation, yet currently only points are supported.

A challenging issue in video annotation concerns the representation of structural and by consequence temporal information in an effective manner so as to avoid overwhelming volumes of metadata. This issue has been already pointed out in relevant studies on multimedia ontologies and the resulting metadata complexity, while it should be noted that many of the MPEG-7 based video annotation tools follow simplified translations in order to avoid the cumbersome and complex MPEG-7 specifications. Finally, it is interesting to note, that although descriptors representation, if not extraction, constitutes a consideration for image annotation tools, this is not the case for video tools.

Table 2 summarises the comparative study of the examined video annotation tools with respect to the *Input & Output* and *Annotation Level* criteria described in Section 2. Regarding the miscellaneous criteria, as illustrated in the individual tools descriptions, no tool provides support for collaborative annotation and all tools are stand-alone applications, publicly available for non-commercial use[83].

It worths noticing that most annotation tools offer a variety of additional functionalities, in order to satisfy varying user needs. Facilitating the retrieval task seems to be a common demand, since almost all the tools have embedded mechanisms for allowing the user to efficiently search and/or navigate through the annotations. Moreover, the visualisation of annotations is enhanced by the annotated concepts' timeline views that most of the tools support. Concluding, we should add that the choice of a tool depends primarily on the intended context of usage, which provides the specifications regarding the annotation dimensions supported, and subsequently on the desired formality of annotations.

5 Conclusions

In the previous Sections, we reviewed representative examples of well known image and video annotation tools with respect to a number of criteria, such defined

[83] In many cases, the source code is available for research purposes

Tool	**Input & Output**		**Annotation level**			
	Metadata Format	*Annotation Vocabulary*	*Metadata Type*	*Granularity*	*Localisation*	*Expressivity*
VIA	XML	U: domain ontology (OWL), free text T: customised structural XML schema	descriptive, structural administrative	video, video segment, frame, moving region, image, still region	time interval, free hand, polygon, rectangle	concepts
Ontolog	RDF	U: domain ontology (RDFS) T: Dublin Core ES T: Ontolog Schema ontology	descriptive, structural administrative, structural	video, video segment	time interval,	concepts, relations
VideoAnnex	MPEG-7/XML	U: XML, free text T: MPEG-7	descriptive, structural, administrative	video, video segment, frame, still region	time interval, rectangle	concepts relations
Advene	custom XML	U: free text (specific format)	descriptive, structural administrative	video, video segment	time interval,	concepts relations
Elan	custom XML	U: free text, keywords	descriptive	video video segment	time interval	concepts
Anvil	custom XML	U: XML Schema T: customised structural XML schema	descriptive, structural administrative	video, points video segment	time interval	concepts
SVAT	MPEG-7/XML	U: free text, keywords T: MPEG-7	descriptive, structural administrative	video, video segment frame, still region	time interval	concepts

Table 2. Video annotation tools summarisation. In the *A*nnotation Vocabulary field, “U” denotes user-entered vocabularies, while “T” refers to vocabularies embedded within the tool, and thus hidden to the user.

as to provide a common framework of reference for assessing the suitability and interoperability of annotations under different context of usages.

The afore presented overview suggests that semantic image annotation tools appear to follow up with relevant research advances. Domain specific ontologies are supported by the majority of tools for the representation of subject matter descriptions. Moreover, influenced by initiatives addressing multimedia ontologies, many tools utilise corresponding ontologies for the representation of structural, localisation and low-level descriptors information. With the exception of KAT though, the defined ontologies constitute simplified versions of corresponding state of the art initiatives. Consequently, given the detail of modelling provided by the state of the art ontologies, a reasonable expectation would be to investigate the use of those ontologies in manual annotation tools, especially with respect to practical scalability and complexity concerns [38, 39].

Semantic video annotation tools on the contrary, present a rather gloomy scenery with respect to interoperability concerns both at semantic and syntactic level. Almost none of the examined tools supports the use of ontologies for descriptive annotations. The case is similar for structural and localisation information, where proprietary schemas are used in proprietary formats. VideoAnnEx and SVAT following the MPEG-7 specifications alleviate to an extend interoperability issues by promoting specific annotation vocabularies and schemes. Yet, apart from the XML-based issues regarding the lack of declarative semantics, the free text formats of MPEG-7 semantic descriptions perpetuate the limitations related to keyword-based search and retrieval. Consequently, a general subject of consideration relates to the low outreach and uptake of results in multimedia annotation research to practical video annotation systems [40].

However, the level of correspondence between research outcomes and implemented annotation tools is not the sole subject for further investigation. Research in multimedia annotation, and by consequence into multimedia ontologies, is not restricted to the representation of the different annotation dimensions involved. A critical issue is the delineation of multimedia specific annotation schemes, i.e. the conceptualisation and modelling of how the various annotations pertaining to multimedia assets can be interlinked in a scalable, yet effective manner. Apart from research activities conducted individually [9, 41, 42, 30, 13, 12] collective initiatives have been pursued. The W3C Multimedia Semantics Incubator Group[84] (MMSEM), constitutes a prominent such activity that has produced a number of comprehensive reports including "Image annotation on the Semantic Web"[85], Multimedia Vocabularies[86] and Tools&Resources[87], as well as a proposal towards a "Multimedia Annotation Interoperability Framework"[88]. As a continuation of the efforts initiated within MMSEM, further manifesting the strong emphasis placed upon achieving cross community multimedia data integration, two new

[84] http://www.w3.org/2005/Incubator/mmsem/

[85] http://www.w3.org/2005/Incubator/mmsem/XGR-image-annotation/

[86] http://www.w3.org/2005/Incubator/mmsem/XGR-vocabularies/

[87] http://www.w3.org/2005/Incubator/mmsem/wiki/Tools_and_Resources

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

W3C Working Groups have been charted, the Media Annotation[89] and Media Fragments[90] WGs. The objective of the Media Annotation WG is to provide an ontology infrastructure to facilitate cross-community data integration of information related to multimedia objects in the Web, while the Media Fragments one addresses the identification of temporal and spatial media fragments in the Web using URIs.

Concluding, semantic image and video annotation constitute particularly active research fields, faced with intricate challenges. Such challenges issue not only from implications related to the sheer volume of content available, but also from the dynamically evolving context of intelligent content management services as delineated by the growth of Semantic Web technologies, as well as by new powerful and exciting concepts introduced by initiatives such as Web 2.0, Linked-Data[91] and Web Services.

6 Acknowledgement

This work was partially supported by the European Commission under contract FP6-027538 BOEMIE. We would also like to express our gratitude to the various tool authors who provided us with useful feedback.

References

1. Smeulders, A., Worring, M., .Santini, S., .Gupta, A., .Jain, R.: Content-based image retrieval at the end of the early years. IEEE Trans. Pattern Anal. Mach. Intell. **22** (2000) 1349–1380
2. Hauptmann, A., Yan, R., Lin, W.: How many high-level concepts will fill the semantic gap in news video retrieval? In: 6th ACM International Conference on Image and Video Retrieval (CIVR), Amsterdam, The Netherlands. (2007) 627–634
3. Snoek, C., Huurnink, B., Hollink, L., de Rijke, M., Schreiber, G., Worring, M.: Adding semantics to detectors for video retrieval. IEEE Transactions on Multimedia **9** (2007) 975–986
4. Hanjalic, A., Lienhart, R., Ma, W., Smith, J.: The holy grail of multimedia information retrieval: So close or yet so far away. IEEE Proceedings, Special Issue on Multimedia Information Retrieval **96** (2008) 541–547
5. Nack, J.: Mpeg-7: Overview of description tools. IEEE MultiMedia **9** (2002) 83–93
6. Salembier, P., Manjunath, B., Sikora, T.: Introduction to MPEG 7: Multimedia Content Description Language. (2002)
7. van Ossenbruggen, J., Nack, F., Hardman, L.: That obscure object of desire: Multimedia metadata on the web, part 1. IEEE MultiMedia **11** (2004) 38–48
8. Nack, F., van Ossenbruggen, J., Hardman, L.: That obscure object of desire: Multimedia metadata on the web, part 2. IEEE MultiMedia **12** (2005) 54–63
9. Hunter, J.: Adding Multimedia to the Semantic Web: Building an MPEG-7 Ontology. In: Proc. The First Semantic Web Working Symposium (SWWS), California, USA. (2001)

[89] http://www.w3.org/2008/01/media-annotations-wg.html

[90] http://www.w3.org/2008/WebVideo/Fragments/

[91] http://linkeddata.org/

10. Tsinaraki, C., Polydoros, P., Christodoulakis, S.: Integration of owl ontologies in mpeg-7 and tv-anytime compliant semantic indexing. In: 16th International Conference on Advanced Information Systems Engineering (CAiSE), Riga, Latvia, June 7-11. (2004) 398–413
11. R. Garcia, O.C.: Semantic Integration and Retrieval of Multimedia Metadata. In: Proc. International Semantic Web Conference (ISWC), Galway, Ireland. (2005)
12. Dasiopoulou, S., Tzouvaras, V., I.Kompatsiaris, Strintzis, M.: Capturing mpeg-7 semantics. In: Proc. International Conference on Metadata and Semantics (MTSR), Corfu, Greece. (2007)
13. Arndt, R., Troncy, R., Staab, S., Hardman, L., Vacura, M.: Comm: designing a well-founded multimedia ontology for the web. In: Proc. International Semantic Web Conference (ISWC), Busan, Korea. (2007)
14. Jorgensen, C., Jaimes, A., Benitez, A., Chang, S.: A conceptual framework and empirical reserach for classifying visual descriptors. J. of the American Society for Information Science and Technology (JASIST) **52** (2001) 938–947
15. Hollink, L., Schreiber, G., Wielinga, B., Worring, M.: Classification of user image descriptions. Int. J. Hum.-Comput. Stud. **61** (2006) 601–626
16. Saathoff, C., Schenk, S., Scherp, A.: Kat: the k-space annotation tool. In: Poster Session, Int. Conf. on Semantic and Digital Media Technologies (SAMT), Koblenz, Germany. (2008)
17. Gangemi, A., Guarino, N., Masolo, C., Oltramari, A., Schneider, L.: Sweetening ontologies with dolce. In: 13th International Conference on Knowledge Engineering and Knowledge Management (EKAW), Siguenza, Spain, October 1-4. (2002) 166–181
18. Gangemi, A.: Ontology design patterns for semantic web content. In: 4th International Semantic Web Conference (ISWC), Galway, Ireland, November 6-10. (2005) 262–276
19. MPEG-7 MDS: ISO/IEC 15938-5:2003 information technology. Multimedia Content Description Interface - Part 5: Multimedia Description Schemes, First Edition (2001)
20. MPEG-7 Visual: ISO/IEC 15938-3:2001 information technology. Multimedia Content Description Interface - Part 3: Visual, First Edition (2001)
21. Halaschek-Wiener, C., Golbeck, J., Schain, A., Grove, M., Parsia, B., Hendler, J.: Annotation and provenance tracking in semantic web photo libraries. In: Proc. International Provenance and Annotation Workshop (IPAW), Chicago, IL, USA. (2006) 82–89
22. Chakravarthy, A., Ciravegna, F., Lanfranchi, V.: Aktivemedia: Cross-media document annotation and enrichment. In: Poster Proceedings of 5th International Semantic Web Conference (ISWC), Athens, GA, USA. (2006)
23. Petridis, K., Anastasopoulos, D., Saathoff, C., Timmermann, N., Kompatsiaris, I., Staab, S.: M-ontomat-annotizer:image annotation. linking ontologies and multimedia low-level features. In: 10th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES), Engineered Applications of Semantic Web Session (SWEA), Bournemouth, U.K. (2006)
24. Simou, N., Tzouvaras, V., Avrithis, Y., Stamou, G., Kollias, S.: A visual descriptor ontology for multimedia reasoning. In: Proc. of Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS), Montreux, Switzerland. (2005)
25. Lux, M., Becker, J., Krottmaier, H.: Caliph & emir: Semantic annota-tion and retrieval in personal digital photo libraries. In: Proc. of Advanced Information Systems Engineering (CAiSE), Velden, Austria. (2003)

26. MPEG-7: ISO/IEC 15938. Multimedia Content Descritpion Interface (2001)
27. Miller, M., McCathieNevile, C.: Semantic web tools to help authoring: A semantic web image annotation tool. In: SWAD-Europe Deliverable 9.3. (2001)
28. Russell, B., Torralba, A., Murphy, K., Freeman, W.: Labelme: A database and web-based tool for image annotation. International Journal of Computer Vision **77** (2008) 157–173
29. Rubin, D., Rodriguez, C., Shah, P., Beaulieu, C.: ipad: Semantic annotation and markup of radiological images. In: in Proc. of Annual American Medical Informatics Association (AMIA) Symposium, Washington, DC. (2008) 626–630
30. Tsinaraki, C., Polydoros, P., Christodoulakis, S.: Interoperability support between mpeg-7/21 and owl in ds-mirf. IEEE Trans. Knowl. Data Eng. **19** (2007) 219–232
31. Troncy, R., Celma, O., Little, S., Garcia, R., Tsinaraki, C.: Mpeg-7 based multimedia ontologies: Interoperability support or interoperability issue? In: Proc. Workshop on Multimedia Annotation and Retrieval enabled by Shared Ontologies (MARESO), Genova, Italy. (2007) 2–16
32. MPEG-7 XM: MPEG-7 Visual eXperimentation Model (XM), Version 10.0, Doc. N4062. ISO/IEC/JTC1/SC29/WG11 (2001)
33. Rutledge, L.: Smil 2.0: Xml for web multimedia. Internet Computing **5** (2001) 78–84
34. Kipp, M.: Anvil - a generic annotation tool for multimodal dialogue. In: in Proc. 7th European Conf. on Speech Communication and Technology (Eurospeech), Aalborg, Denmark. (2001)
35. Kipp, M.: Spatiotemporal coding in anvil. In: in Proc. 6th international conference on Language Resources and Evaluation (LREC), Marrakech, Morocco. (2008)
36. Schallauer, P., Ober, S., Neuschmied, H.: Efficient semantic video annotation by object and shot re-detection. In: Posters and Demos Session, 2nd International Conference on Semantic and Digital Media Technologies (SAMT), Koblenz, Germany. (2008)
37. Schroeter, R., Hunter, J., Kosovic, D.: Vannotea - a collaborative video indexing, annotation and discussion system for broadband networks. In: in Proc. of Workshop on "Knowledge Markup and Semantic Annotation (K-CAP), Florida, US. (2003)
38. Hausenblas, M., Bailer, W., Bürger, T., Troncy, R.: Deploying multimedia metadata on the semantic web. In: Posters and Demos Session, 2nd International Conference on Semantic and Digital Media Technologies (SAMT), Genoa, Italy. (2007)
39. Vacura, M., Svátek, V., Saathoff, C., ranz, T., Troncy, R.: Describing low-level image features using the comm ontology. In: in Proc. 15th International Conference on Image Processing (ICIP), San Diego, California, USA. (2008) 49–52
40. Bürger, T., Hausenblas, M.: Why real-world multimedia assets fail to enter the semantic web. In: in Proc. of the Semantic Authoring, Annotation and Knowledge Markup Workshop (SAAKM), Whistler, British Columbia, Canada. (2007)
41. Lagoze, C., Hunter, J.: The abc ontology and model. Journal of Digital Information **2** (2001)
42. Troncy, R., Bailer, W., Hausenblas, M., Hofmair, P., Schlatte, R.: Enabling multimedia metadata interoperability by defining formal semantics of mpeg-7 profiles. In: Proc. International Conference on Semantics And digital Media Technology (SAMT), Athens, Greece. (2006) 41–55