Multimed Tools Appl (2017) 76:14437–14460
DOI 10.1007/s11042-016-3705-7

RDF-powered semantic video annotation tools with concept mapping to Linked Data for next-generation video indexing: a comprehensive review

Leslie F. Sikos [1]

Received: 5 December 2015 / Revised: 24 April 2016 / Accepted: 24 June 2016 /
Published online: 19 August 2016

Abstract Video annotation tools are often compared in the literature, however, most reviews mix unstructured, semi-structured, and the very few structured annotation software. This paper is a comprehensive review of video annotations tools generating structured data output for video clips, regions of interest, frames, and media fragments, with a focus on Linked Data support. The tools are compared in terms of supported input and output data formats, expressivity, annotation specificity, spatial and temporal fragmentation, the concept mapping sources used for Linked Open Data (LOD) interlinking, provenance data support, and standards alignment. Practicality and usability aspects of the user interface of these tools are highlighted. Moreover, this review distinguishes extensively researched yet discontinued semantic video annotation software from promising state-of-the-art tools that show new directions in this increasingly important field.

Keywords Video annotation · Multimedia semantics · Spatiotemporal fragmentation · Video scene interpretation · Multimedia ontologies · Hypervideo application

1 Introduction

While there are embedded metadata formats available for images, audio files, and videos, they are often limited to technical characteristics and many of them are not structured. In contrast to MP3 files, which often provide information about the album, singer or band, release year, genre, and might even include the lyrics, video files typically do not have any information embedded to them about the depicted concepts, actors, or the plot. Online video information

✉ Leslie F. Sikos
leslie.sikos@flinders.edu.au

[1] Centre for Knowledge and Interaction Technologies, School of Computer Science, Engineering and Mathematics, Flinders University, GPO Box 2100, Adelaide, SA 5001, Australia

retrieval often relies on the text surrounding the media files embedded to web pages, mainly due to the huge "semantic gap" between what computers and humans understand (automatically extractable low-level features and sophisticated high-level content descriptors) [36].

While some semantic image annotation tools (e.g., *K-Space Annotation Tool*, *PhotoStuff*, *AktiveMedia*, *M-OntoMat-Annotizer*, *SWAD*, *Annotorius*, *Pundit*, *ImageSnippets*) could be used for annotating video frames (as still images), semantic video annotation requires specialized tools for representing temporal and other information unique to videos.[1] For this reason, manual, semi-automatic, and automatic annotation tools have been introduced over the years for the semantic enrichment of audiovisual contents.

Video management systems date back to the early 1990s with video databases and conceptual models annotating audiovisual contents with unstructured data (e.g., *OVID* [33], *Vane* [10]), all of which were different in terms of spatial and temporal data representation, semantic expressiveness, and flexibility.

Less than a decade after the introduction of video annotation software tools, they began to support structured data. While video annotation software generating semi-structured output, such as *MuViNo,*[2] *EXMARaLDA,*[3] the *VideoAnnEx Annotation Tool*[4], *ELAN*[5], the *Video Image Annotation Tool (VIA),*[6] the *Semantic Video Annotation Suite (SVAS),*[7] *VAnalyzer,*[8] the *Semantic Video Content Annotation Tool (SVCAT),*[9] *Anvil,*[10] and the video annotation tool of Aydınlılar and Yazıcı [1], have been developed in parallel with tools powered by the Resource Description Framework (RDF), this paper focuses only on those video annotation tools that produce output in structured data formats including RDF or exclusively in RDF.

2 Semantic video annotation

State-of-the-art structured video annotation incorporates multimedia signal processing and formally grounded knowledge representation including, but not limited to, video feature extraction, machine learning, ontology engineering, and multimedia reasoning.

2.1 Feature extraction for concept mapping

A wide range of well-established algorithms exists for automatically extracting low-level video features, as for example, fast color quantization to extract the dominant colors [44] or Gabor filter banks to extract homogeneous texture descriptors [43]. There are also advanced algorithms for video content analysis, such as the Viola-Jones and Lienhart-Maydt object detection algorithms [26, 41], and the SIFT, SURF, and ORB keypoint detection algorithms [23, 28, 35]. The corresponding

[1] There are also cross-media annotation tools, such as IMAS and YUMA, which provide annotations for multiple media types (see Section 3).
[2] http://vitooki.sourceforge.net/components/muvino/code/index.html
[3] http://www.exmaralda.org/en/tool/exmaralda/
[4] http://www.research.ibm.com/VideoAnnEx/
[5] https://tla.mpi.nl/tools/tla-tools/elan/
[6] http://sourceforge.net/projects/via-tool/
[7] http://www.joanneum.at/en/digital/productssolutions/sematic-video-annotation.html
[8] https://www.dimis.fim.uni-passau.de/iris/index.php?view=vanalyzer
[9] https://www.dimis.fim.uni-passau.de/MDPS/de/mitglieder/30-german-articles/forschung/projekte/33-svcat.html
[10] http://www.anvil-software.org

descriptors can be used as positive and negative examples in machine learning, such as support vector machines (SVM) and Bayesian networks, for keyframe analysis, face recognition, and video scene understanding.

While useful, many automatically extracted low-level video features are inadequate for representing video semantics. For example, annotating the dominant color or color distribution of a frame does not provide the meaning of the visual content. In contrast, high-level descriptors are suitable for video concept mapping, but they often rely on human knowledge, experience, and judgment. However, manual video concept tagging is very time-consuming, might be biased, too generic, or inappropriate, which has led to the introduction of collaborative semantic video annotation, where multiple users annotate the same resources and improve each other's annotations [12]. User-supplied annotations can be curated using natural language processing to eliminate duplicates and typos, and filter out incorrectly mapped concepts. The integrity of manual annotations captured as structured data can be confirmed automatically using LOD definitions. Research results for high-level concept mapping in constrained videos, such as medical videos [16] or sport videos [3], are already promising, however, concept mapping in unconstrained videos is still a challenge [22].

The next section details multimedia ontology engineering best practices to create machine-interpretable high-level descriptors and reuse de facto standard definitions to formally represent human knowledge suitable for the automated interpretation of video contents.

2.2 Knowledge representation of video scenes

Logical formalization of video contents can be used for video indexing, scene interpretation, and video understanding [37]. Structured knowledge representations are usually expressed in, or based on, RDF, which can describe machine-readable statements in the form of subject-predicate-object triples, e.g., scene-depicts-person. The corresponding concepts are defined in 1) controlled vocabularies, consisting of three countably finite sets of symbols: a set N_C of concept names, a set N_R of role names, and a set N_I of individual names, or 2) ontologies, i.e., quadruples expressed as $O=(\mathbf{C}, \Sigma, \mathbf{R}, \mathbf{A})$, where $\mathbf{C}$ is a set of concept expressions, $\mathbf{R}$ is a set of binary relationships between concepts from $\mathbf{C}$, $\langle \mathbf{C}, \Sigma \rangle$ is the taxonomic structure of concepts from $\mathbf{C}$, and $\mathbf{A}$ is a set of axioms. Vocabularies are defined in RDF Schema (RDFS), an extension of RDF to create vocabularies and taxonomies, and complex ontologies are defined in the fully-featured ontology language OWL (Web Ontology Language). Related terms and factual data might also be derived from other structured data sources, such as knowledge bases and LOD datasets. For example, to declare a video clip depicting a person in a machine-readable format, a vocabulary or ontology which provides the formal definition of video clips and their features is needed, such as the Clip vocabulary from Schema.org,[11] because it is suitable for declaring the director, file format, language, encoding, etc. of video clips (`schema:Clip`). The "depicts" relationship is defined by the Friend of a Friend (FOAF) vocabulary[12] (`foaf:depicts`). The definition of "Person" can be used from `schema:Person`, which defines typical properties of a person, including, but not limited to, name, gender, birthdate, and nationality.[13]

[11] http://schema.org/Clip
[12] http://xmlns.com/foaf/spec/
[13] It is a common practice to abbreviate terms using the namespace mechanism, which relies on a prefix to eliminate long (often symbolic) URIs, such that `schema:` abbreviates http://schema.org/ and `foaf:` abbreviates http://xmlns.com/foaf/0.1/. For example, `foaf:depicts` abbreviates http://xmlns.com/foaf/0.1/depicts.

Some well-established commonsense knowledge bases and their corresponding general-purpose upper ontologies that can be used for describing the concepts depicted in videos are Wordnet[14] and OpenCyc.[15] There are also more specific ontologies for this purpose, such as the Large-Scale Concept Ontology for Multimedia (LSCOM) [31].

The spatiotemporal annotation of video events requires even more specialized ontologies, such as the SWRL Temporal Ontology[16] and VidOnt,[17] along with Media Fragment URI 1.0 identifiers.[18]

If the concepts to describe belong to a knowledge domain not covered by existing ontologies, one can create a new ontology by formally defining the classes, properties, and their relationships, preferably in OWL, with a logical underpinning in description logics (DL). DL-based ontologies do not specify a particular interpretation based on a default assumption; instead, they consider all possible cases in which the axioms are satisfied.

In the case of the most expressive OWL 2 ontologies, the set of *role expressions* **R** over a signature is defined as $\mathbf{R} ::= U \mid \mathrm{N}_R \mid \mathrm{N}_{\bar{R}}$, where U represents the universal role, N_R is a set of roles, and $\mathrm{N}_{\bar{R}}$ is a set of negated role assertions. The concept expressions of an OWL 2 ontology are defined as the set $\mathbf{C} ::= \mathrm{N}_C \mid (C \sqcap C) \mid (C \sqcup C) \mid \neg C \mid \top \mid \bot \mid \exists R.C \mid \forall R.C \mid \geqslant n\ R.C \mid \leqslant n\ R.C \mid \exists R.Self \mid \{\mathrm{N}_I\}$, where n is a non-negative integer, C, D represent concepts, and R represents roles. Based on these sets, $\mathcal{SROIQ}^{(\mathcal{D})}$ axioms can be defined as general concept inclusions (GCIs) of the form $C \sqsubseteq D$ and $C \equiv D$ for concepts C and D (terminological knowledge, TBox), individual assertions of the form $C(\mathrm{N}_I)$, $R(\mathrm{N}_I, \mathrm{N}_I)$, $\mathrm{N}_I \approx \mathrm{N}_I$, or $\mathrm{N}_I \not\approx \mathrm{N}_I$ (assertional knowledge, ABox), and role assertions of the form $R \sqsubseteq S$, $R \equiv S$, $R_1 \circ \ldots \circ R_n \sqsubseteq S$, $Asy(R)$, $Ref(R)$, $Irr(R)$, $Dis(R, S)$ for roles R, R_i, and S (role box, RBox) [24], as summarized in Table 1.

Interpretation $\mathcal{I}$ consists of a set $\Delta^{\mathcal{I}}$ (the domain of $\mathcal{I}$) and an interpretation function $\cdot^{\mathcal{I}}$, which maps each atomic concept A to a set $A^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}}$, each atomic role R to a binary relation $R^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}}$, and each individual name a to an element $a^{\mathcal{I}} \in \Delta^{\mathcal{I}}$. Similar to the constructors, the formal meaning of the axioms is defined by their model-theoretic semantics, as shown in Table 2.

As an example, assume a file of a video scene, namely the climax of the movie "The Good, the Bad, and the Ugly" with the trio, Tuco, Blondie, and Angel Eyes, portrayed by Eli Wallach, Clint Eastwood, and Lee Van Cleef, respectively (United Artists, 1966). The aim is to describe the video scene with spatiotemporal data, maintain provenance data, and annotate the movie characters depicted in the various regions of the scene, along with the actors who played in the corresponding roles. Using description logic, the knowledge representation of this video scene can be formalized as follows:[19]

```
Scene(TRIO)
Movie(THEGOODTHEBADANDTHEUGLY)
sceneFrom(TRIO, THEGOODTHEBADANDTHEUGLY)
hasStartTime(TRIO, 02:40:28)
```

[14] http://wordnet-rdf.princeton.edu/ontology
[15] https://sourceforge.net/projects/texai/files/open-cyc-rdf/1.1/
[16] http://swrl.stanford.edu/ontologies/built-ins/3.3/temporal.owl
[17] http://vidont.org/vidont.ttl
[18] https://www.w3.org/TR/media-frags/
[19] In the example, concept names are written in PascalCase, role names in camelCase, and individual names in ALL CAPS, as per description logic best practices.

Table 1 Syntax and semantics of $\mathcal{SROIQ}$ constructors

	Syntax	Semantics
Atomic concept	A	$A^{\mathcal{I}}$
Intersection	$C \sqcap D$	$C^{\mathcal{I}} \cap D^{\mathcal{I}}$
Union	$C \sqcup D$	$C^{\mathcal{I}} \cup D^{\mathcal{I}}$
Complement	$\neg C$	$\Delta^{\mathcal{I}} \setminus C^{\mathcal{I}}$
Top concept	$\top$	$\Delta^{\mathcal{I}}$
Bottom concept	$\bot$	$\varnothing$
Existential quantification	$\exists R.C$	$\{x \mid \text{some } R^{\mathcal{I}}\text{-successor of } x \text{ is in } C^{\mathcal{I}}\}$
Universal quantification	$\forall R.C$	$\{x \mid \text{all } R^{\mathcal{I}}\text{-successors of } x \text{ are in } C^{\mathcal{I}}\}$
At-least restriction	$\geqslant n\ R.C$	$\{x \mid \text{at least } n\ R^{\mathcal{I}}\text{-successors of } x \text{ are in } C^{\mathcal{I}}\}$
At-most restriction	$\leqslant n\ R.C$	$\{x \mid \text{at most } n\ R^{\mathcal{I}}\text{-successors of } x \text{ are in } C^{\mathcal{I}}\}$
Local reflexivity	$\exists R.Self$	$\{x \mid \langle x, x\rangle \in R^{\mathcal{I}}\}$
Nominal	$\{a\}$	$\{a^{\mathcal{I}}\}$
Atomic role	R	$R^{\mathcal{I}}$
Inverse role	R^{-}	$\{\langle x, y\rangle \mid \langle y, x\rangle \in R^{\mathcal{I}}\}$
Universal role	U	$\Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}}$
Individual name	a	$a^{\mathcal{I}}$

$C, D \in \mathbf{C}$ concepts, $A \in \mathrm{N}_C$ concept name, $R \in \mathbf{R}$ role, $a, b \in \mathrm{N}_I$ individual names

```
duration(TRIO, 00:04:40)
hasFinishTime(TRIO, 02:45:08)
depicts(TRIO, MexicanStandoff)
MovieCharacter(TUCO)
portrayedBy(TUCO, ELIWALLACH)
MovieCharacter(BLONDIE)
portrayedBy(BLONDIE, CLINTEASTWOOD)
BLONDIE ≈ MANWITHNONAME
MovieCharacter(ANGELEYES)
```

Table 2 Syntax and semantics of $\mathcal{SROIQ}$ axioms

		Syntax	Semantics
TBox	Concept inclusion	$C \sqsubseteq D$	$C^{\mathcal{I}} \subseteq D^{\mathcal{I}}$
	Concept equivalence	$C \equiv D$	$C^{\mathcal{I}} = D^{\mathcal{I}}$
ABox	Concept assertion	$C(a)$	$a^{\mathcal{I}} \in C^{\mathcal{I}}$
	Role assertion	$R(a, b)$	$\langle a^{\mathcal{I}}, b^{\mathcal{I}}\rangle \in R^{\mathcal{I}}$
	Individual equality	$a \approx b$	$a^{\mathcal{I}} = b^{\mathcal{I}}$
	Individual inequality	$a \not\approx b$	$a^{\mathcal{I}} \neq b^{\mathcal{I}}$
RBox	Role inclusion	$R \sqsubseteq S$	$R^{\mathcal{I}} \subseteq S^{\mathcal{I}}$
	Role equivalence	$R \equiv S$	$R^{\mathcal{I}} = S^{\mathcal{I}}$
	Complex role inclusion	$R_1 \circ R_2 \sqsubseteq S$	$R_1^{\mathcal{I}} \circ R_2^{\mathcal{I}} \subseteq S^{\mathcal{I}}$
	Role disjointness	$Dis(R, S)$	$R^{\mathcal{I}} \cap S^{\mathcal{I}} = \varnothing$

```
portrayedBy(ANGELEYES, LEEVANCLEEF)
depicts(TRIOROI1, TUCO)
depicts(TRIOROI2, BLONDIE)
depicts(TRIOROI3, ANGELEYES)
```

The concepts, roles, and individuals of this example are defined by multiple ontologies, which have to be declared in order to obtain the full identifiers according to the corresponding namespaces. Due to the relationship between DLs and OWL, the above example can be translated to any RDF serialization. In Turtle, for example, a shot of the Mexican standoff scene of "The Good, the Bad, and the Ugly" can be described as follows:

```
@prefix dbpedia: <http://dbpedia.org/resource/> .
@prefix foaf: <http://xmlns.org/foaf/0.1/> .
@prefix temporal:
<http://swrl.stanford.edu/ontologies/built-ins/3.3/temporal.owl#> .
@prefix schema: <http://schema.org/> .
@prefix vidont: <http://vidont.org/> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
vidont:TheGoodTheBadAndTheUgly a schema:Movie .
<http://example.com/trio.mp4> a vidont:Scene ;
vidont:sceneFrom vidont:TheGoodTheBadAndTheUgly ;
temporal:hasStartTime "02:40:28"^^xsd:dateTime ;
temporal:duration "P4M40S"^^xsd:duration ;
temporal:hasFinishTime "02:45:08"^^xsd:dateTime ;
foaf:depicts dbpedia:Mexican_standoff .
vidont:Tuco a vidont:MovieCharacter ; vidont:portrayedBy
vidont:EliWallach .
vidont:Blondie a vidont:MovieCharacter ;
vidont:portrayedBy vidont:ClintEastwood ;
owl:sameIndividualAs dbpedia:Man_with_No_Name .
vidont:AngelEyes a vidont:MovieCharacter ;
vidont:portrayedBy vidont:LeeVanCleef .
<http://example.com/trio.mp4#t=45,46&xywh=538,258,105,511>
foaf:depicts vidont:Tuco .
<http://example.com/trio.mp4#t=45,46&xywh=1161,286,47,157>
foaf:depicts vidont:Blondie .
<http://example.com/trio.mp4#t=45,46&xywh=1306,206,166,530>
foaf:depicts vidont:AngelEyes .
```

This example incorporates concept and role definitions and individuals from DBpedia,[20] FOAF, the SWRL Temporal Ontology, Schema.org, and VidOnt, as well as XML Schema datatypes. The namespaces are declared using `@prefix`. Note that `a` is a shorthand notation for the `rdf:type` predicate. Also note that a series of RDF triples sharing the same subject are abbreviated by stating the subject once, and then separating each predicate-object pair using a semicolon.

[20] http://dbpedia.org

Spatial information is declared using Media Fragment URI 1.0 identifiers. In this example, the position of the selected shot is specified as Normal Play Time according to RFC 2326, which is the default time scheme for media fragment URIs. The movie characters are represented by the top left corner coordinates and the dimensions of the imaginary surrounding rectangles, as shown in Fig. 1.

In contrast to the tree structure of XML documents, RDF-based knowledge representations can be visualized as graphs. RDF graphs are directed, labeled graphs in which the nodes are the resources and values, and the arrows assign the predicates (see Fig. 2).

Because the RDF graphs that share the same resource identifiers naturally merge together, interlinking LOD concepts and individuals (e.g., `dbpedia:Mexican_standoff`, `dbpedia:Man_with_No_Name`) makes the above graph part of the LOD Cloud.[21]

2.3 Ontology-based video indexing and retrieval

Concept relationships are proven to be valuable knowledge resources that can enhance the effectiveness of video retrieval even for ambiguous queries [46]. RDF-based data is inherently machine-interpretable and unambiguous, which can be exploited in video indexing and retrieval. Video annotation tools often apply concept detection scores for a region, keyframe, shot, video clip, or entire video, which tend to perform better than feature extraction based on local descriptors (e.g., SIFT, HoG, HoF) [29, 30]. Each score value between 0 and 1 indicates the presence or absence of a concept. The higher the score, the higher the likelihood of the depiction of the concept. To improve concept detection accuracy, the relation between depicted concepts can be analyzed by computing co-occurrence, visual descriptors, and hybrid semantic similarity, which leverages contextual information for video classification [5]. Description logic-based semantic video annotations can also be complemented by rule-based representations to improve the integrity and correctness of the interpretation [45]. For example, Mexican standoffs can be described with SWRL rules as follows:

```
foaf:depicts(?s,?p1) ∧ foaf:depicts(?s,?p2) ∧
foaf:depicts(?s,?p3) ∧ vidont:isHolding(?p1,
dbpedia:pistol) ∧ vidont:isHolding(?p2, dbpedia:pistol) ∧
vidont:isHolding(?p3, dbpedia:pistol) ∧
vidont:isLookingAt(?p1,?p2) ∨ vidont:isLookingAt(?p1,?p3)
∧ vidont:isLookingAt(?p2,?p1) ∨
vidont:isLookingAt(?p2,?p3) ∧ vidont:isLookingAt(?p3,?p1)
∨ vidont:isLookingAt(?p3,?p2) ∧
temporal:hasStartTime(?e1,?Ste1) ∧
temporal:hasFinishTime(?e1,?ste1) ∧
temporal:hasStartTime(?e2,?Ste2) ∧
temporal:hasFinishTime(?e2,?ste2) ∧ temporal:before(e1,
e2) → foaf:depicts(?c, dbpedia:Mexican_standoff)
```

The semantically enriched representation can be used by automated mechanisms to recognize the same type of video scenes in different video resources. Moreover, reasoners can use such machine-interpretable descriptions to automatically infer new statements to achieve knowledge discovery.

Once correctly identified, concepts can be interlinked with related data across LOD datasets. In contrast to website contents retrieved through keyphrase-based web search, RDF-based

[21] http://lod-cloud.net

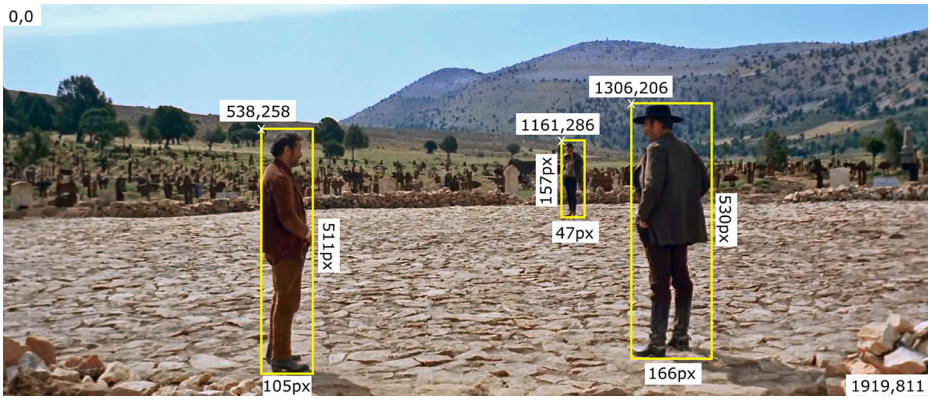

Fig. 1 The top left corner coordinates and dimensions of RoIs can be used for spatial annotation of movie characters. Movie scene by United Artists

knowledge representations can be queried and manipulated manually or programmatically through the very powerful SPARQL query language [25]. SPARQL queries might include multiple questions in a single query to answer complex questions that cannot be formulated as keywords used in traditional keyphrase-based web search. Furthermore, they can be executed not only on a single dataset, but also across multiple datasets using federated queries.

2.4 Primary application areas

Multimedia ontologies can be used for high-level scene interpretation, such as event detection [2], moving object detection and tracking [14], and even human intention detection [27]. High-level scene interpretation is suitable for, among others, classification, video surveillance [34], intelligent video analytics, and real-time activity monitoring [9]. Most of these tasks are performed by reasoning over the video contents to recognize situations and temporal events based on human knowledge formally described as ontology concepts, roles, individuals, and rules. By representing fuzzy relationships between the context and depicted concepts of video contents, both deductive and abductive reasoning can be performed [13].

3 A retrospective survey of semantic video annotation tools

Veggie, one of the first video annotation tools to generate RDF output, was introduced by Hunter and Newmarch in 1999 [20]. The Java application produced Dublin Core-based metadata descriptions and video summaries for MPEG-1 videos.

In 2002, Heggland developed *OntoLog*, an application for searching and browsing temporal media metadata by leveraging metadata exchange using RDF, and SMIL for interoperability between different media players [19]. The software supported high-level descriptors not only for entire videos, but also for video shots and frames. Ontolog incorporated RDFS for representing depicted concepts and the relationships between them.

Vannotea, also released in 2002, was a prototype system for real-time collaborative, synchronous indexing, browsing, annotation, and commentary of MPEG-2 videos (see Fig. 3).

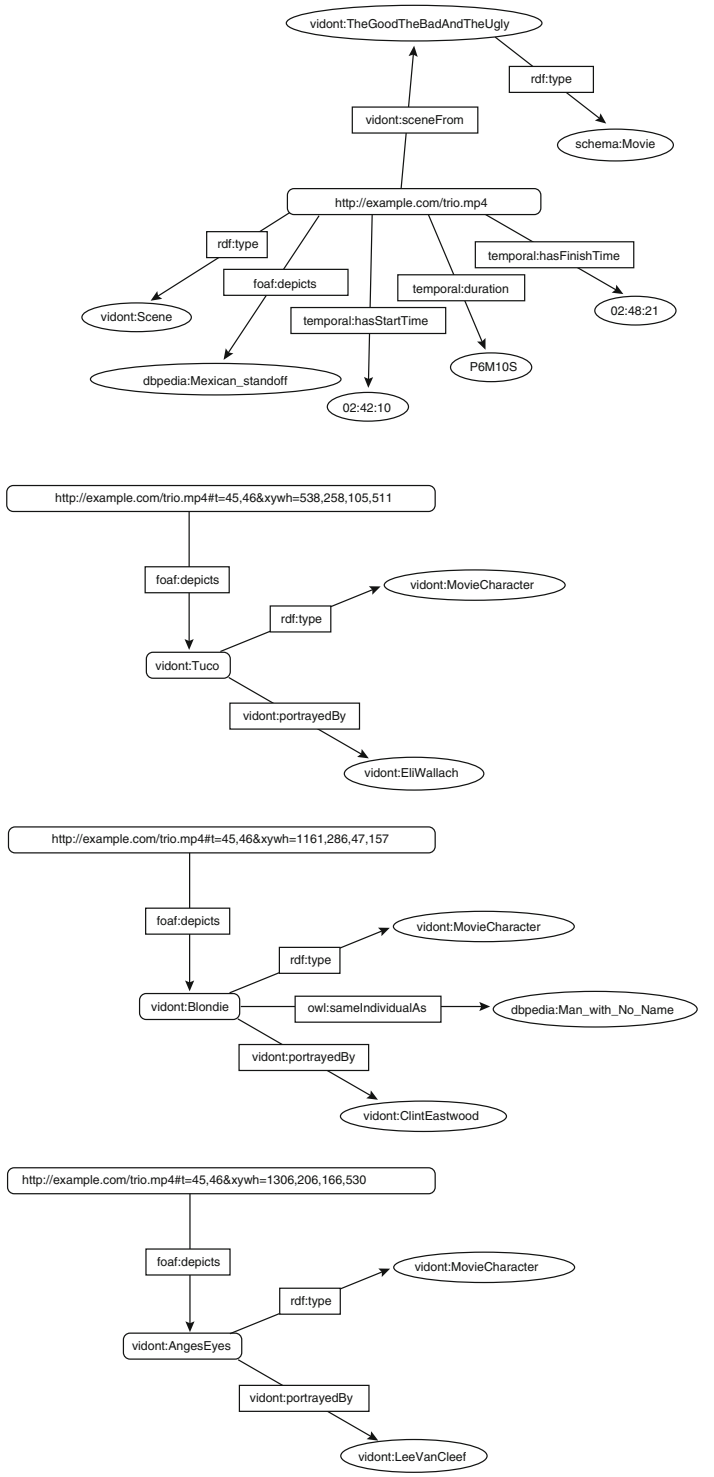

Fig. 2 A graph visualizing the RDF triples

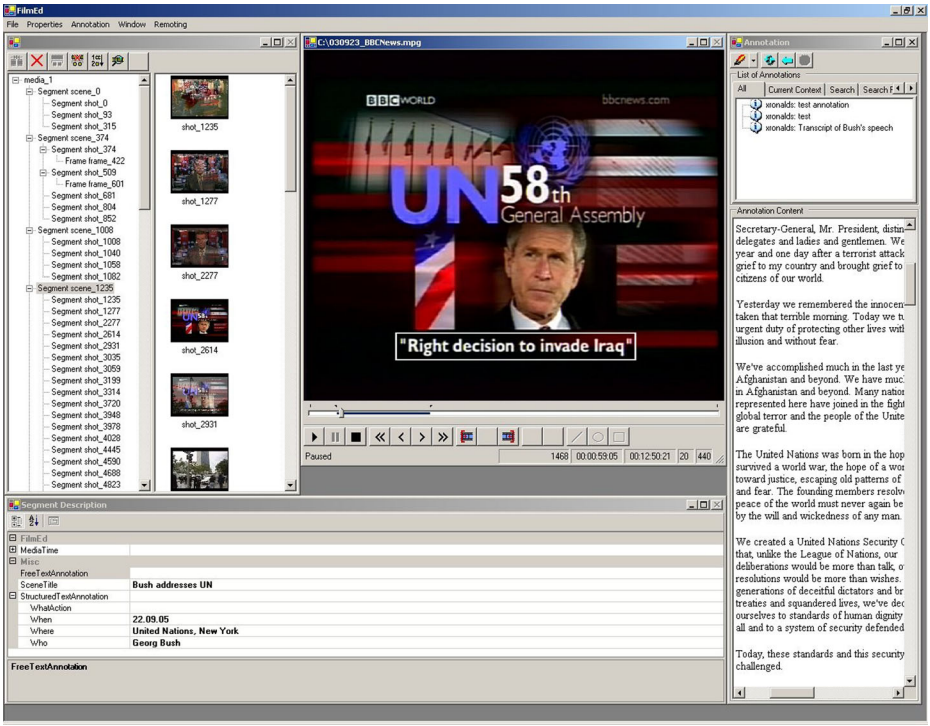

Fig. 3 Vannotea, an early implementation of structured video annotation tools [21]

Vannotea was based on W3C's *Annotea,*[22] and used RDF for knowledge representation and XPointer to link the annotations to the video resources.

Advene[23] (*Annotate Digital Video, Exchange on the Net*), also released in 2002, was developed over a decade, and is still available to download today. In Advene, users can annotate video fragments at arbitrary positions, save the semantically enriched videos, and play the videos with the associated semantics (see Fig. 4).

Being a proprietary binary format, the native file format of Advene is not ideal. Nevertheless, the software is Linked Data-ready, because every annotation, relation, and view is identified by a URI and RDF/XML output is supported. The software does not incorporate multimedia ontologies beyond FOAF and Dublin Core though.

OntoMedia[24] was developed in 2006 for large multimedia collection management using Semantic Web technologies. The graphical user interface of this standalone Java application offered easy metadata indexing and video retrieval. OntoMedia accepted any input media supported by QuickTime or the Java Media Framework, and could generate RDF and relational database output.

Also in 2006, Bertini and his colleagues developed the *Multimedia Ontology Manager (MOM)* to combine multimedia ontology engineering with automatic annotation, and generate textual and auditory commentary for video sequences [7]. The automatic video annotation was performed for entire video clips by using similarity checking between visual ontology concepts and extracted

[22] https://www.w3.org/2001/Annotea/
[23] http://advene.org
[24] http://www.ontomedia.de

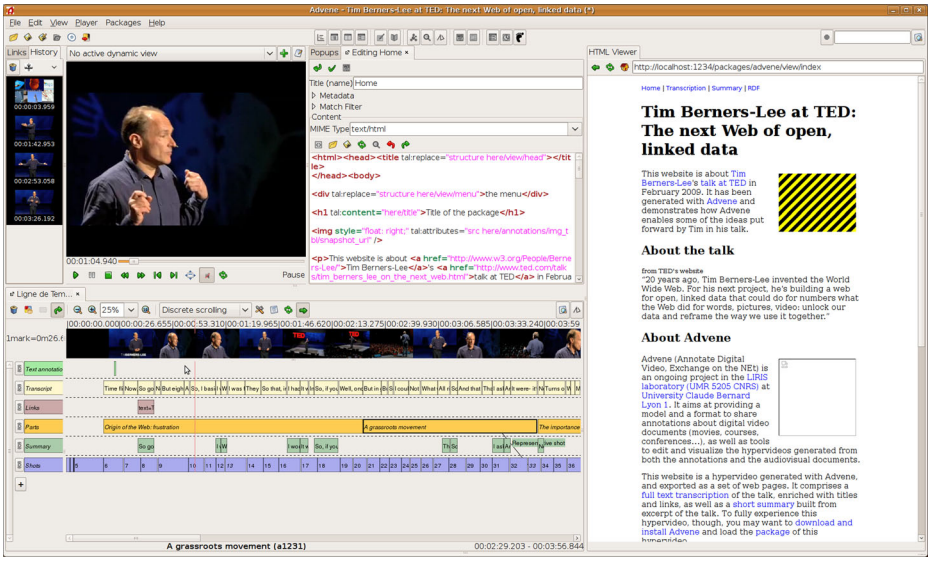

Fig. 4 Temporal segment annotation with Advene

clips, and for video sequences by using composite concept patterns. Video clip sequences were annotated with predefined articulated sentences curated by the RACER reasoner.

Annomation,[25] published in 2008 as a collaborative Linked Data-driven narrative hypervideo application, allowed users to semantically annotate video resources using controlled vocabularies defined in the LOD Cloud. It was restricted to predefined videos hosted by the service, and the semantic annotations were saved in a local repository, making them inaccessible to external semantic agents.

In 2009, the *LEMO Annotation Framework* was released, providing a uniform, multimedia-enabled annotation model. LEMO addressed video fragments using the MPEG-21 Part 17 (Fragment Identification of MPEG Resources) standard, and exposed data as Linked Data [17].

IMAS, also published in 2009, was a web-based annotation tool for media resources that generated annotations using a set of proprietary ontologies [42]. IMAS imported images and videos from a media repository, but did not support media fragments. The output of IMAS was suitable for producers only, rather than general-purpose online publishing.

SemWebVid[26] was an Ajax web application released in 2010, which automatically generated YouTube[27] video descriptions in RDF, taking manually added tags and closed captions into account. SemWebVid implemented natural language processing APIs to analyze the descriptors, and mapped the results to LOD concepts, using the DBpedia, Uberblic, Any23, and rdf:about APIs, and the now-discontinued Sindice API. Provenance data was color-coded, which was an original idea, however, the resulting text was not always easy to read (see Fig. 5).

The application implemented YouTube Data API v2, which has been replaced by the backward incompatible YouTube Data API v3 in April 2015. Consequently, SemWebVid is not working anymore.

[25] http://annomation.open.ac.uk
[26] http://tomayac.com/semwebvid/
[27] https://www.youtube.com

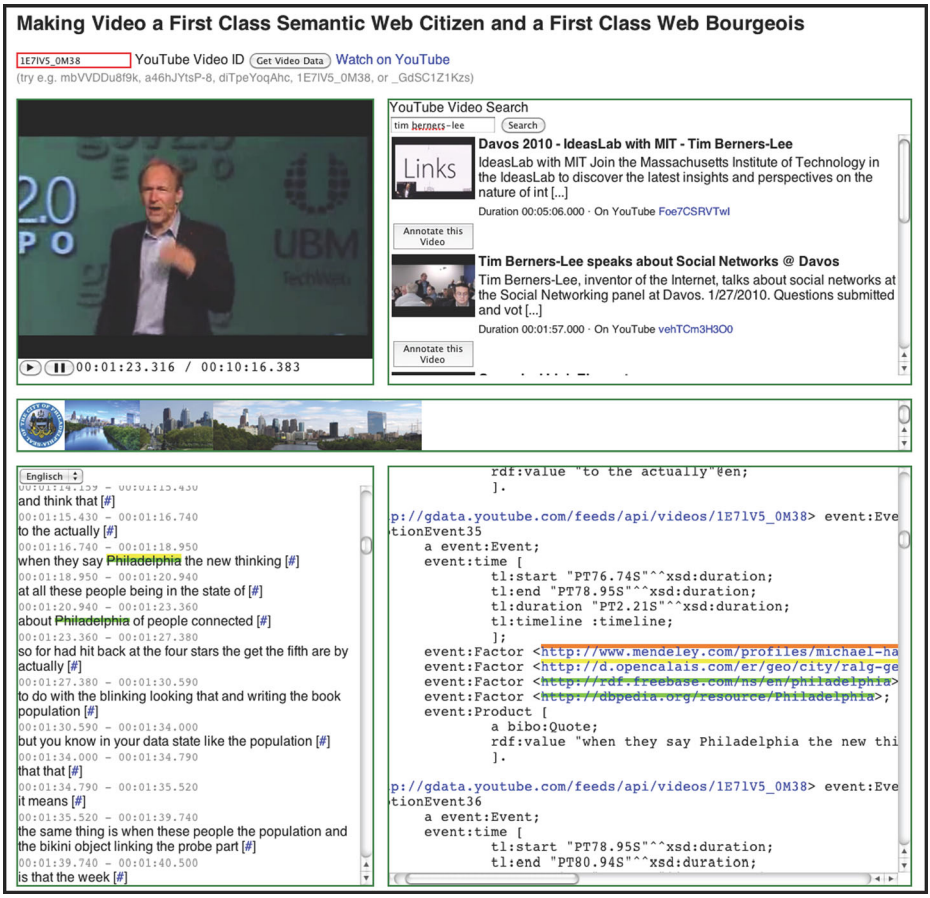

Fig. 5 Comprehensive concept mapping to LOD in SemWebVid [40]

Also in 2010, Choudhury and Breslin introduced a framework to annotate and retrieve online videos with light semantics, and integrate structured video annotations into the Linked Open Data Cloud by reusing important terms from Dublin Core, FOAF, and SKOS [11]. In the same year, the *EuropeanaConnect Media Annotation Suite (ECMAS)* was released, which used both plain text and semantic tags for the knowledge representation of depicted concepts [18].

Pan, an ontology-based online video annotation tool to import and edit OWL ontologies with MPEG-7 alignment, was also developed in 2010 [6]. Pan can browse videos, provide a mechanism for the user to select concepts from an ontology structure, add and edit annotations, and load previously saved annotations. The annotations are managed by another tool, Orione, an ontology-based search engine. Pan is not future-ready, because it was written in Adobe Flex and ActionScript 3, i.e., it requires the Flash plugin, which is now deprecated in favor of HTML5 and JavaScript.

YUMA[28] was an annotation software released in 2011, which supported image, audio, and video files [39]. YUMA suggested DBpedia and GeoNames[29] terms, and exported the results to RDF using a proprietary vocabulary, along with LEMO and Open Annotation.[30]

[28] https://github.com/paulweichhart/client-suite
[29] http://www.geonames.org/ontology/
[30] http://www.openannotation.org/spec/core/

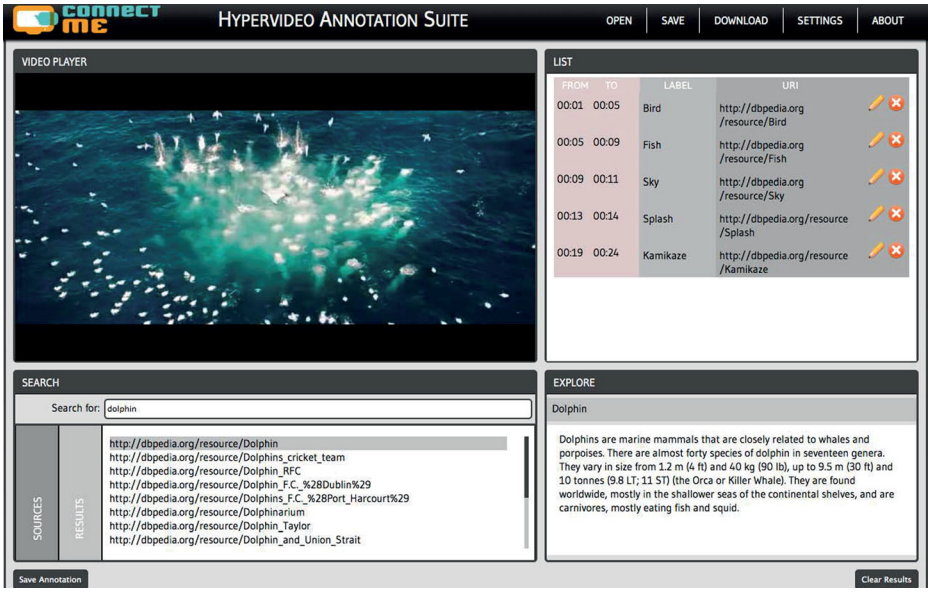

Fig. 6 The ConnectME hypervideo annotation suite incorporated temporal information with labels and LOD concepts

The *ConnectME* toolset was released in 2012, comprising of an HTML5-based semantically enriched video player and an online video annotation tool. The ConnectME framework identified, annotated, and deployed video concepts as Linked Data [32]. The user interface displayed timestamps next to the video player, along with the corresponding labels and LOD URIs (see Fig. 6), although using prefixes would have made the URIs more compact, easier to read, and easier to fit in the program window (more space would have been reserved for the video player, the search box, and the explorer).

SemTube, a YouTube video annotation prototype, was also released in 2012. It expressed the context of YouTube videos in RDF/OWL and OAC [15]. SemTube used RDF, Linked Data, SPARQL, and RESTful APIs for data import and export. The data retrieval from SemTube annotations supported keyword-based and Linked Data-powered faceted search, and SPARQL queries. One of the preferred LOD datasets for concept mapping in both SemTube and SemWebVid was Freebase, which has been discontinued in 2015, with some of its articles transferred to Wikidata.[31]

Many of the annotation tools discussed above were built with proprietary APIs, which have been changed over the years, breaking the functionality of the original program code. The original version of those tools that have not been updated to reflect these API changes stopped working partially or completely. Also, support is limited for most software prototypes, which often had a domain name registered at the time of their release, but have later been discontinued. Veggie, OntoLog, OntoMedia, MOM, the LEMO Annotation Framework, IMAS, ECMAS, ConnectME, SemTube, YUMA, and Vannotea are not available online anymore, while SemWebVid and Annomation were available at the time of writing, but were not working.

[31] https://www.wikidata.org

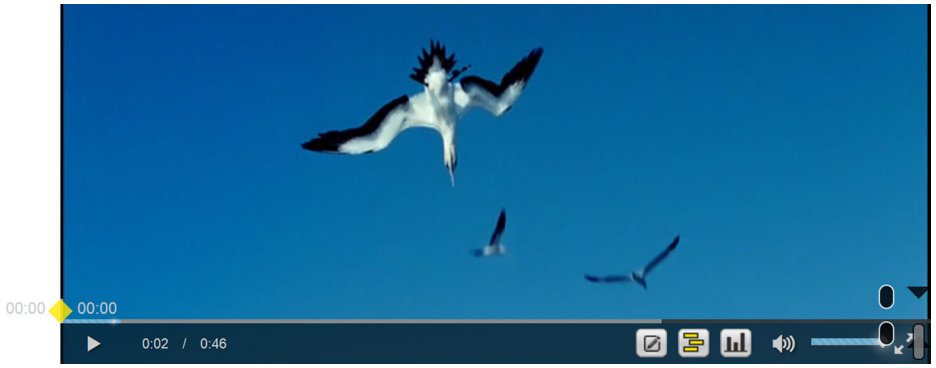

Fig. 7 In Open Video Annotation, users can take notes on the timeline, view existing annotations, and play annotated video fragments individually

4 State-of-the-art structured video annotation tools

The *TV Metadata Generator*[32] was released by Eurecom as part of the LinkedTV project in 2011. Based on the local or online input video file, TV-Anytime or EXMARaLDA metadata files, or SRT subtitle files, the software automatically converts television content metadata into RDF. However, the software cannot generate RDF based on the video content alone, and is basically limited to the serialization of existing textual data as structured data. The *LinkedTV Editor*[33] provides a user interface for broadcasting services, which uses the automatically generated annotations of LinkedTV for the rapid generation of contextual information queues.

Open Video Annotation[34] is based on open source JavaScript libraries, such as Video.js,[35] Annotator,[36] and RangeSlider.[37] The developers claim that the software is compliant with W3C's Open Annotation data formats. Open Video Annotation was designed to provide an intuitive interface for semantic tagging and the playback of semantically enriched videos (see Fig. 7).

At the time of writing, the Open Video Annotation was still under development, with many functionalities of the demo not yet working.

MyStoryPlayer is a video player capable of the semantic enrichment of multi-angle videos, and was specifically designed for educational videos. It provides an interface for interactive user annotations to be used in action, gesture, and posture analysis, with a focus on the formal representation of relationships between depicted elements in RDF [4]. MyStoryPlayer powers the website of the European eLibrary for Performing Arts (ECLAP),[38] and provides not only general and technical metadata, such as title and duration, but also timestamp-based data, which can be used to annotate presentations, human dialogues, and arbitrary video events (see Fig. 8).

SemVidLOD[39] is a software prototype for the semantic enrichment of online video resources, video files, and streaming media with high-level descriptors using terms from the LOD Cloud. SemVidLOD implements VidOnt, the most expressive decidable multimedia

[32] http://linkedtv.eurecom.fr/tv2rdf
[33] http://editortoolv2.linkedtv.eu
[34] http://www.openvideoannotation.org
[35] http://videojs.com
[36] http://annotatorjs.org
[37] https://github.com/andreruffert/rangeslider.js
[38] http://www.eclap.eu
[39] http://vidont.org/semvidlod/

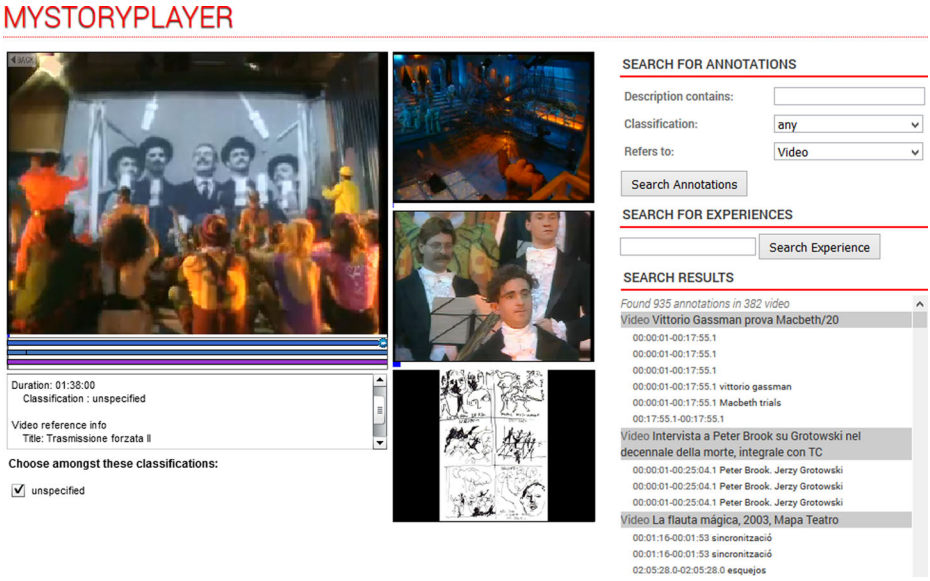

Fig. 8 In MyStoryPlayer, metadata and classification are coupled with timestamp-based snapshot comments

ontology to date [38], to express administrative, technical, and licensing metadata, as well as sophisticated high-level content descriptions in RDF.

5 Comparison of structured video annotation tools

Based on the review of the state of the art, semantic video annotation tools differ in terms of characteristics and functionality due to the following technical features:

- *Expressivity.* The semantic richness of annotations is determined by the expressivity of the controlled vocabularies and ontologies used for the knowledge representation of the depicted concepts. Some tools are restricted to proprietary controlled vocabulary terms, while others do not provide suggestions but accept arbitrary data.
- *Annotation level.* Annotation software usually specialize in particular types of metadata (technical, administrative, licensing), content descriptors (high-level descriptors), multimedia descriptors (low-level descriptors), structural descriptors (spatial, temporal, and spatiotemporal descriptors), or a combination of these.

  - *Low-level descriptor support.* Capability to annotate automatically extractable low-level features of videos, such as motion trajectory.
  - *High-level descriptor support.* Capability to precisely annotate depicted concepts and individuals, such as a person, a car, or a building.
  - *Spatial fragment support.* Enables working with a portion of the media (Region of Interest, RoI) to represent information about the depicted space, for example to annotate a tumor in a medical video or an actor in a movie.
  - *Temporal fragment support.* Enables frame sequence segmentation within videos to represent time and events, such as video scenes or a goal in a soccer match video.

- *Standards alignment.* Standards make it possible for various platforms and computer systems to communicate with each other and exchange data efficiently, regardless of their structural and functional differences. Standards alignment determines whether standards and de facto standards are implemented (e.g., MPEG-7, Dublin Core, Open Annotation). Video annotation software prototypes may use proprietary formats and mechanisms, which are difficult to implement in large-scale, heterogeneous multimedia systems. Poor standard support, including proprietary vocabulary use, negatively affects interoperability. Open standards are likely to be implemented globally, so they should be preferred.
- *Supported input and output data formats.* Some annotation tools are designed for a particular video compression or codec only (MPEG-1, MPEG-2, MPEG-4/AVC H.264, etc.), or accept nothing else but YouTube videos by URL. Ideally, the set of supported formats would include at least the current industry standard video file formats. Some video annotation software can handle any kind of video file format, as long as the related codecs are installed on the system.
- *Signal processing integration.* By integrating signal processing algorithms to video annotation tools, the annotation of low-level features becomes seamless, although the majority of automatically extracted low-level descriptors cannot be used for high-level scene interpretation, as mentioned earlier.
- *Linked Data support.* Supporting best practices for publishing structured data, called Linked Data [8], is crucial for semantic multimedia applications. Linked Data provides unique URIs for each video object, media fragment, keyframe, and RoI, along with a mechanism to interlink depicted concepts with arbitrary definitions from the LOD Cloud, and differentiates media files from web resources that convey information about them. Linked Data support is crucial for future multimedia applications.
- *Automation.* While manual annotations can be the most sophisticated and accurate annotations, they depend on the experience and background of the user, can be misspelt and ambiguous, do not always incorporate the most relevant keywords, and might be biased by personal preferences. Semi-automatic (supervised) and automated (unsupervised) annotation would be desired to address the above issues of manual annotations and to efficiently generate annotations to the rapidly growing number of online videos.
- *Provenance data support.* Storing data source information (preferably by using the PROV Ontology)[40] is beneficial for video annotations derived from diverse data sources. Provenance data makes data quality assessment easier, can be used to find similar or related resources, and makes LOD concept interlinking more efficient.
- *RDF output.* All structured video annotation software must support RDF output in a standard serialization, such as RDF/XML or Turtle. HTML5 Microdata, RDFa, and JSON-LD are also desirable, which can be directly embedded to the website markup.
- *Architecture.* Web-based semantic video annotation tools are preferred to their desktop counterparts due to benefits such as platform-independence, interoperability, and global availability.
- *Built-in Video Player.* Ideally, video annotation tools are embedded to a video player for seamless annotation and hypermedia playback.

[40] http://www.w3.org/TR/prov-o/

Less objective features include user-friendliness, documentation quality and coverage, user support (examples, tutorial videos, contact), long-term availability, licensing, and whether the software is open source.

The following sections compare structured video annotation tools from the four main perspectives: standards support, input and output data formats, concept mapping sources, and spatiotemporal fragmentation.

5.1 Standards alignment

Several multimedia and web standards are required, and often implemented, in structured video annotation tools to provide backward- and forward-compatibility and interoperability. Standards are vital to gain widespread use, obtain optimality in terms of file structure and code length, and consider global needs. The most common international standards in semantic video annotation are DVD-Video (media and format are defined by multiple standards, e.g., ISO/IEC 16448:2002[41] and ECMA-267,[42] ISO/IEC 25434:2008),[43] MPEG-7 (ISO/IEC 15938),[44] MPEG-21 Part 17 (ISO/IEC 14496–17:2006),[45] Uniform Resource Locators (IETF RFC 1738),[46] and Dublin Core (IETF RFC 5013,[47] ISO 15836:2009,[48] ANSI/NISO Z39.85).[49] The technical specifications used by semantic video annotation tools that have not been standardized officially by a standardization body yet are used globally are known as de facto standards; these include W3C recommendations, such as RDF[50] and SKOS,[51] Open Annotation, and the Media Fragment URI (see Table 3).

Open standards are preferred to proprietary implementations, such as the temporal annotation of Annomation, the spatial and temporal fragmentation of Advene and SemTube, and proprietary ontologies, e.g., the SALERO ontologies used by IMAS or the LinkedTV ontology implemented by the LinkedTV Editor.

Standards alignment is a necessary but not always sufficient requirement for the long-term viability of software tools. For example, the development of Vannotea has been discontinued regardless of its MPEG-7 compliance, but the implementation of de facto standards, such as that of the Media Fragment URI or Open Annotation, can explain the continuing success and ongoing development of LinkedTV Editor, SemVidLOD, and TV Metadata Generator.

5.2 Supported data formats

The supported input and output file formats can be crucial for the usability of a software tool, especially with the large variety of video container formats, file formats, and codecs. Open formats are preferred to proprietary formats for two reasons. Firstly, open formats make software development easier and interoperability wider. Secondly, the popularity of tools that support only proprietary file formats tends to decrease faster than the ones that implement

[41] http://standards.iso.org/ittf/PubliclyAvailableStandards/c035641_ISO_IEC_16448_2002%28E%29.zip
[42] http://www.ecma-international.org/publications/files/ECMA-ST/Ecma-267.pdf
[43] http://www.iso.org/iso/iso_catalogue/catalogue_ics/catalogue_detail_ics.htm?csnumber=51140
[44] http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=34228
[45] http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=39478
[46] https://www.ietf.org/rfc/rfc1738.txt
[47] https://www.ietf.org/rfc/rfc5013.txt
[48] http://www.iso.org/iso/catalogue_detail.htm?csnumber=52142
[49] http://www.niso.org/apps/group_public/project/details.php?project_id=105
[50] https://www.w3.org/TR/rdf11-concepts/
[51] https://www.w3.org/TR/skos-reference/

Table 3 Standards supported by structured video annotation tools

Tool	Standards	De Facto Standards
Advene	DVD-Video, Dublin Core	XML, RDF
Annomation	Dublin Core	RDF, SKOS
ConnectME	–	RDF, Open Annotation, Media Fragment URI
IMAS	–	RDF
LEMO	MPEG-21 Part 17	RDF, FLV
LinkedTV Editor	–	RDF, Open Annotation, Media Fragment URI
MyStoryPlayer	Dublin Core	RDF, Open Annotation
Open Video Annotation	–	RDF, Open Annotation
SemTube	–	RDF
SemVidLOD	–	RDF, Media Fragment URI
SemWebVid	–	RDF
TV Metadata Generator	–	RDF, Open Annotation, Media Fragment URI
Vannotea	MPEG-7, MPEG-21, Dublin Core	RDF, ABC
YUMA	–	RDF, Open Annotation

standardized and open formats. This might be the reason behind the discontinuation of LEMO, which was designed for the Flash Video format now replaced by HTML5.

The support for multiple input data formats is a user expectation, which is why many video annotation tools can open a variety of video files and handle video streams (see Table 4).

While Linked Data output is expected from semantic video annotation tools, dependence on a particular LOD dataset can be a major design issue. A good example is the now-discontinued SemTube, which implemented Freebase as the primary LOD dataset for interlinking, which became obsolete and succeeded by Wikidata. However, the still popular DBpedia and GeoNames were the primary LOD datasets of Annomation, ConnectME, and YUMA, all of which have also been discontinued. This suggests that the long-term viability of LOD dataset URLs generated by semantic video annotation tools does not guarantee the success of these tools.

The tools for annotating YouTube videos (SemTube, SemWebVid) rely on the proprietary YouTube API. Consequently, such tools cannot be used for annotating videos stored on other video sharing portals, such as Vimeo[52] and LiveLeak,[53] and since the corresponding API might change over time, future updates are required for the upcoming versions of the API or else the tools will stop working.

Those tools that accept video input via URL can use the corresponding URLs directly to add context to RDF triples and provide a graph identifier for quads (subject-predicate-object-graph name) so that they become globally interpretable. The software tools that open local files only do not have this kind of unique web identifier for the media resources by default.

[52] https://vimeo.com
[53] http://www.liveleak.com

Table 4 Supported data formats of structured video annotation tools

Tool	Input	Output
Advene	Video file, DVD-Video, video stream	Hypervideo with RDF description
Annomation	Video file from local repository	LOD (DBpedia, Dewey, and GeoNames concept links)
ConnectME	Video file via URL or from repository	LOD
IMAS	Image or video from media repository	RDF
LEMO	Flash Video	RDF
LinkedTV Editor	Video registered on the LinkedTV Platform	LOD (DBpedia suggestions)
MyStoryPlayer	Image, audio, or video file	LOD
Open Video Annotation	Arbitrary video, description, and tags	Semantically enriched hypervideo
SemTube	YouTube video via URL	LOD
SemVidLOD	Video file, video via URL, or streaming media	LOD
SemWebVid	YouTube video via URL	LOD
TV Metadata Generator	Video via URL	LOD (DBpedia)
Vannotea	Video file	RDF
YUMA	Audio, image, or video by URL	LOD

5.3 Ontology use

The primary concept mapping sources vary greatly among structured video annotation tools, and include ontologies such as Dublin Core,[54] the Ontology for Media Resources,[55] FOAF,[56] Open Annotation,[57] and Representing Content in RDF.[58] Some tools also allow arbitrary ontologies so that they are not limited to the concepts of the primary concept mapping sources (see Table 5).

As shown above, not all annotation tools allow arbitrary ontologies, which is a huge limitation even if standardized ontologies are used as the primary concept mapping sources. However, arbitrary ontology support is not necessarily sufficient to gain global adoption, as was the case of Advene, ConnectME, IMAS, and SemTube.

5.4 Spatiotemporal annotation support

Structured video annotation tools support either spatial or temporal annotations, both spatial and temporal annotations, or neither (see Table 6).

The most common spatiotemporal annotation format in semantic video annotation is W3C's Media Fragment URI. LEMO used the MPEG-21 Part 17 standard for the same

[54] http://dublincore.org/documents/dcmi-terms/
[55] https://www.w3.org/TR/mediaont-10/
[56] http://xmlns.com/foaf/spec/
[57] http://www.openannotation.org/ns/
[58] https://www.w3.org/2011/content

Table 5 Ontology use of semantic video annotation tools

Tool	Primary Vocabularies and Ontologies	Arbitrary Ontology
Advene	Dublin Core, FOAF	+
Annomation	Dublin Core, FOAF, SKOS	–
ConnectME	Proprietary, Ontology for Media Resources, Annotation Ontology, Dublin Core, FOAF, Open Annotation, Representing Content in RDF	+
IMAS	SALERO ontologies	+
LEMO	Proprietary	
LinkedTV Editor	LinkedTV, Ontology for Media Resources, Annotation Ontology, Dublin Core, FOAF, Open Annotation, Representing Content in RDF	–
MyStoryPlayer	Annotation Ontology, Dublin Core, FOAF, Open Annotation, Representing Content in RDF	–
Open Video Annotation	Annotation Ontology, Dublin Core, FOAF, Open Annotation, Representing Content in RDF	–
SemTube	–	+
SemVidLOD	VidOnt, Schema, Dublin Core, FOAF	+
SemWebVid	–	–
TV Metadata Generator	LinkedTV, Ontology for Media Resources, Annotation Ontology, Dublin Core, FOAF, Open Annotation, Representing Content in RDF	–
Vannotea	Dublin Core	–
YUMA	LEMO, Annotation Ontology, Dublin Core, FOAF, Open Annotation, Representing Content in RDF	–

purpose. Some tools (Advene, Annomation, Vannotea) implemented proprietary mechanisms that cannot be processed by any other software tool but the ones that introduced them.

Table 6 Spatiotemporal annotation support of semantic video annotation tools

Tool	Spatial Fragmentation	Temporal Fragmentation
Advene	Proprietary	Proprietary
Annomation	–	Proprietary
ConnectME	Media Fragment URI	Media Fragment URI
IMAS	–	–
LEMO	MPEG-21 Part 17	MPEG-21 Part 17
LinkedTV Editor	Media Fragment URI	Media Fragment URI
MyStoryPlayer	–	–
Open Video Annotation	–	Media Fragment URI
SemTube	Proprietary	Proprietary
SemVidLOD	W3C Media Fragment URI	W3C Media Fragment URI
SemWebVid	–	–
TV Metadata Generator	Media Fragment URI	Media Fragment URI
Vannotea	Proprietary	Proprietary
YUMA	Proprietary	Proprietary

6 Conclusions

In contrast to review papers of multimedia annotation tools that mix the annotation of still images and videos, or do not differentiate between semi-structured and structured output, this comprehensive review explicitly enumerates the milestones of structured video annotation tools, highlights their limitations, and suggests required features for upcoming software tools.

Semantic video annotation tools face many challenges including, but not limited to, the wide variety of video codecs, the lack of standardized video ontologies, the vast number of video resources, not to mention the inherent ambiguity of audiovisual contents. Unstructured comments, labels, and tags of traditional video annotation systems come with a degree of formalism inadequate for efficient automated processing. To address this limitation, OWL ontologies and Linked Data can be used for structured video annotation, which can be generated semi-automatically or automatically with semantic video annotation tools. Multimedia ontology engineering has been demonstrated through structured video annotations that leverage standardized definitions as well as concepts from a state-of-the-art ontology, VidOnt, to combine the representation of video fragments, regions of interest, depicted concepts, and spatiotemporal information. The strengths and weaknesses of ontology-based video scene representation have also been discussed, and the limitations of structured video annotation tools have been highlighted. Despite the potential of these software tools, the development, maintenance, and support of most semantic multimedia annotation software prototypes mentioned in the literature have been discontinued. There are very few structured video annotation tools that are being actively developed. The state-of-the-art tools differ significantly in terms of supported input data formats, ontology use, standards alignment, Media Fragment URI implementation, and Linked Data support. Some software tools rely heavily on proprietary APIs and software libraries that might change over time. Fortunately, the implementation of Open Annotation and other de facto standard ontologies is more and more common. Based on this review it can be concluded that the global adoption of semantic video annotation tools depends on a number of characteristics, including the implemented technologies and standards, the supported input and output file formats, the primary concept mapping sources, Linked Data integration, and the option to use arbitrary ontologies and spatiotemporal fragmentation.

To meet the challenges of future web applications and improve the efficiency of concept mapping, information fusion would be desired, so that manually added tags, closed captions, and audio analysis could support the selection of the most relevant concepts. To provide Linked Data-powered structured annotations for video resources, online semantic multimedia annotation tools are preferred to desktop tools, using technologies such as HTML5, JavaScript, and Ajax in combination with Semantic Web standards. This can be achieved by a paradigm shift in the software design of semantic multimedia annotation tools, namely by adding the capability to open videos by URL (as opposed to opening video files from local repositories), supporting Linked Data and spatiotemporal fragmentation, and using modern multimedia ontologies for high-level concept descriptors. The interoperable video annotation output leverages Semantic Web standards for easy data distribution, sharing, reuse, and personalization, setting a new direction for online video sharing and next-generation video retrieval.

References

1. Aydınlılar M, Yazıcı A (2013) Semi-automatic semantic video annotation tool. In: Gelenbe E, Lent R (eds) Computer and information sciences III, pp 303–310. doi:10.1007/978-1-4471-4594-3_31
2. Ballan L, Bertini M, Del Bimbo A, Seidenari L, Serra G (2011) Event detection and recognition for semantic annotation of video. Multimed Tools Appl 51(1):279–302. doi:10.1007/s11042-010-0643-7
3. Ballan L, Bertini M, Del Bimbo A, Serra G (2010) Semantic annotation of soccer videos by visual instance clustering and spatial/temporal reasoning in ontologies. Multimed Tools Appl 48: 313–337. doi:10.1007/s11042-009-0342-4
4. Bellini P, Nesi P, Serena M (2015) MyStoryPlayer: experiencing multiple audiovisual content for education and training. Multimed Tools Appl 74:8219–8259. doi:10.1007/s11042-014-2052-9
5. Benmokhtar R, Huet B (2014) An ontology-based evidential framework for video indexing using high-level multimodal fusion. Multimed Tools Appl 73(2):663–689. doi:10.1007/s11042-011-0936-5
6. Bertini M, d'Amico G, Ferracani A, Meoni M, Serra G (2010) Sirio, Orione and Pan: an integrated web system for ontology-based video search and annotation. In: ACM international conference on multimedia, Firenze, Oct 25–29, 2010, pp 1625–1628. doi:10.1145/1873951.1874305
7. Bertini M, Del Bimbo A, Torniai C, Cucchiara R, Grana C (2006) MOM: multimedia ontology manager. A framework for automatic annotation and semantic retrieval of video sequences. In: ACM Multimedia 2006, Santa Barbara, Oct 23–27, 2006, pp 787–788
8. Bizer C, Heath T, Berners-Lee T (2009) Linked Data—the story so far. Int J Semant Web Inform Syst 5(3):1–22. doi:10.4018/jswis.2009081901
9. Bohlken W, Neumann B, Hotz L, Koopmann P (2011) Ontology-based realtime activity monitoring using beam search. Lect Notes Comput Sci 6962:112–121. doi:10.1007/978-3-642-23968-7_12
10. Carrer M, Ligresti L, Ahanger G, Little TDC (1998) An annotation engine for supporting video database population. Springer Int Series Eng Comput Sci 431:161–184. doi:10.1007/978-0-585-28767-6_7
11. Choudhury S, Breslin JG (2010) Enriching videos with light semantics. In: Fourth international conference on advances in semantic processing, Florence, Oct 25–30, 2010, pp 126–131
12. Duong TH, Nguyen NT, Truong HB, Nguyen VH (2015) A collaborative algorithm for semantic video annotation using a consensus-based social network analysis. Expert Syst Appl 42(1):246–258. doi:10.1016/j.eswa.2014.07.046
13. Elleuch N, Zarka M, Ammar AB, Alimi AM (2011) A fuzzy ontology-based framework for reasoning in visual video content analysis and indexing. In: Eleventh international workshop on multimedia data mining, San Diego, Aug 21–24, 2011, Article 1. doi:10.1145/2237827.2237828
14. Gómez-Romero J, Patricio MA, García J, Molina JM (2010) Ontology-based context representation and reasoning for object tracking and scene interpretation in video. Expert Syst Appl 38:7494–7510. doi:10.1016/j.eswa.2010.12.118
15. Grassi M, Morbidoni C, Nucci M (2012) A collaborative video annotation system based on semantic web technologies. Cogn Comput 4(4):497–514. doi:10.1007/s12559-012-9172-1
16. Guo K, Zhang S (2013) A semantic medical multimedia retrieval approach using ontology information hiding. Computational and Mathematical Methods in Medicine, Volume 2013, Article ID 407917, Hindawi Publishing Corporation. doi:10.1155/2013/407917
17. Haslhofer B, Jochum W, King R, Sadilek C, Schellner K (2009) The LEMO annotation framework: weaving multimedia annotations with the web. Int J Digit Libr 10(1):15–32. doi:10.1007/s00799-009-0050-8
18. Haslhofer B, Momeni E, Gay M, Simon R (2010) Augmenting Europeana content with Linked Data resources. In: 6th international conference on semantic systems, Graz, Sep 1–3, 2010, Article 40. doi:10.1145/1839707.1839757
19. Heggland J (2002) Ontolog: temporal annotation using ad hoc ontologies and application profiles. Lect Notes Comput Sci 2458:118–128. doi:10.1007/3-540-45747-X_9
20. Hunter J, Newmarch J (1999) An indexing, browsing, search and retrieval system for audiovisual libraries. Lect Notes Comput Sci 1696:76–91. doi:10.1007/3-540-48155-9_7
21. Hunter J, Schroeter R, Henderson M (2003) Vannotea screenshot. University of Queensland. http://www.itee.uq.edu.au/eresearch/filething/images/get/projects/vannotea/031014_Screenshot_FilmEd_v2.jpg. Accessed 4 April 2016
22. Jiang Y-G, Bhattacharya S, Chang S-F, Shah M (2013) High-level event recognition in unconstrained videos. Int J Multimed Info Retr 2:73–101. doi:10.1007/s13735-012-0024-2
23. Khedher MI, El Yacoubi MA (2015) Local sparse representation based interest point matching for person re-identification. Lect Notes Comput Sci 9491:241–250. doi:10.1007/978-3-319-26555-1_28
24. Krötzsch M, Simančík F, Horrocks I (2013) A description logic primer. arXiv:1201.4089v3
25. Lee M-H, Rho S, Choi E-I (2014) Ontology-based user query interpretation for semantic multimedia contents retrieval. Multimed Tools Appl 73(2):901–915. doi:10.1007/s11042-013-1383-2

26. Lienhart R, Maydt J (2002) An extended set of Haar-like features for rapid object detection. In: 2002 International conference on image processing, New York, Sep 22–25, 2002, pp 900–903. doi:10.1109/ICIP.2002.1038171
27. Lombardo V, Pizzo A (2014) Ontology–based visualization of characters' intentions. Lect Notes Comput Sci 8832:176–187. doi:10.1007/978-3-319-12337-0_18
28. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2):91–110. doi:10.1023/B:VISI.0000029664.99615.94
29. Mazloom M, Habibian A, Snoek CG (2013) Querying for video events by semantic signatures from few examples. In: 21st ACM international conference on multimedia, Barcelona, Oct 21–25, 2013, pp 609–612. doi:10.1145/2502081.2502160
30. Merler M, Huang B, Xie L, Hua G, Natsev A (2012) Semantic model vectors for complex video event recognition. IEEE Trans Multimed 14(1):88–101. doi:10.1109/TMM.2011.2168948
31. Naphade M, Smith JR, Tesic J, Chang S-F, Hsu W, Kennedy L, Hauptmann A, Curtis J (2006) Large-scale concept ontology for multimedia. IEEE Multimedia 13(3):86–91. doi:10.1109/MMUL.2006.63
32. Nixon L, Bauer M, Bara C, Kurz T, Pereira J (2012) ConnectME: semantic tools for enriching online video with web content. In: 8th international conference on semantic systems, Graz, Sep 5–7, 2012, pp 55–62
33. Oomoto E, Tanaka K (1993) OVID: design and implementation of a video-object database system. IEEE T Knowl Data En 5(4):629–643. doi:10.1109/69.234775
34. Poppe C, Martens G, De Potter P, Van de Walle R (2012) Semantic web technologies for video surveillance metadata. Multimed Tools Appl 56(3):439–467. doi:10.1007/s11042-010-0600-5
35. Rublee E, Rabaud V, Konolige K, Bradski G (2011) ORB: an efficient alternative to SIFT or SURF. In: 2011 I.E. international conference on computer vision, Barcelona, Nov 6–13, 2011, pp 2564–2571. doi:10.1109/ICCV.2011.6126544
36. Sikos LF (2015) Mastering structured data on the Semantic Web: from HTML5 Microdata to Linked Open Data. Apress Media, New York. doi:10.1007/978-1-4842-1049-9
37. Sikos LF (2016) A novel approach to multimedia ontology engineering for automated reasoning over audiovisual LOD datasets. Lect Notes Comput Sci 9621:3–12. doi:10.1007/978-3-662-49381-6_1
38. Sikos LF, Powers DMW (2015) Knowledge-driven video information retrieval with LOD: from semi-structured to structured video metadata. In: Exploiting semantic annotations in information retrieval, Melbourne, Oct 23, 2015, pp 35–37. doi:10.1145/2810133.2810141
39. Simon R, Jung J, Haslhofer B (2011) The YUMA media annotation framework. Lect Notes Comput Sci 6966:434–437. doi:10.1007/978-3-642-24469-8_43
40. Steiner T, Hausenblas M (2010) SemWebVid—making video a first class semantic web citizen and a first class web Bourgeois. In: Ninth international semantic web conference, Shanghai, Nov 7–11, 2010
41. Viola P, Jones M (2001) Rapid object detection using a boosted cascade of simple features. In: IEEE computer society conference on computer vision and pattern recognition, Kauai, Dec 8–14, 2001, pp 511–518. doi:10.1109/CVPR.2001.990517
42. Weiss W, Bürger T, Villa R, Punitha P, Halb W (2009) Statement-based semantic annotation of media resources. Int J Digital Libr 5887:52–64. doi:10.1007/978-3-642-10543-2_7
43. Xu F, Zhang Y-J (2006) Evaluation and comparison of texture descriptors proposed in MPEG-7. J Vis Commun Image Represent 17:701–716. doi:10.1016/j.jvcir.2005.10.002
44. Yang N-C, Chang W-H, Kuo C-M, Li T-H (2008) A fast MPEG-7 dominant color extraction with new similarity measure for image retrieval. J Vis Commun Image Represent 19:92–105. doi:10.1016/j.jvcir.2007.05.003
45. Yıldırım Y, Yazıcı A, Yılmaz T (2013) Automatic semantic content extraction in videos using a fuzzy ontology and rule-based model. IEEE T Knowl Data En 25(1):47–61. doi:10.1109/TKDE.2011.189
46. Zarka M, Ammar AB, Alimi AM (2015) Fuzzy reasoning framework to improve semantic video interpretation. Multimed Tools Appl. doi:10.1007/s11042-015-2537-1

Leslie F. Sikos, Ph.D., is a researcher at Flinders University, Australia, specializing in semantic video annotations, multimedia ontology engineering, and software tools leveraging Linked Open Data. He is the author of 15 textbooks covering a wide range of topics from Internet technologies to video authoring. Dr. Sikos works on the standardization of multimedia reasoning and video concept mapping to Linked Data, advancing the traditional video annotation techniques. Inspired by the creation and exploitation of rich LOD datasets, he actively contributes to the development of open standards and the Open Data initiative. For more information, visit http://www.lesliesikos.com.