Integrating structure and semantics into audio-visual documents

Raphaël Troncy

Integrating Structure and Semantics into Audio-visual Documents

Raphaël Troncy[1,2]

[1] Institut National de l'Audiovisuel, Direction de la Recherche, Équipe DCA
4, Av. de l'Europe - 94366 Bry-sur-Marne
rtroncy@ina.fr
http://www.ina.fr/
[2] INRIA Rhône-Alpes, Équipe EXMO
http://www.inrialpes.fr/exmo

Abstract. Describing audio-visual documents amounts to consider documentary aspects (the structure) as well as conceptual aspects (the content). In this paper, we propose an architecture which describes formally the content of the videos and which constrains the structure of their descriptions. This work is based on languages and technologies underlying the Semantic Web and in particular ontologies. Therefore, we propose to combine emerging Web standards, namely MPEG-7/XML Schema for the structural part and OWL/RDF for the knowledge part of the description. Finally, our work offers reasoning support on both aspects when querying a database of videos.

1 Introduction

The French institute INA[3] has to manage large multimedia and audio-visual (AV) databases. In order to allow an efficient access to the data stored, we describe and index most of the parts of these video documents by their content. This description process can be briefly summarized in three steps:

1. Identification of the content creator (author, title, producer, length, rights, etc.) and the content provider (broadcast channel, broadcast hour, title of the program group, etc.).
2. Structural decomposition in video segments (i.e. a spatial and/or temporal unit of multimedia) corresponding to the logical structure of the TV program.
3. Semantic description of these segments.

Putting this database on the Web raises typical Semantic Web [4] issues because the Web contains more and more multimedia data. For instance, allowing any individuals (i.e. non audio-visual professional) to retrieve and to access

[3] INA (*Institut National de l'Audiovisuel*) has been archiving TV documents for 50 years and radio documents for 65 years. It stores more than 840 000 hours of broadcast programs (including 76 000 hours digitalized) and 940 000 hours of radio programs.

particular audio-visual sequences requires a precise and controlled description of the content of the documents. Moreover, browsing the database to retrieve the broadcast context of an AV sequence requires a controlled description of the logical structure of these documents. For that purpose, we propose in this article an architecture based on ontologies to describe formally the content of the videos, and documentary tools (XML languages family) to constrain their structure. This architecture combines Web emerging standards, namely MPEG-7/XML Schema for the structural part and OWL/RDF for the knowledge part of the description, and finally offers reasoning support on both aspects when querying the database.

This paper is organized as follows. In the next section, we give a complete example on how both structure and content have to be related. In section 3, we briefly introduce some basic components of MPEG-7 [8], that is, the Multimedia Description Schemes (MDS) and XML Schema which has been adopted as the schema language for the multimedia standard. Readers who are already familiar with these languages can skip this section. In section 4 we discuss which language is suitable for describing both the structure and the semantics of audio-visual materials, and for performing some reasoning on these descriptions. We introduce in section 5 our architecture proposal: the modeling of an AV ontology which aims to express the logical constrains for the structure of the videos, and the modeling of another domain-specific ontology to describe the content of each sequence (detailed in section 6). We show how the temporal decomposition of a video and the descriptions of each segment provide the statements that feed a knowledge base (section 7). Finally, we give our conclusions and outline future work in section 8.

2 Example and Requirements

For describing audio-visual materials, librarians currently watch the video, provide its metadata, decompose it into segments, and describe, in natural language, the meaning of each sequence. Moreover, a few keywords controlled by a thesaurus can be used as descriptors for the whole document. For instance, the example given in Table 1 contains a sample description of a French weekly sports magazine named "*Stade2*". To describe this magazine, the INA's documentalists identify its *report* and its *studio* sequences, and write short summaries including information such as events, participants, places, context, etc.

Now, let us imagine that someone wish to retrieve *all audio-visual sequences concerning the Paris-Nice cycling race in which Sandy Casar is interviewed.* A full text search on the description given in Table 1, with the keywords *Paris-Nice*, *Sandy Casar* and *interview* will effectively give the three sequences, but several limitations occur:

1. The sequence numbered 13 does not match precisely the query since it contains also some shots from the studio dealing with other sports and events. The problem raised here comes from the logical model that does not allow more than two levels (the whole document and its sequences).

11 [**Outdoor set with guests** : Laurent Jalabert and Sandy Casar] at 18:37:56:00 - 00:02:43:00. - France 2 *Live from Nice, Jean René GODART interviews Laurent JALABERT and Sandy CASAR about the last stage of the Paris-Nice cycling race.*
12 [**Cycling** : Paris-Nice, the French who success] at 18:40:39:00 - 00:03:17:00 Rodolphe Gaudin. - France 2 *Report devoted to the Paris-Nice cycling race. For the French teams and for the riders, this competition is very important for preparing the Tour de France event. Commentaries on race pictures with some interviews with Jean Marie LEBLANC, Didier ROUS and Sandy CASAR.*
13 [**Indoor Set** : 6^{th} part] at 18:43:56:00 - 00:09:06:00. - Eurosport *In studio, the second part of the interview, from Nice, of Sandy CASAR by Jean René GODART about the Paris-Nice cycling race and a few sports news with pictures commented by Alexandre BOYON and Laurent PUYAT.*

Table 1. A sample description of the French sports magazine "*Stade2*"

2. The sequences numbered 11 and 13 are in fact the same interview, broadcasted in two parts, but this piece of knowledge is lost. The problem comes from the structural model since a sequence cannot be a non-connected temporal entity.
3. Finally, the query cannot be extended. For instance, if the query had been to retrieve *all the audio-visual sequences where a rider gives an interview about any cycling race with several stages* then the full text search engine would answer "no match".

One way to alleviate this problem is to make the AV descriptions more readily accessible to both human readers and automated processes. The Semantic Web aims at providing machine-processable data that improve the automatic manipulation of the information, in particular, offering some reasoning support when querying these data. Thus, we propose to use the languages and the technologies underlying the Semantic Web to improve the current way of describing. As seen with the example given above, our architecture has to fulfill the following requirements:

- To express models that constrain the logical structure of a description (e.g., identify an *interview* inside a particular *report* of a given *sports magazine*).
- To represent the meaning contained in this structure in a machine-accessible fashion to make inferences (e.g., deduce that a *Cartoon* is a *Fiction* with no real characters).
- To describe semantically the content of each sequence, again in a formal way, to perform some reasoning (e.g., deduce that in case of a cycling race with several stages the *Prologue* is always an *individual time trial* numbered *stage 0*).

The question arising is then: which languages are the most suitable to perform all these tasks? We will introduce briefly in the next section the natural candidate

from an audio-visual standpoint, MPEG-7, which aims at describing the content of multimedia documents. But if this language is well suited for document manipulation, it falls short at embedding semantic constructs.

3 MPEG-7: The Standard for Describing Multimedia Documents

According to [3], the most important barrier for accessing efficiently multimedia data has been the lack of a standard, comprehensive and flexible representation that enables intelligent and interoperable multimedia applications. In 1996, the MPEG committee at the Tampere meeting has emphasized the need for a powerful solution for quick and efficient identification (searching, filtering, etc.) of these information. A new standard for "*Multimedia Content Description*", widely known as MPEG-7 [8], emerged from the result of their discussion.

MPEG-7 standardizes *tools* or ways to define multimedia *Descriptors* (Ds), *Description Schemes* (DSs) and the relationships between them. The descriptors correspond to the data features themselves, generally low-level features such as visual (e.g. texture, camera motion) or audio (e.g. melody), while the description schemes refer to more abstract description entities. These tools as well as their relationships are represented using the *Description Definition Language* (DDL), the core part of the language. According to the MPEG-7 Requirements Document, the DDL allows the modification of existing Description Schemes and the creation of new ones. Finally, the W3C XML Schema recommendation [16] has been adopted as the most appropriate schema for the MPEG-7 DDL[4] [15].

MPEG-7 provides the tools for describing any multimedia document, but the standard puts largely emphasis on audio-visual data. The Part 5 of the standard, named *Multimedia Description Schemes*, covers a broad range of descriptors including tools for content description (structure and semantics), but also for content management (media, creation and production, access rights), navigation and access (browsing, summarization), content organization (collection) or even user interaction with multimedia content (preferences, usage history).

3.1 Structure and Semantics in MPEG-7

Intuitively, the content of video data is adequately described if both their *structural* and *semantic* aspects are highlighted. For this reason, the integration of both structure and semantic features is generally considered as the most important contribution of the MPEG-7 standard. More precisely, the structural description of the content is based on the idea of *segment*. The abstract segment type specified in the *SegmentDS* is specialized into concrete types depending on the media (audio, image, video, multimedia). These types use either media locator or masks (for non-connected segments in space and time) to describe

[4] Several extensions (array and matrix datatypes) have been added in order to satisfy specific MPEG-7 requirements.

the formation and boundaries of the segments. Furthermore, the segments can be recursively decomposed temporally, spatially or by their constituents like the audio tracks and thus, form a tree hierarchy of segments, that is the hierarchical organization of the video.

The semantic description (*SemanticDS*) deals with the narrative world depicted or related to the audio-visual content. The MPEG-7 approach is focused on the events, understood as occasions upon which something happens. Objects, people and places can populate such occasions, as well as their date. Furthermore, all these entities can have properties. Finally, the background, the other events and entities provide the context for the description [3]. Therefore, the components of the semantic descriptions roughly fall into entities of the narrative world (*ObjectDS*, *EventDS*, *SemanticPlaceDS*, *SemanticDateDS*), their attributes (label, comments, abstraction level) and their relations. Normative semantic relations describe how semantic entities relate in a story (*agent*, *cause*, *goal*), how they relate to each other (*specializes*, *exemplifies*) and how they are linked to the media (*depicts*, *symbolizes*).

3.2 MPEG-7 Extension Mechanisms

We have seen that the MPEG-7 framework consists in a huge amount of descriptors and descriptions schemes. Moreover, the standard provides two ways to extend them. The first is to use the extension mechanism provided by XML Schema which has been chosen as the MPEG-7 DDL. The XML Schema language provides powerful and flexible means for specifying constrains on XML documents. Basically, XML Schema allows to define simple and complex types, to declare the elements and their attributes with some cardinality indicators that will be used in an XML instance document, and to import or reuse schema components previously defined in external documents. The derivation mechanism makes easy to construct type definitions by specifying only the method of derivation and the differences between the base and derived type definitions. With this simple inheritance mechanism, new user-defined data types can be created, extending or restricting the built-in data types. The standardized Description Schemes of the MPEG-7 language are mostly defined with this facility, and should be themselves, in theory, expandable. However, the conformity and the validity of the new "derived" Description Schemes is rather fuzzy in the standard and should be studied more precisely in future version.

The second extension facility of MPEG-7 deals with the definition of simple taxonomy named *Classification Schemes* (CS). According to the standard, "*a CS is a set of standard terms that describe some domain and may organize the terms it contains with a set of terms relations*". However, a CS is closer to a thesaurus than a formal ontology because the constructs provided are very limited[5] with respect to the current proposed ontology languages. Furthermore,

[5] Only five relationships can be used to form the classification hierarchy: *use*, *used for*, *broader term*, *narrower term* and *related term*.

if CS can be used inside a description, they cannot be used in models to constrain the semantics of collection of multimedia documents.

Despite these two extension mechanisms, it is still impossible to define precisely the semantics of the descriptors added. For instance, one may want to define with necessary and sufficient conditions a `StudioProgram` *as exactly* a `HomogeneousProgram` whose all sequences are `StudioSequence`. Let us consider the `Program#123`, described with the theme `Cycling` and whose sequences are all identified with the genre `StudioSequence`. If a user looks for all `StudioProgram` with the theme `Sports`, he expects to retrieve the `Program#123`. The problem is that this simple result is impossible to infer with MPEG-7 or XML Schema.

To conclude, we notice that the descriptors standardized by MPEG-7 or obtained by extension are not formally defined, but formal semantics is necessary in order to enable the access and the exchange of the multimedia content. XML Schema allows to add some structure, but it cannot express the meaning of this structure. Which language is then suitable for representing the video content both in a machine and a human-understandable format and for performing reasoning on these descriptions?

4 Which Language for Reasoning on Audio-visual Descriptions?

The MPEG-7 descriptors are closely related to the physical features of audio-visual data. For instance, it is not possible to constrain, at the schema level, the segments according to their genre (e.g. report, studio, interview) or their general themes (e.g. sports, sciences, politics, economy). Similarly, the standardized descriptors are too restrictive to cover all the possible descriptions of the content and specially to describe precisely a particular scene of the video. In this section, we will discuss how we can create new descriptors, with a formal meaning, in order to perform some reasoning: with the extension mechanism included in MPEG-7 (section 4.1) or with a knowledge representation language like OWL+RDF (section 4.2). This will lead us to conclude that the best solution is to combine these two approaches.

4.1 MPEG-7 + XML Schema

As we have seen in section 3.2, the MPEG-7 language can be extended in two ways: using the *DDL* (that is, XML Schema) or using the *Classification Schemes* (CS).

XML Schema allows to construct new type definitions extending or restricting already defined data types. For instance, TV Anytime[6] uses this language for

[6] The TV Anytime Forum (`http://www.tv-anytime.org/`) is an association of organizations which seeks to develop specifications to provide value-added interactive services, such as the electronic program guide, in the context of TV digital broad-

building higher-level descriptors, such as the intended audience of a program or its broadcast conditions [12]. We can also use this mechanism for building our own description schemes. For example, the code below defines a type for `TVNews`, a kind of `TVProgram`, `hosted` by somebody, and which contains an unbounded ordered sequence of `studio` and `report` segments.

```
<xsd:complexType name="TVNewsType">
 <xsd:complexContent>
  <xsd:extension base="TVProgramType">
   <xsd:sequence maxOccurs="unbounded">
    <xsd:element name="Studio" type="StudioType"/>
    <xsd:element name="Report" type="ReportType"/>
   </xsd:sequence>
   <xsd:attribute name="host" type="xsd:string" use="required"/>
  </xsd:extension>
 </xsd:complexContent>
</xsd:complexType>
```

However, the problem is not solved because it is still impossible to reason, here, on the structural part of the audio-visual description. Actually, XML Schema provides the means to have very rich structure but is rather limited when expressing the meaning of this structure as the language is (only) concerned with providing typing and structuring information for isolated chunk of data. As we have seen, it is not possible to define (with necessary and sufficient conditions) a `StudioProgram` *as exactly* a `HomogeneousProgram` whose all sequences are `StudioSequence`. It is also not possible to express basic axioms such as "a `TVProgram` cannot be both a `HomogeneousProgram` and a `HeterogeneousProgram`[7]."

The other way to extend the set of MPEG-7 descriptors is the Classification Schemes. TV Anytime has defined in this way a thesaurus describing the possible genres and themes of a TV program. The COALA (Content-Oriented Audio-visual Library Access) project uses also the CS to build a framework for the retrieval of audio-visual TV news documents [6]. But as we have seen in section 3.2, the CS can only be referred to inside a description and they cannot be used to constrain the structure of a collection of documents. More important, the constructs provided by the CS are thesauric relations and cannot express simple axioms allowing efficient inference calculus. Consequently, using only documentary languages like MPEG-7 and XML Schema to describe formally the structure and the content of audio-visual documents seems not enough. However, the knowledge representation languages under development by the Semantic Web Activity of the W3C, are good candidates for solving our problem.

---

casting. The forum identified the metadata as one of the key technologies enabling their vision and have adopted MPEG-7 as the description language.

[7] A definition of `HomogeneousProgram` and `HeterogeneousProgram` is given in section 6.1.

4.2 OWL + RDF

For alleviating the lack of semantics of MPEG-7, Jane Hunter has already proposed an ontology expressing the semantics of the MPEG-7 metadata terms [7]. This ontology, represented in DAML+OIL[8], is built by reverse-engineering of the existing XML Schema definitions together with the interpretation of the english-text semantic descriptions. It contains the class and property hierarchies corresponding to the segment entities classified by media, their possible decomposition, the non-multimedia entities (e.g. Agent, Role, Place, Instrument), etc.

In the same way, we can define the classes and properties still missing in MPEG-7. OWL (*Ontology Web Language*) [10], a language under the W3C recommendation process, can be used to explicitly represent the meaning of terms in vocabularies and their relationships. It is therefore a language for representing, encoding and sharing ontologies on the Web. The expressive power of OWL allows us to define (in the Description Logics sense) classes. For instance, the code below illustrates the definition of the `StudioProgram` class in OWL.

```
<owl:Class rdf:ID="StudioProgram">
 <owl:intersectionOf rdf:parseType="Collection">
  <owl:Class rdf:about="#HomogeneousProgram"/>
  <owl:Restriction>
   <owl:onProperty rdf:resource="#hasSequence"/>
   <owl:allValuesFrom rdf:resource="#StudioSequence"/>
  </owl:Restriction>
 </owl:intersectionOf>
</owl:Class>
```

Furthermore, it is possible to express basic axioms such as "a `TVProgram` cannot be both a `HomogeneousProgram` and a `HeterogeneousProgram`" (with the construct `owl:disjointWith`). OWL ontologies are encoded and written as RDF graphs, which are in turn sets of RDF triples [13]. These triples are the statements feeding a knowledge base on which it is now possible to perform some reasoning.

So, the use of a knowledge representation language seems to satisfy our needs from the knowledge standpoint. However, the structure of the audio-visual documents is completely lost. Now, it is impossible to retrieve the sequence preceding, following, embedding or embracing a given sequence, that is, the necessary context to interpret any video sequence. The problem, as shown in [11], comes here from the modeling foundations of the languages used: whereas XML (and XML Schema) are based on a tree model where nodes are totally ordered, OWL (and RDF) are based on a directed graph model where edges are unordered. In other words, the notion of regular expression, used to constrain the structure of XML documents, is missing in OWL. Patel-Schneider and Siméon [11] proposed to represent the ordering information between the relationships that are related to

[8] DAML+OIL results from the joint efforts of the DAML group (US) and the OIL proposal (EU), and provides a language for expressing far more sophisticated classifications and properties of resources than RDFS.

a common resource. Thus, the formal definition of their SWOL proposal language allows to capture document order in XML documents, but this is only a partial solution for the documentary problem.

Consequently, we propose to use MPEG-7/XML Schema for expressing the structural meaning of audio-visual documents, while OWL/RDF is used for expressing the ontological meaning of the constructs. A transformation, from the ontological syntax (OWL) into the schema syntax (XML Schema), is performed. The following sections detail our proposed architecture and its resources.

5 Architecture Proposal

We have already expressed:

1. The need to make inferences on both structural and conceptual aspects of the audio-visual descriptions.
2. The need to have a good documentary model, that is, sufficiently expressive to constrain the structure of the documents.

To achieve the first goal, we propose to build two ontologies: an audio-visual ontology modeling the classes and properties useful to describe the audio-visual structure of the documents (e.g. genre, general theme), and a domain-specific ontology modeling the classes and properties useful to describe the content of each sequence. These classes are derived in XML Schema types, which are combined in order to form description schemes. This process provides the material for constraining collection of audio-visual descriptions. An overview of our proposed architecture is shown in Figure 1.

The audio-visual ontology aims at representing the meaning of the structural constructs of the descriptions. For instance, the ontology contains the definition of the `StudioProgram` class seen previously. Parts of this ontology are then translated into XML Schema types. At this stage, these types may include additional knowledge that is not modeled in the audio-visual ontology. For instance, the type `TVNewsType` points out that this kind of program is composed of an unbounded sequence (ordered) of `studio` and `report sequences`. As a result, the structural knowledge is divided between the ontology and the description schemes composed of XML Schema types. That does not really matter since it is still possible to make inferences with the ontology and since the description schemes obtained are in a final state. Finally, the XML Schema types derived from the ontology are linked (when it is possible) with the existing MPEG-7 types. For instance, the `TVNewsType` becomes a subtype of the `VideoSegmentType` defined in MPEG-7.

The temporal decomposition of a particular audio-visual program provides a skeleton of description. It is an XML document that has to be validated with the description scheme on which it relies. Sequences of interest are *time-coded* (i.e. spatio-temporally located), and are described in terms of concepts derived from the audio-visual ontology. So, they provide the instances that are encoded in RDF triples that feed our knowledge base. It is now possible to make inferences

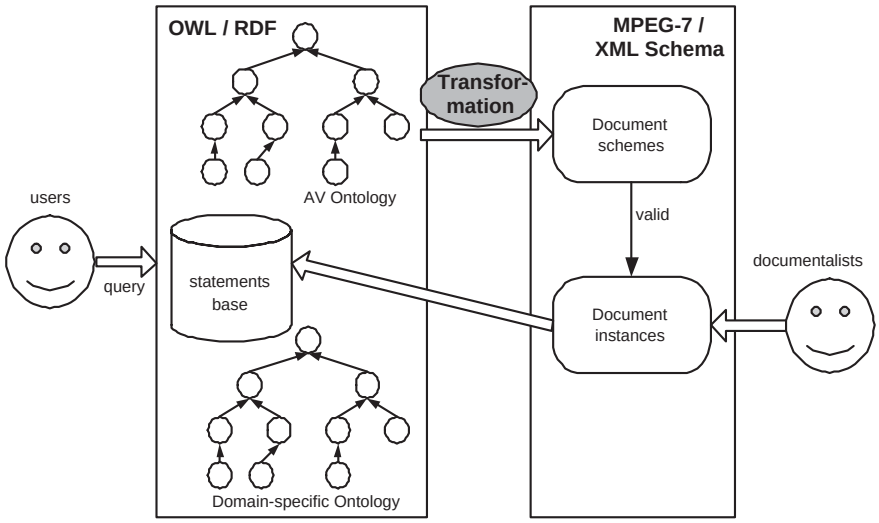

Fig. 1. Overall architecture offering reasoning support on audio-visual descriptions. The transformation between the AV ontology (OWL) and the document schemes (XML Schema) could be carried out with XSLT (see discussion in Section 6.2)

on the structural elements of the description. Moreover, it is still possible to retrieve the context of a given sequence through the documentary structure of the program. Finally, the domain-specific ontology aims at representing the content of each audio-visual sequences. Again, instances of concepts used in the description provide the statements that will be a part of the knowledge base.

6 Ontologies and Documents

The general architecture described above needs some resources. In section 6.1, we present how we have modeled and formalized the audio-visual ontology. We show in section 6.2 how parts of this ontology are translated into XML Schema types that are the building blocks of document schemes. Finally, we briefly present in section 6.3 another ontology dealing with the cycling sport. This domain-specific ontology provides the necessary concepts for describing the content of audio-visual sequences.

6.1 The Audio-visual Ontology

For constraining the structure of the descriptions while performing reasoning on this structure, we have proposed to model and formalize an audio-visual ontology. Many approaches (for a complete survey, the reader can refer to the OntoWeb Technical RoadMap[9]) have been reported to build ontologies, but few

[9] `http://babage.dia.fi.upm.es/ontoweb/wp1/OntoRoadMap/index.html`

fully detail the steps needed to obtain and structure the taxonomies. This observation has lead us to propose a methodology entailing a semantic commitment to normalize the meaning of the concepts. This methodology, detailed in [1], is implemented into the **DOE**[10] (*Differential Ontology Editor*) tool, an ontology editor complementary to other existing tools like **OilEd** or **Protégé2000** (Figure 2).

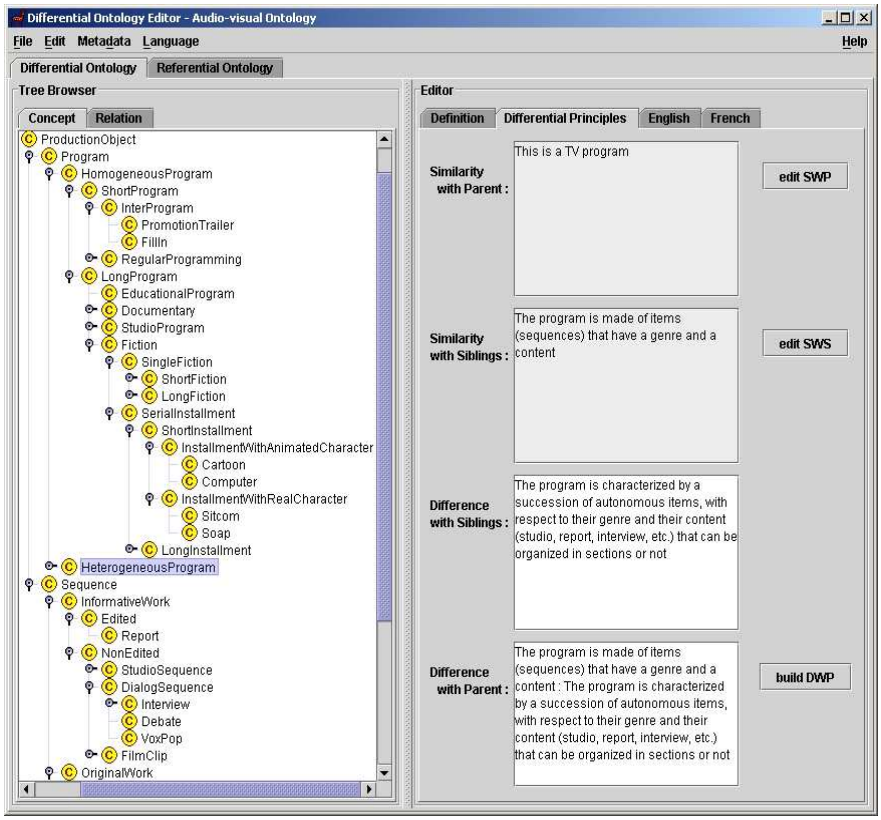

Fig. 2. The differential principles bound to the concept *HeterogenousProgram* in the *DOE* tool

The audio-visual ontology[11] has been elaborated from real guidelines used to teach professional documentalists how to describe the TV and radio programs.

[10] DOE, available for free at `http://opales.ina.fr/public/`, has been partially funded by the OPALES consortium under a PRIAMM grant from the French Ministry of Industry.

[11] This ontology is available in several formats (RDFS, OWL) at `http://opales.ina.fr/public/ontologies/`.

The ontology can roughly be divided into three parts. One contains all the concepts that characterize a program or its sequences. The first criteria used in the ontology modeling process is the audio-visual genre. In the first move, we distinguish the "whole" program from its sequences. A program can be homogeneous or heterogeneous. Actually, a *heterogeneous program* is characterized by a sequence of autonomous elements in form and in content (e.g. *studio*, *report*, *interview*) unlike a *homogeneous program* (Figure 2). Then, the classes *magazine*, *TV news* or *best-of* will be subsumed by the *heterogeneous program* while the classes *documentary*, *fiction* or *studio program* will subsume the *homogeneous program*. The second criteria used to describe an audio-visual sequence is its general themes. We have then classify about 40 themes such as *sports*, *sciences*, *art* or *tourism* under the branch *qualia* of the ontology. Finally, classes dealing with the intended audience, the public participation or the broadcast conditions allow to complete the characterization of the video sequences. The second part of the ontology lists numerous roles of persons involved in the making of the programs while the third part deals with production activities. Here, these classes concern the shooting and recording (e.g. *sound pick-up*, *camera motion*), the editing (e.g. *fade*, *cut*) or the post-production (e.g. *special effects*, *subtitle*).

The audio-visual ontology, once well-conceptualized in DOE, is exported into the RDFS language. We use then the **OilEd** editor [2] to complete the formalization process, that is, to define some classes (see the `StudioProgram` previous example) or to add some basic axioms, and to produce the final OWL files.

6.2 Building Formalized Description Schemes

The audio-visual ontology contains formal definitions for all the concepts used in the structural descriptions of the documents. We have proposed then to translate parts of this ontology into XML Schema types. This translation can be made from the ontology editor or from the formal definition of each element expressed in the ontology web language. For instance, OWL *classes* become *complex types*, the *subClassOf* relationship becomes an *extension* of a base type, the *property restriction* becomes an *element* of a type content model, or the *union* of class expressions becomes a *choice* of this content model, etc. We wonder if this transformation can be made automatically through simple XSLT stylesheet but it seems that this process is not deterministic.

These XML Schema types are linked (when it is possible) with the existing MPEG-7 types. For instance, all the types related to audio-visual genres are considered as a specialization of the MPEG-7 `VideoSegmentType`. Thus, these new types inherit its content model and benefit from the available decomposition for the segments. Finally, types are combined in order to form description schemes[12]. The expressive power of XML Schema allows us to add knowledge

[12] These schemes could also contain some constructs from the VRA Core categories (`http://www.vraweb.org`) in order to describe all the necessary metadata for each sequence, thus enhancing the interoperability between processors aware of at least a part of the description.

which is not modeled in the ontology, such as a regular expression of sequences symbolizing the structural constrains of a collection of programs. The description schemes contain embedded MPEG-7 constructs, thus making easier the exchange of descriptions between MPEG-7 aware applications.

6.3 Domain Ontology and Statements

Besides the audio-visual ontology, we have proposed to model a domain-specific ontology in order to provide the necessary concepts for describing the content itself of audio-visual sequences. For example, we have modeled a cycling ontology [1] which describes the *Tour de France* event or the related sports news and magazines broadcasted on TV.

Furthermore, we use a base of conceptual facts bound to the cycling ontology. These statements were obtained automatically by natural language extraction on textual resources dealing with this event [9]. For instance, the statement: *the rider Sandy Casar has the position 2 in the overall results of the Paris-Nice cycling race* can be represented, in terms of the cycling ontology, as a RDF graph (Figure 3).

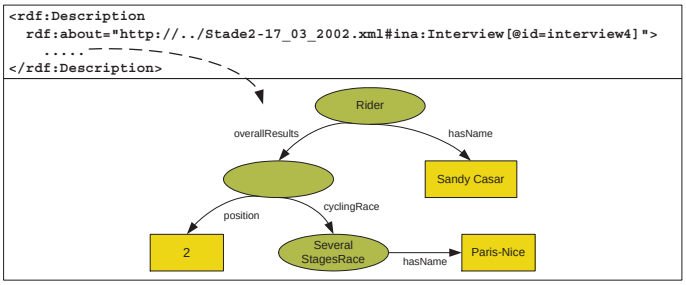

Fig. 3. Graph for RDF/XML example

7 Implementation

The final goal of the architecture proposed in section 5 is the creation of a base of statements. Actually, it provides the instances of the concepts for the two ontologies (audio-visual and cycling) that will be used inside the descriptions of the video documents. Two sources allow us to enlarge the knowledge base. As seen in the previous section, the first source is obtained automatically by conceptual information extraction on texts dealing with cycling. Moreover, the temporal decomposition of a specific program provides the instance of a description scheme and consequently, the instances of the audio-visual ontology.

Actually, the description of an audio-visual document begins with the localization of entities of interest. Each segment is located in space and time in order to be characterized in terms of genre and theme and then, be described by the content. The description has also to be valid with respect to the description

scheme on which the video document belongs. At INA, we use **SegmenTool**[13] to segment temporally the audio-visual documents and to build MPEG-7 descriptions (Figure 4).

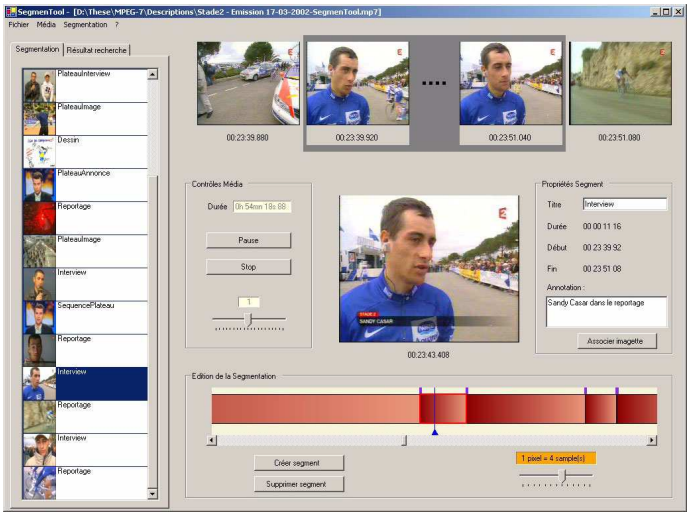

Fig. 4. Video segmentation inside **SegmenTool**

Each segment is time-coded and characterized according to a genre and a general theme. The resulting description reflects then the logical structure of the program. For instance, the code below shows the skeleton of a description where an *interview* of Sandy Casar is located temporally inside a *report* of the sports magazine *Stade2*.

```
<ina:Report id="aa23c647c-6517-4aee-8bce-870ae52a01af">
 ...
 <mp7:TemporalDecomposition>
  <ina:Interview id="adb23ab65-f8e7-4b2a-8c98-807197da600a">
   <mp7:MediaTime>
    <mp7:MediaTimePoint>T00:24:19:10160640F14112000</MediaTimePoint>
    <MediaDuration>PT00H00M07S9031680N14112000F</MediaDuration>
   </mp7:MediaTime>
   <ina:Themes value="Cycling"/>
  </ina:Interview>
 </mp7:TemporalDecomposition>
</ina:Report>
```

This fragment of description contains some descriptors that have their corresponding concept in the audio-visual ontology. Hence, we can generate auto-

[13] SegmenTool is developed by the DCA team of INA and has been funded by the CHAPERON project under a PRIAMM grant from the French Ministry of Industry.

matically the instances of these classes from this fragment. For instance, from the code above, we can generate an instance of the class `Interview` and give a value for its attributes:

```
<Interview rdf:ID="interview4">
 <hasStartTime rdf:datatype="xsd:string">T00:24:19</hasStartTime>
 <hasDuration rdf:datatype="xsd:string">PT00H00M07S9</hasDuration>
 <hasThemes rdf:resource="#Cycling"/>
 ...
</Interview>
```

Finally, other statements describing the content of video sequences (such as the one shown in Figure 3) can be linked – via an XPATH expression – with a particular fragment of the structural description. Hence, Figure 3 asserts that it exists an *interview* in the program where Sandy Casar talks about his position in the overall results of the Paris-Nice cycling race. All these statements can be represented by RDF triples that feed the knowledge base on which we will make inferences.

We use the **Sesame** architecture [5] for storing and querying the ontologies and the statements. Currently, this architecture supports only RDF Schema as the ontology language, but support for OWL Lite semantics is planned for the next releases. Enhanced inference services are possible with the **BOR** reasoner [14] that has been integrated with Sesame within the On-To-Knowledge project. Actually, the BOR reasoner implements the semantics of DAML+OIL. Thus, it is closed to what could be achieved with OWL inference engines in terms of reasoning performance. With the Sesame-integrated-BOR system, answers for the query given at the end of section 3.2 can be retrieved. The system can infer that the `Program#123` is a `StudioProgram` (thanks to the ontology definition) and deals with `Sports` (as its theme is `Cycling`). In the same way, for the query given in section 2 – *find all the audio-visual sequences where a rider gives an interview about any cycling race with several stages* – the system can infer that the sequence *interview*, as described in the Figure 3, matches the query.

8 Conclusion and Future Work

In this paper we have proposed a general architecture for reasoning on descriptions of video documents. We have modeled an audio-visual ontology that contains the concepts of *genre*, of *theme* and of *technical process* for the production of TV programs. We have also translated parts of this ontology into XML Schema types which specialize MPEG-7 descriptors. Those XML Schema types can be combined inside description schemes that constrain the logical structure of collections of documents. Describing a particular TV program boils down to cut it out temporally in order to instantiate a description scheme, that provides the instances of the concepts of the audio-visual ontology which finally, feed the knowledge base.

We have now to test this architecture on a larger scale, that is, a few hours of video annotated and real users querying the database with the ontologies. We

expect also tools that implement the semantics of OWL in order to exploit the full reasoning power of the ontologies modeled. Finally, we plan to compare and merge the MPEG-7 ontology proposed by Hunter with our audio-visual ontology.

References

1. B. Bachimont, A. Isaac, and R. Troncy. Semantic Commitment for Designing Ontologies: A Proposal. In *Proc. of the 13th International Conference on Knowledge Engineering and Knowledge Management (EKAW'02)*, LNAI 2473, p. 114-121, Sigüenza, Spain, 2002.
2. S. Bechhofer, I. Horrocks, C. Goble, and R. Stevens. OilEd: a Reason-able Ontology Editor for the Semantic Web. In *Proc. of KI2001, Joint German/Austrian conference on Artificial Intelligence*, LNAI 2174, p. 396-408, Vienna, Austria, 2001.
3. Ana B. Benitez, H. Rising, and C. Jörgensen. Semantics of Multimedia in MPEG-7. In *IEEE Conference on Image Processing (ICIP'02)*, Rochester, New York, 2002.
4. T. Berners-Lee, J. Hendler, and O. Lassila. The Semantic Web. In *Scientific American*, 284(5):34-43, 2001.
5. J. Broekstra, A. Kampman, and F. van Harmelen. Sesame: a Generic Architecture for Storing and Querying RDF and RDF Schema. In Ian Horrocks and James Hendler, editors, *Proc. of the first International Semantic Web Conference (ISWC'02)*, LNCS 2342, p. 54-68, Sardinia, Italia, 2002.
6. N. Fatemi, and O. Abou Khaled. COALA: Content-Oriented Audiovisual Library Access. In *Proc. of the 8th International Conference on Multimedia Modeling (MMM'2001)*, p. 59-71, Amsterdam, The Netherlands, 2001.
7. J. Hunter. Adding Multimedia to the Semantic Web - Building an MPEG-7 Ontology. In *Proc. of the First International Semantic Web Working Symposium (SWWS'01)*, p. 261-283, Stanford, California, 2001.
8. ISO. Information Technology - Multimedia Content Description Interface (MPEG-7). Standard No. ISO/IEC 15938:2001, International Organization for Standardization(ISO), 2001.
9. E. Le Roux. Extraction d'information dans des textes libres guidée par une ontologie. PhD Thesis, 2003.
10. OWL, Web Ontology Language Reference Version 1.0. W3C Working Draft, 31 March 2003. `http://www.w3.org/TR/owl-ref/`
11. P. F. Patel-Schneider, and J. Siméon. Building the Semantic Web on XML. In Ian Horrocks and James Hendler, editors, *Proc. of the first International Semantic Web Conference (ISWC'02)*, LNCS 2342, p. 147-161, Sardinia, Italia, 2002.
12. S. Pfeiffer, and U. Srinivasan. TV Anytime as an application scenario for MPEG-7. In *Proc. of Workshop on Standards, Interoperability and Practice of the 8th International Conference on Multimedia*, ACM Multimedia, Los Angeles, California, 2000.
13. RDF, Ressource Description Framework Primer. W3C Working Draft, 23 January 2003. `http://www.w3.org/TR/rdf-primer/`
14. K. Simov, and S. Jordanov. BOR: a Pragmatic DAML+OIL Reasoner. Deliverable 40, On-To-Knowledge Project, 2002.
15. E. Terzi, A. Vakali, J. Fan, and M.-S. Hacid. The MPEG-7 Multimedia Content Description Standard and the XML Schema Language. In *Proc. of the 7th International Conference on Distributed Multimedia Systems (DMS'01)*, Taipei, Taiwan, 2001.
16. XML Schema. W3C Recommendation, 2 May 2001. `http://www.w3.org/XML/Schema`