

# A Motivating Scenario for Designing an Extensible Audio-Visual Description Language

Raphaël Troncy, Jean Carrive, Steffen Lalande and Jean-Philippe Poli  
Institut National de l'Audiovisuel, Équipe DCA  
4, Av. de l'Europe, 94366 Bry-sur-Marne, France  
{rtroncy, jcarrive, slalande, jppoli}@ina.fr

## Abstract

*Enabling an intelligent access to multimedia data requires a powerful description language. In this paper, we demonstrate, through a particular scenario, why the MPEG-7 standard fails to fulfill this task. We introduce then our proposition: an audio-visual specific description language (AVDL), modular, reduced, but designed to be extensible. This language is centered on the notions of descriptor and structure. A descriptor can be a low-level feature, automatically extracted from the signal, or a higher semantic concept that will be used to annotate the video documents. It can have some properties including a formal definition encoded in the new Semantic Web languages, which provides a formal semantics. Finally, the descriptors can be combined into structures according to defined models that provide description patterns. This notion of structure is close to what is named Description Schemes in the MPEG-7 standard, but with more expressiveness as our proposed language allows to define its own models. We show how our motivated scenario can henceforth be carried out. We give also some implementation details of our language stressing on its XML syntax and the API developed for accessing the data model. We show finally how it will be used into the FERIA project, whose goal is to provide a generic framework for developing multimedia production applications for new delivery channels such as interactive television.*

## 1 Introduction

Producing multimedia content today is easier than ever before. While digital video documents are more and more available on the Web, their effective processing is still problematic. Generally, the audio-visual document has to be decomposed in smaller parts and then indexed by a representation of its content to be efficiently retrieved and ma-

nipulated. For instance, INA<sup>1</sup> is used to describing manually both the structure and the content of each document of its archive with keywords, thesauri or free text annotations. These textual descriptions can be used for retrieving relevant video sequences or for enhancing their content to produce new rich media documents such as adaptive and interactive presentations. The main problem is then the design and the standardization of a *description language*, flexible enough to enable an intelligent access to the information, but which takes also into account the peculiarities of the audio-visual media. This article proposes to tackle this hard task arguing that the new MPEG-7 standard [8] fails in this challenge.

The “*Multimedia Content Description*” standard, widely known as MPEG-7, standardizes *tools* or ways to define multimedia *Descriptors (D)*, *Description Schemes (DS)* and the relationships between them. The descriptors correspond to the data features themselves, generally low-level features, 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. After an exploration phase where high-level requirements have been gathered [11], a call for DDL proposals was proposed by the MPEG-7 committee. These proposals were evaluated [7] and at the end of the selection process, the XML Schema recommendation [17] has been adopted as the most appropriate schema language. However, we argue that MPEG-7 does not comply with the requirements the committee has itself settled, mainly because of the choice of XML Schema as the MPEG-7 DDL. Actually, the W3C schema language does not reconcile the critical issue of object-oriented semantic expression versus structural validation. Therefore, we propose in this article an alternative description language, audio-visual specific, modular, reduced, but designed to be

---

<sup>1</sup>The *French National Institute of Audio-visual (INA)* has been archiving and indexing the TV and radio programs broadcasted in France for forty years and thus, has to manage huge audio-visual databases.

extensible.

This paper is organized as follows. In the next section, we present a scenario that motivates the design of a new audio-visual description language. We demonstrate in section 3 why MPEG-7 falls short to be the multimedia description language it aims to be since it does not allow to carry out this scenario. We introduce then our proposition in section 4: an audio-visual specific description language centered on the notions of *descriptor* and *structure*. A descriptor can be a low-level feature, automatically extracted from the signal, or a higher semantic concept that will be used to annotate the video documents. It can have some properties including a formal definition encoded in the new Semantic Web languages, which provide a formal semantics. Both descriptors and properties are organized into taxonomies, hence the properties can be inherited and the model can be easily mapped into an object-oriented programming language. Finally, the descriptors can be combined into structures according to defined models that provide description patterns. This notion of structure is close to what is named *Description Schemes* in the MPEG-7 standard, but with more expressiveness as our proposed language allows the definition of its own models. Section 5 comes back to our preliminary scenario and shows how it can now be accomplished with our proposed audio-visual description language. Section 6 gives some implementation details stressing on the XML syntax and the API developed for accessing the data model. We show in section 7 how this language will be used into the FERIA project, whose goal is to provide a generic framework for developing multimedia production applications for new delivery channels such as interactive television. Finally, we give our conclusions and outline future work in section 8.

## 2 Motivating Scenario

The institute INA is faced with the problem of describing a huge amount of hours of video (all the TV programs of 50 channels broadcasted in France). Consequently, the documentalists have to strictly follow well-established documentary practices when indexing all these documents [14]. In this context, our problem is to develop a generic application that could assist the manual description task while guaranteeing the structural and semantical validity of the descriptions. The structural constraints allow the validation of a description with respect to some pre-defined patterns representing the logical structure of a program. The semantical constraints force to have machine understandable description of the content thus offering reasoning support on both aspects when querying a database of videos.

The description language underlying to such an application should then fulfill some requirements. It shall support the hierarchical representation of different descriptors in or-

der to express their common properties at the right level and to allow the inheritance of their behaviour along the taxonomy. Furthermore, these descriptors shall be formally defined (e.g. using the *Ontology Web Language* [9]) thus allowing the application to perform some inferences on the descriptions. It shall be able to define description schemes, that is, kind of patterns that constraint the logical structure of a document (in terms of its components, and the spatial, temporal and logical relationships between them). Finally, this language shall be really machine processable and not only an exchange format.

As an example, let us take the task of describing a collection of a particular program. Table 1 gives, in the EBNF form, the mereological structure for all `Sports-Magazine`. This schema specifies that a sports magazine always begins with a `FirstStudioSequence`, followed by several sequences that could be either a `StudioSequence` or a series of `StudioAnnouncement` and `Report`, and that finally ends with a `LastStudioSequence`. A `StudioSequence` can contain some `LiveSportsExcerpt` commented on studio and short `CaricaturePicture`. Finally, a `Report` can be composed of `Interview` and `LiveSportsExcerpt` sequences.

We will see in the next section why MPEG-7, the natural candidate to be this language, is finally not suitable for this scenario.

## 3 MPEG-7: a Non-Suitable Description Language for Accomplishing this Scenario

MPEG-7 [8] is a large and complex standard that provides the tools (i.e. some Descriptors, Description Schemes and the DDL) for describing any multimedia document, but it puts largely the emphasis on audio-visual data. As part of the MPEG-7 development process, the specification of the requirements play a key role since they allow one to define the scope and the expressiveness of the language. The MPEG-7 Requirements document [11] presents the needs that the tools have to fulfill. Among these requirements, those concerning the DDL are the ability to represent hierarchies of descriptors and spatial, temporal, structural and conceptual relationships between the descriptors, to have compositional capabilities (i.e. creation and modification of description schemes combining descriptors), to validate constraints expressed in these description schemes, to link the descriptions with ontologies and with the media, etc.

All the functionalities listed above are clearly relevant and necessary for accessing and processing audio-visual data. However, not all of these requirements have been met in the current MPEG-7 framework. Furthermore, we argue that some of them cannot be addressed, mainly due to the choice of XML Schema as the MPEG-7 DDL. We give be-

```

SportsMagazine := ( FirstStudioSequence,
                    ( StudioSequence |
                      ( StudioAnnouncement, Report) )+,
                    LastStudioSequence )
StudioSequence := (LiveSportsExcerpt | CaricaturePicture)*
Report := (Interview | LiveSportsExcerpt)*

```

**Table 1. The Sports Magazine temporal structure defined in terms of a regular expression of its sequences (EBNF form)**

low some arguments that prevent the use of MPEG-7 as an effective description language.

1. *A closed set of descriptors*: Even if the language consists in a huge amount of descriptors and descriptions schemes, it is not sufficient to cover all description needs. Yet, it is not possible to define new descriptors. MPEG-7 provides a vocabulary extension facility thanks to the *Classification Schemes (CS)* that allow one to define “a set of standard terms that describe some domain and a set of terms relations for organizing them” [8]. However, these terms are not descriptors, that is, they can only be used as descriptor value or property value but cannot be part of models to constrain the structure and the semantics of collection of multimedia documents. Furthermore, a CS is closer to a thesaurus than a formal ontology because the constructs provided are very limited<sup>2</sup> with respect to the current proposed ontology languages. This lightweight semantics does not allow to bring the access and the exchange of multimedia content between applications to its full potential.
2. *A non object-based data model*: Using XML Schema as the MPEG-7 DDL does not provide a *pure* object-oriented data model. XML Schema allows the derivation of new types from existing ones, either by restriction or by extension. However, even if this derivation mechanism reminds us the inheritance of the object-oriented programming languages, it is actually rather a reuse of the content model defining the super type [4]. Consequently, the use of XML Schema makes difficult the creation of new DS while inheriting some properties. Furthermore, the conformity and the validity of possible new “derived” Description Schemes are rather fuzzy in the standard and should be studied more precisely in future version.
3. *A non modular language*: As a direct consequence of the bad modeling outlined above, the MPEG-7 schema is not modular. Hence, any description has to import

<sup>2</sup>Only five relationships can be used to form the classification hierarchy: *use*, *used for*, *broader term*, *narrower term* and *related term*.

the whole schema to be valid, which makes difficult the development of lightweight and portable application dealing with the metadata. The descriptions themselves are very complex because of the way they have to be made. Indeed, for manually describing a particular audio-visual segment of interest, a documentalist has to select the whole document and apply numerous decompositions until he reaches the desired portion of video to be annotated. However, the strictly hierarchical resulting description is not suitable in a lot of cases and other syntactical constructions should be allowed.

4. *No formal semantics provided*: MPEG-7 is based on XML Schema that not only impacts syntax level aspects. Since no formal semantics are provided, one needs to read the English prose in the standard to understand the meaning of the schemata. The applications cannot access to the meaning of the descriptions, which is obviously a major drawback for their interoperability. For alleviating the lack of semantics of MPEG-7, J. Hunter has already proposed an ontology expressing formally the semantics of the MPEG-7 metadata terms [6]. This ontology, built by reverse-engineering of the existing XML Schema definitions together with the interpretation of the English-text semantic descriptions, is represented using Semantic Web languages (OWL/RDF)<sup>3</sup>. However, this ontology covers only the descriptors standardized by MPEG-7, that are mainly related to the physical features of audio-visual data. For instance, it is not possible to type video segments according to their genre (e.g. report, studio, interview) or their general themes (e.g. sports, sciences, politics, economy). A very detailed comparison of MPEG-7 and Semantic Web technology is proposed in [15, 16]. The authors give the pros and the cons of each languages with respect to their abilities to define structures for describing media semantics. Previous work by the authors [13] described a more general architecture based on ontologies to describe formally the content of the videos

<sup>3</sup>This MPEG-7 ontology is available at <http://metadata.net/harmony/MPEG7/mpeg7.owl>.

(OWL/RDF [9, 12]), and documentary tools (MPEG-7/XML Schema) to constrain their structure, to finally offer reasoning support on both aspects when querying the database. However, if this architecture allows to bring more semantics to the description, it does not solve the problems expressed above (real object-oriented data model, modularity, audio-visual specific language, extensibility, etc.).

Going back now on our simple scenario presented in the section 2, we need to create new descriptors for representing the notions of `Interview`, `Report` or `LiveSportsExcerpt`. These descriptors have also to be defined formally in order to be machine understandable. Finally, we need to construct a new description scheme to control the temporal structure of all programs belonging to this Sports Magazine category. For creating new descriptors, we can use the MPEG-7 *Classification Schemes* mechanism, but as a result, we obtain new terms that cannot be used in description schemes (*see argument 1*) and with a rather fuzzy formal semantics (*see argument 4*). Another possibility consists in creating new descriptors by extending existing ones with the XML Schema derivation mechanism. However, due to the limitation of XML Schema, we cannot define properties for some descriptors and inheriting them along a new hierarchy (*see argument 2*). Actually, we face here the critical issue of object-oriented semantic expression versus structural validation: by choosing XML Schema as the DDL, the MPEG-7 committee has clearly opted for the latter.

To summarize this discussion, the MPEG-7 language appears to be messy when one tries to use it. Its goal is unclear and one can wonder if it aims to be an exchange format or a real machine understandable and processable representation of the multimedia description. However, we think that all the requirements stated in prelude to the design of the language are necessary. We propose then to start again from these requirements for designing a core description language, audio-visual specific, reduced but extensible and coming with a formal semantics. We argue that the issue stated above cannot be reconciled with an extension of the current MPEG-7 proposal since it implies a complete change of paradigm on which the DDL is based. We justify hence the design of an alternative language. In the following sections, we detail this language and how it can be used to carry out our proposed scenario.

## 4 An Audio-Visual Description Language Proposal

In this section, we present our proposed audio-visual description language. First, we introduce its basic concepts, that is, the notions of *Description Scheme* and *Description*

(section 4.1). Second, we give the meta-model of the language consisting in a core hierarchies of descriptors and properties, and we show how these taxonomies can be extended (section 4.2). Third, we describe how to combine descriptors into structures according to defined models in order to provide description patterns. Again, it is possible to extend the language with new models and appropriate tools (section 4.3). Finally, we complete the presentation of the language with the necessary constructs for locating descriptors in audio-visual documents when annotating them (section 4.4). In the following, this language will be named AVDL for *Audio-Visual Description Language*.

### 4.1 The Basic Concepts

The notion of *document* is central for any information system. According to [10], a document is a set of traces, written by one or several authors on a medium, meaningful for potential senders and receivers, and that exhibits an intentional structure. When the document is digitalized, its physical nature becomes less important and we can observe that it shares some common peculiarities with the audio-visual document: both are encoded in a particular format that cannot be accessed directly (unlike the traditional textual document), but has to be calculated in order to provide a meaningful representation. As reported by [1, 2], old confusions between medium, message, and meaning are then renewed. Hence, when describing audio-visual documents, the system should distinguish the document from its content or the media itself.

The AVDL includes these three notions:

- A **document** is considered from a classical documentary standpoint and will serve as basic material for the description process.
- A **content** is an abstraction of a media. It can be *physical* (i.e. an autonomous file stored on a disk and unambiguously identified) or *virtual* (i.e. the result of elementary editing operations on physical pieces of content, thus consisting in references to physical contents).
- A **media** is an interface between the document and its content. Its main purpose is to allow the decomposition into its media constituents such as audio tracks or viewpoints from several cameras.

As the MPEG-7 DDL, our proposed language allows one to define *Description Schemes* (DS). However, these are not equivalent to the corresponding MPEG-7 tool. In our case, a DS has a broader sense, since it can be considered as a set of definitions gathering together:

- The **descriptors**: they can be low-level features, automatically extracted from the signal (e.g. Tracking,

Detection, etc.) or higher semantic concepts (e.g. audio-visual genres like *Report*, *Debate*, etc.), and they are used to annotate the audio-visual documents. The descriptors are organized in taxonomies and can be defined formally in order to be machine understandable.

- The **properties**: they help to define the descriptors and are also organized in taxonomies. Attached to a particular descriptor, a property is typed, that is, a range is provided. Moreover, a property can be defined intentionally with a rule allowing to calculate dynamically its value.
- The **structures**: they are the real description patterns. A structure is defined with a model that gives the possible descriptors to be instantiated and the way they are combined. As we will see in section 4.3, some model definitions are already provided with the language, but the user can define their own models as soon as the tool to process them is provided.

A *Description* is then an instance of a *Description Scheme*. An instance of a descriptor is a set of values of the types of the properties defining the class, whereas an instance of a structure is a combination of descriptor instances, controlled by its model.

## 4.2 An Extensible Set of Descriptors and Properties

Extensibility is a key feature of our proposed description language. Therefore, the AVDL allows the definition of its own descriptors and properties. Actually, all the new descriptor and property definitions have to be classified under a hierarchy of built-in concepts that are part of the meta-model of the language. Figure 1 shows this meta-model in the UML formalism.

In the language, these concepts have a well-defined semantics:

- **Descriptor** is the top concept of the hierarchy. It can have a parent descriptor and inherits then some super properties. It can also contain its own properties and be part of structure definitions. Finally, a formal definition can be given in order to represent the meaning of the descriptor in a machine-accessible fashion, which allows to perform some reasoning on the future descriptions. Currently, the Ontology Web Language (OWL) [9] is supported by the AVDL for representing this formal semantics.
- **LocatedDescriptor** is a kind of descriptor which has the ability to be located in an audio-visual content. All the necessary constructs for locating descriptors in space and/or in time are presented in section 4.4.

- **Document** and **Excerpt** are rather documentary units that characterize a whole piece of content.
- **Collection** allows to introduce the notion of collection, that is, a set of *Excerpt* (or *Documents*) that shares some common characteristics such as the authors, the thematic, the characters and the settings.

In the same way, new properties can be defined, specializing the **Property** concept, either inside a descriptor definition or outside. In the latter case, the domain of the property has to be specified. In both cases, the range of the property is mandatory since each property is typed. Finally, a property can specify its super property and then inherits its domain and range. A special kind of property named **IntensionalProperty** allows in addition to dynamically infer its value thanks to the specification of a calculus rule.

## 4.3 An Open Way to Design Its Own Structure

Describing audio-visual documents amounts to consider conceptual aspects (the content) as well as documentary aspects (the structure). In order to constraint the descriptions, it is important to represent and control this structure, that is, a combination of descriptors according to a given model. The AVDL proposes two basic models:

- **Containment** specifies that an audio-visual segment can be decomposed into other segments and recursively.
- **Regular Expression** specifies that an audio-visual segment can be decomposed temporally as a regular expression of other audio-visual segments.

Besides these schemes, the AVDL allows the definition of its own model as soon as the specific tool dealing with this part of the description is specified. The main issue does not come from the representation, but rather from the instrumentation of the model, that is the control of the structure. In particular, we investigate the possibility to define models in terms of temporal constraints such as the ones expressed in [3].

The inheritance of structure is also an open issue even if, most of the time, this notion does not make sense. Actually, in particular cases, it can be necessary to specialize existing structures such as the one described in Table 1. That could be done by using more specific concepts or by restricting their number of occurrence in the model.

## 4.4 Location and Media

As we have seen in Figure 1, some descriptors have the ability to be located in the audio-visual content. For that purpose, the AVDL provides the necessary constructs for

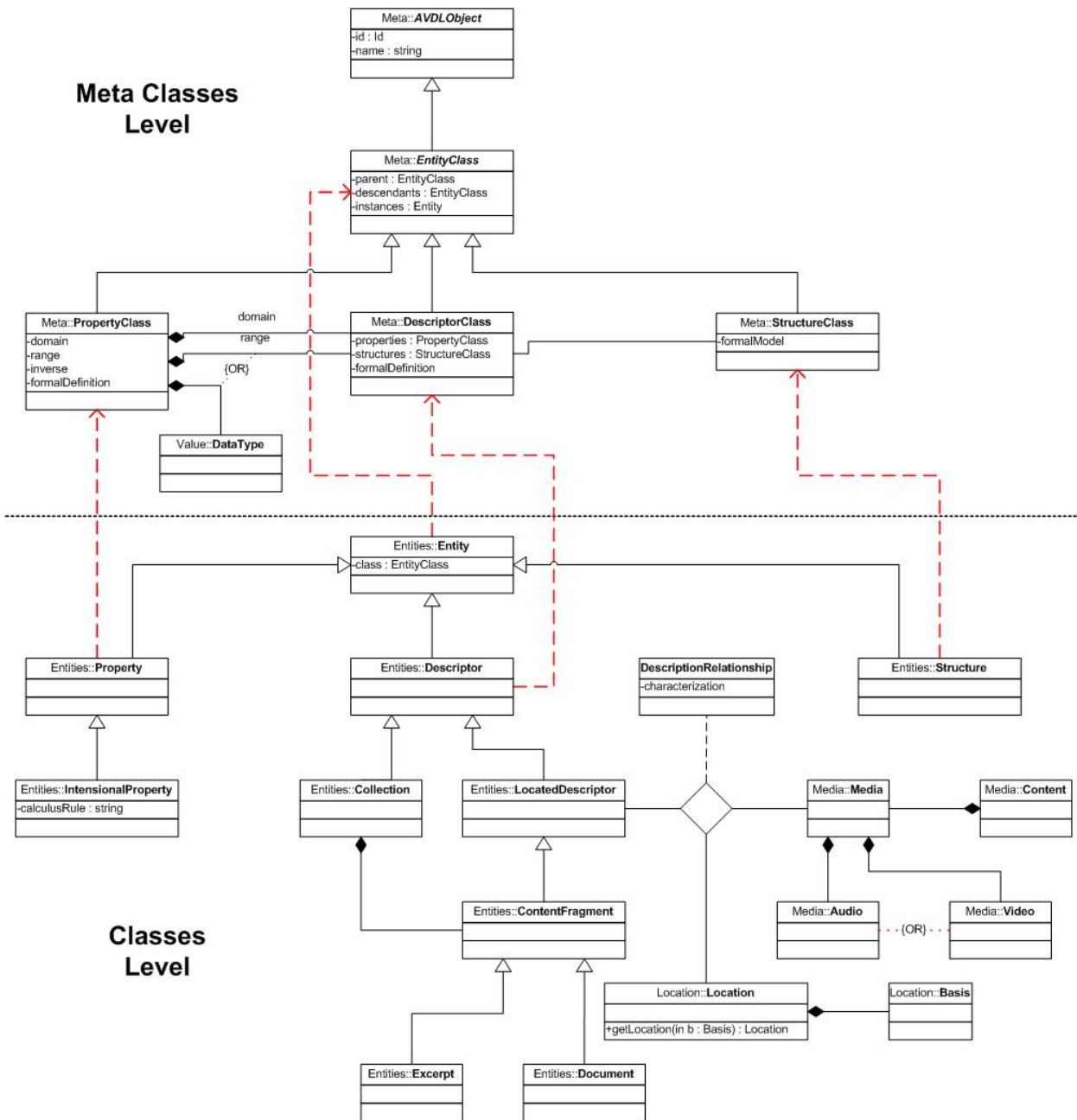


Figure 1. The hierarchy of concepts provided by our proposed AVDL meta-model

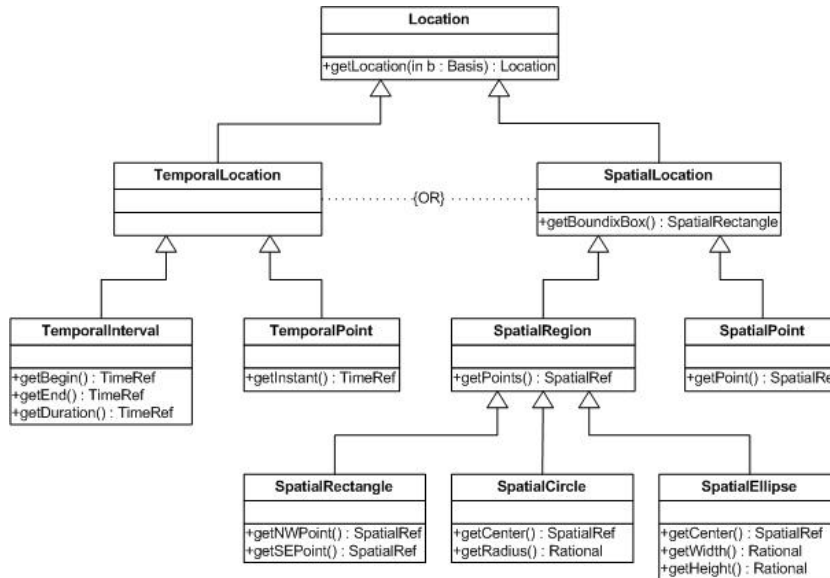


Figure 2. The basic constructs for locating descriptors in a media

representing the spatial, temporal and spatio-temporal location (Figure 2). A temporal location can be a *temporal interval* (defined by its boundaries) or an *instant*. A spatial location can be a *point* in the image, a *region* (defined with  $n$  points), or a *rectangle*, a *circle* or an *ellipse* (defined with two or three points). For each spatial location, a bounding box can be calculated. Finally, each location can be made absolutely (the default case) or relatively to a basis.

The AVDL proposes also a new time measure in order to solve the problems coming from the way the physical contents are encoded. Hence, two special datatypes are provided with the language:

- The **TimeRef** class allows an accurate representation of temporal points and intervals on a timeline. Its internal representation of time uses the least common multiple between the usual sound sample rate (96000 and sub-multiple or 44100 and sub-multiple) and video frame rate (30, 25, 24), that is 14112000. This integer defines then an universal common sample rate (*i.e.* 14112000 corresponds to 1 second) and any temporal point in an audio-visual content will be represented as an integer on this temporal basis. This representation avoids the well-known “precision lost” drawback encountered in real number arithmetic since the temporal bounds of audio-visual samples are unambiguously represented and their retrieval can be achieved accurately as soon as rational calculus are performed.
- The **SpaceRef** class represents a point and it is defined as a couple  $(x,y)$  of rational. Again, the idea behind the use of rational numbers (*i.e.* a couple of integer)

is to avoid the rounded errors when one has to make some calculus on the points of an image (*e.g.* due to a picture scaling or a complete change of resolution).

## 5 Carrying out The Scenario Using the AVDL

As we have seen in this paper, the meta-model of our proposed AVDL allows the definition of new descriptors using the subsumption relationship. For instance, a face detection and tracking application needs the concepts of *Face*, *Eyes*, *Mouth* and *Nose* that are instances of *LocatedDescriptor* as they use either a rectangle (the bounding box of the object) or a point to be located. A *Tracking* is an instance of *LocatedDescriptor* located in a temporal interval. It is also linked to a media, and can then be played on a particular device.

Going back over the particular scenario given in section 2, we can now define the concepts of *StudioSequence*, *Report*, *Interview*, etc. For that purpose, we have modeled an audio-visual ontology<sup>4</sup> defining the classes and properties useful to describe the *genre*, the *theme* or the *technical process* for the production of TV programs [14]. For instance, the knowledge that the *Synopsis* class is equivalent to the *FirstStudioSequence* class, or that a *CaricaturePicture* is exactly a *Sequence* whose author is a *Caricaturist* and whose duration is less than 8 seconds can be easily represented using description

<sup>4</sup>This ontology is available in several formats (RDFS, OWL) at <http://opales.ina.fr/public/ontologies/>

logics such as the OWL language. These concepts and properties are then linked to the meta-model of our proposed audio-visual description language and can be used in the definition of new description schemes that will constraint the logical structure of these documents.

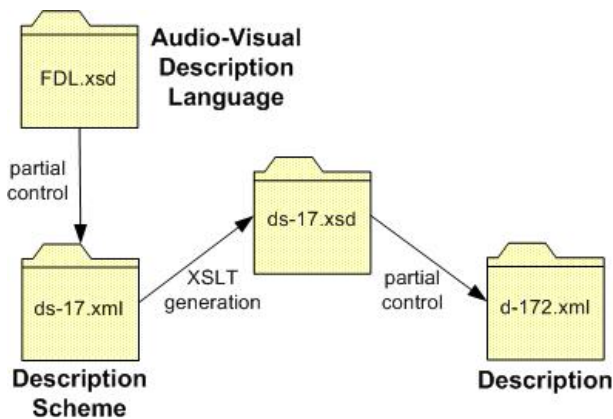
Finally, all these descriptors can be combined to define the formal model of the temporal structure of all programs belonging to the Sports Magazine category. The *Regular Expression* model presented in the section 4.3 is enough to accomplish our scenario, but one can define a more complex model (e.g. with temporal constraints) to strengthen the structural validity of the description of the programs. In the following sections, we give some implementation details of our proposed description language and we show how it will be used into the FERIA project.

## 6 Implementation

The proposed audio-visual description language described above is still under implementation. In the following, we give its current XML syntax (section 6.1) and the API developed to access the data model (section 6.2).

### 6.1 XML Serialization

The AVDL adopts an XML syntax. The Description Schemes and Descriptions are then plain XML files. More precisely, the meta-model expressed in section 4 is an XML Schema file (named AVDL.xsd). Therefore, any DS (i.e. set of descriptor, property and structure definitions) has to be a valid instance with respect to this schema file.



**Figure 3. The XML serialization of our proposed Audio-Visual Description Language**

A partial control on the syntactical aspects of these definitions can be done using an XML Schema compliant

parser. The semantics of the model can be validated using the formal part of each definition. The reasoning task can be accomplished by any OWL-aware inference engine such as RACER [5].

Description schemes are then translated automatically into other XML Schema files, thanks to XSLT transformations. A particular description is thus a valid XML file with respect to this schema. Again, this syntactic sugar allows to benefit from the validation control carrying out by a XML Schema compliant parser, the semantics validation task remaining to do by other specific tools. Figure 3 summarizes this overall process.

### 6.2 Mapping to an Object Programming Language

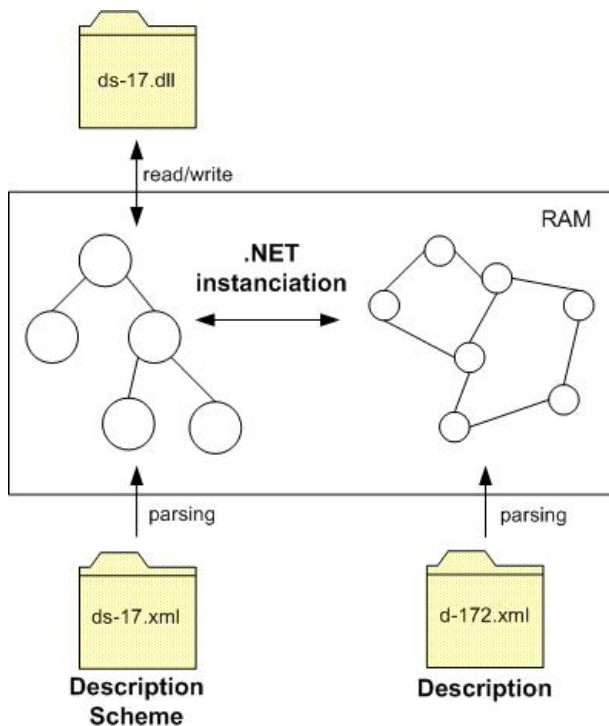
The hierarchy of descriptors forms a classical taxonomy in the sense of the description logics. The formal definition of each descriptors allows to use a classifier such as RACER [5] to perform some reasoning on the descriptions. Moreover, we have chosen to use an object programming language, namely C# in the Microsoft .NET framework, for interfacing the descriptors with visual editors, media players, etc. This language is also a core component of the FERIA project (see section 7). Figure 4 describes how C# is used. The kernel classes of the AVDL are implemented once forever as C# assemblies (or dynamic libraries, also known as DLLs). In order to be able to express new descriptor classes, we used the introspection mechanisms of the .NET framework to dynamically generate C# classes in memory, for instance from a XML representation (as shown in section 6.1). The new created classes can then be saved in an assembly and can be reused in an application as a standard .NET library.

There are two ways to create instances of descriptors. In the first case, the introspection mechanisms of .NET can be used to dynamically create instances from dynamic created classes. In the second case, the generated DLL is used in a .NET Integrated Development Environment (IDE) and the descriptor classes are thus made visible to the programmer.

## 7 Application

The audio-visual description language described above is at the heart of an ongoing project named FERIA (*Framework for the Experimentation and the Development of Industrial multimedia Applications*) whose objective is to provide a generic framework for developing multimedia production applications for new delivery channels such as interactive television. The descriptions of audio-visual documents are the starting point for the class of applications that can be developed within the FERIA framework, as shown in Figure 5. In a first step, raw audio-visual materials are

processed by automatic and manual tools for obtaining descriptions. In a second step, the documents and their descriptions are combined to produce new contents, such as a web site, an interactive television program or a DVD.

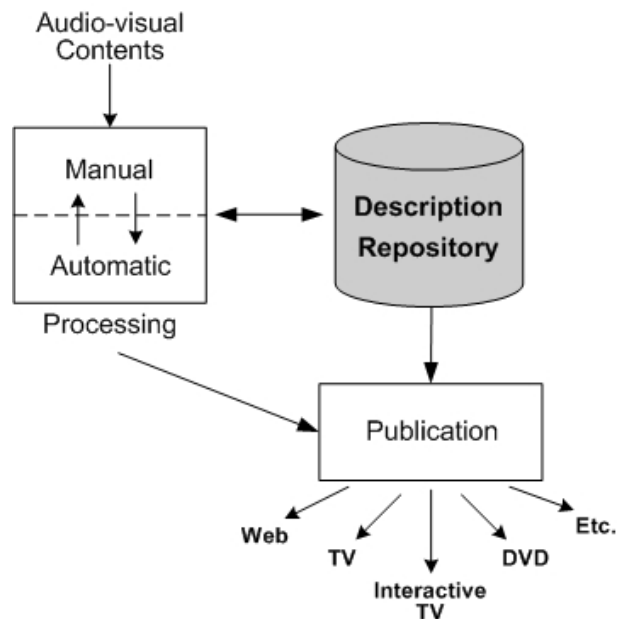


**Figure 4. Dynamic creation of C# classes for representing the descriptor hierarchy**

The framework approach has been adopted to fulfill the first goal of the project, that is, facilitating the industrial development of the set of applications described above. Generally, a framework is considered as a semi-complete application defined by a hierarchy of classes collaborating according to a predefined schema. The framework has then to be completed to produce a fully operational application. In the framework, descriptions are stored and managed in a description server which is one of the main components of the whole system and which relies on a XML native database. Other functionalities provided by the framework are:

- Automatic indexing tools, such as shot segmentation, face detection and recognition, text detection and recognition, speech recognition and language processing of the transcription, sound segmentation into speech, music, laughs, jingles, etc. The project aims at stressing particularly on the joint use of low-level descriptors extracted by each tool and higher-level semantics concepts.

- Organization of audio-visual contents, stream delivery for visualization and frame and sound sample access for manipulation.
- Unified identification of documents to assure independence of format and localization of media.



**Figure 5. Global schema for developing applications in FERIA**

To demonstrate the validity of the approach, two applications will be developed within this framework. The first one is an application of multiple delivery production that should help creating descriptions to produce an enriched presentation of a filmed opera to be delivered on different channels such as DVD, web and especially interactive TV. These descriptions consist in synchronized summaries, synchronized libretto, presentations of characters with their photos, biographies of artists, etc. Once produced, all these metadata will be published with the film into interactive multimedia presentations in the appropriated format for each channel. The second application aims at automatically creating a web site for a collection a programs. By collection, we mean a set of programs which share some characteristics, such as all the installments of a soap, some periodic magazines with the same set and anchor person, the evening news for a given channel, etc. A variety show from the INA's funds has been selected to test this application that will finally apply some automatic indexing tools and publish the results into a web site, offering research and navigation facilities inside the videos of the collection.

## 8 Conclusion

Enabling an intelligent access to audio-visual data requires a powerful description language. In this paper, we have proposed an audiovisual description language which aims to be an alternative to the MPEG-7 standard since we have argued that the latter presents major drawbacks. The main characteristics of this new language are its extensibility and modularity. Moreover, each component can be formally defined using Semantic Web technology, thus providing machine understandable and processable description of the audio-visual contents.

We have now to finalize the implementation of this language and test it in real projects. For that purpose, we plan to represent the full audio-visual ontology described in [13, 14] with this language. The formal definition of each descriptors joint with the offered expressiveness for defining new models should allow to make some reasoning on both the content and the structure of the audio-visual descriptions.

## Acknowledgments

Part of the research described here was funded by the FERIA project under a RIAM grant from the French Ministry of Industry.

## References

- [1] M. Buckland. What is a "document"? *Journal of the American Society of Information Science*, 48(9):804–809, 1997.
- [2] M. Buckland. What is a "digital document"? *Document Numérique*, 2(2):221–230, 1998.
- [3] J. Carrive, F. Pachet, and R. Ronfard. Clavis: a temporal reasoning system for classification of audiovisual sequences. In *Content-Based Multimedia Access (RIAO'2000)*, pages 1400–1415, Paris, France, 12-14 April 2000.
- [4] J. Euzenat, A. Napoli, and J.-F. Baget. XML et les objets (Objectif XML). *RSTI L'Objet*, 9(3):11–37, 2003.
- [5] V. Haarslev and R. Möller. Racer system description. In *International Joint Conference on Automated Reasoning (IJCAR'01)*, pages 701–705, Siena, Italia, 18-23 June 2001.
- [6] J. Hunter. Adding Multimedia to the Semantic Web - Building an MPEG-7 Ontology. In *First International Semantic Web Working Symposium (SWWS'01)*, Stanford, Californie, USA, August 2001.
- [7] J. Hunter and F. Nack. An overview of the MPEG-7 Description Definition Language (DDL) proposals. *Signal Processing: Image Communication*, 16(1-2):271–293, 2000.
- [8] MPEG-7. Information Technology - Multimedia Content Description Interface. Standard No. ISO/IEC n°15938, December 2001.
- [9] OWL. Web Ontology Language Reference. W3C Recommendation, 10 February 2004. <http://www.w3.org/TR/owl-ref/>.
- [10] R. T. Pédaque. Document: Form, Sign and Medium, As Reformulated for Electronic Documents. Technical report, STIC-CNRS, 8 July 2003. [http://rtp-doc.enssib.fr/rtpenglish/pedauque\\_en.html](http://rtp-doc.enssib.fr/rtpenglish/pedauque_en.html).
- [11] F. Pereira. MPEG-7 Requirements Document V.16. ISO/IEC JTC1/SC29/WG11/N4510. Pattaya, Thailand, December 2001.
- [12] RDF. Resource Description Framework Primer. W3C Recommendation, 10 February 2004. <http://www.w3.org/TR/rdf-primer/>.
- [13] R. Troncy. Integrating Structure and Semantics into Audio-visual Documents. In D. Fensel, K. Sycara, and J. Mylopoulos, editors, *2<sup>nd</sup> International Semantic Web Conference (ISWC'03)*, volume (2870) of *Lecture Notes in Computer Science*, pages 566–581, Sanibel Island, Florida, USA, 20-23 October 2003.
- [14] R. Troncy. *Formalization of Documentary Knowledge and Conceptual Knowledge With Ontologies: Applying to The Description of Audio-visual Documents*. PhD thesis, University Joseph Fourier, Grenoble, France, 2004.
- [15] J. van Ossenbruggen, F. Nack, and L. Hardman. That Obscure Object of Desire: Multimedia Metadata on the Web (Part I). *IEEE Multimedia*, 11(4), October-December 2004.
- [16] J. van Ossenbruggen, F. Nack, and L. Hardman. That Obscure Object of Desire: Multimedia Metadata on the Web (Part II). *IEEE Multimedia*, 12(1), January-March 2005.
- [17] XML Schema. W3C Recommendation, 2 May 2001. <http://www.w3.org/XML/Schema>.