Designing and Using an Audio-Visual Description Core Ontology

Antoine Isaac[1,2] and Raphaël Troncy[1]

[1] Institut National de l'Audiovisuel,
Direction de la Recherche, Équipe DCA
4, Av. de l'Europe - 94366 Bry-sur-Marne
{aisaac,rtroncy}@ina.fr & http://www.ina.fr/
[2] Université de Paris-Sorbonne, LaLICC,
http://www.lalic.paris4.sorbonne.fr

Abstract. Describing audio-visual contents necessitates appropriate terminologies and practices. In this paper, we detail the content of an audio-visual description core ontology that contains all the concepts and relationships necessary for a wide range of applications using the audio-visual descriptions. This ontology is based on well-established documentary practices and standardized vocabularies. We show how it can be linked to foundational ontologies and how it can be specialized for specific applications.

1 Introduction

While it is easier than ever before to produce audio-visual (AV) material, it is still problematic to process it efficiently. Digital video documents are now widely available: homemade or produced by professionals, they are broadcasted on the Web (TV on the Web, trailers, advertisements, etc.) or on peer-to-peer networks. AV documents are used in various contexts and for different purposes: in education, they are part of the teaching material; in humanities research, they serve as illustrations for theory presentations; in the general production life cycle, they are more and more re-used for making new documents. However, despite this wide acceptance, the video content is not easily processable.

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 manipulated. The MPEG-7 standard [7] has been recently created for such a purpose, offering tools (including vocabulary and grammatical specifications) to represent AV content by the means of structured textual descriptions. Besides well-defined terminologies, it is desirable to formalize additional guidelines and description best practices in order to guarantee the interoperability of the metadata. For instance, INA[3] investigates the use of formal ontologies and description

[3] The *French National Institute of Audiovisual* (INA) has been archiving and indexing the TV and radio programs broadcasted in France for thirty years and thus, has to manage huge audio-visual databases.

patterns for creating relevant content descriptions of the video documents [11, 12]. Actually, the use of these formal ontologies allows one to improve the automatic manipulation of the metadata since the semantics of the information become machine-accessible. In particular, reasoning support can be provided when querying these data. However, the design of an audio-visual ontology is a complex task as the use of video descriptions may differ a lot. This article proposes to tackle this problem by proposing an audio-visual description core ontology useful for a wide range of applications using AV material.

In the next section, we give several application examples that use descriptions of video documents but that do not focus on the same features. However, all these applications share common concepts and properties and may benefit from a formalized description process, motivating here the need for an audio-visual core ontology. In section 3, we present the sources of information we have used to design such an ontology (existing terminologies and ontologies, documentalist best practices, etc.) as well as the methodology of ontology construction we have followed. In section 4, we detail our audio-visual description core ontology proposal, and how it is linked to a foundational ontology. We show then how this core ontology can itself be specialized for specific applications (section 5). Finally, we give our conclusions and outline future work in section 6.

2 Using Audio-Visual Documents for Various Purposes

The applications that use audio-visual document descriptions are interested in different aspects. They have their own viewpoint on this complex media and usually they are just concerned with selected pieces of information corresponding to their needs. For instance:

- Many tools aim at indexing automatically audio-visual content by extracting low-level features from the signal. These features concern video segmentation (in shots or in sequences), speech transcription, detection and recognition of camera motions, faces, texts, etc. This family of applications needs a common vocabulary to store and exchange the results of their algorithms. The MPEG-7 standard [7] defines such descriptors, without giving them a formal semantics.
- A TV (or radio) broadcaster may want to publish the program listings on its web site. Therefore, it is interested in identifying and cataloguing its programs. The channel would like also to know the detail of the audience and the peak viewing times in order to adapt its advertisement rates. Broadcasters have recently adopted the TV Anytime[4] format and its terminologies to exchange all these metadata [9].

[4] 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 in the context of TV digital broadcasting. The forum identified metadata as one of the key technologies enabling their vision and have adopted MPEG-7 as the description language.

– A news agency may aim at delivering program information to newspapers. It could receive the TV Anytime metadata, and enrich them with the cast or the recommended audience of the program, the last minute changes in the program listings, etc. The ProgramGuideML[5] format is currently developed for this purpose.
– Education or humanities research use more and more the audio-visual media. Their needs concern the possibility to analyse its production (*e.g.* number, position and angle of the camera, sound recording) and to select and describe deeply some excerpts according to domain theories, focusing for example on action analysis (*i.e.* a praxeological viewpoint).
– Finally, an institute like INA has to collect and describe an audio-visual cultural heritage. It is interested in all the aspects given above, with a strong emphasis on a documentary archive viewpoint.

Despite this variety, all these specific applications share common concepts and properties when describing an AV document. For instance, the concept of *genre* or some *production* and *broadcast* properties are always necessary, either for cataloguing and indexing the document, or to parameterize an algorithm whose goal is to extract automatically some features from the signal. We observe also that the *archive* point of view is an aggregation of the usual description facets. We have therefore formalized the practices of the documentalists of INA as well as the terminology they use, in order to design an audio-visual description core ontology. Before presenting the result of this formalization, we introduce briefly in the next section our information sources and the methodology of ontology construction followed.

3 Methodology

As we have seen in the previous section, several terminologies have been recently standardized (MPEG-7, TV Anytime) or are still under development (ProgramGuideML). The INA institute, with the help of the English BBC and the Italian RAI channels, has also developed a complete terminology for describing radio and TV programs. However, none of these terminologies have a formal semantics, that is, the English prose in the documentation is the only way to understand the meaning of the vocabulary. Hence, the applications cannot easily access the semantics of the descriptions, which is obviously a major drawback for their interoperability.

Our approach was then to further formalize these terminologies and documentary practices in order to give a formal semantics for all the concepts used while describing audio-visual content. The resulting audio-visual ontology benefits also from older attempts in this way. For instance, for alleviating the lack of

[5] The ProgramGuideML initiative is developed by the International Press Telecommunications Council (IPTC) (`http://www.programguideml.org`) and aims to be the global XML standard for the interchange of Radio/TV Program Information.

semantics of MPEG-7, Hunter [5] and Tsinaraki [13] have already proposed ontologies expressing formally the semantics of the MPEG-7 metadata terms. These ontologies, built by reverse-engineering of the existing XML Schema definitions together with the interpretation of the English-text semantic descriptions, are represented using Semantic Web languages (OWL/RDF, [8, 10])[6]. However, they cover only the descriptors standardized by MPEG-7, that are mainly related to the physical features or the structural decomposition 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). Previous work by the authors [11] described a more general architecture based on ontologies to describe formally the content of videos (OWL/RDF), and documentary tools (MPEG-7/XML Schema) to constrain their structure, to finally offer reasoning support on both aspects when querying the database. If this architecture is promising, it is very dependent on a well-founded AV description core ontology, which motivates the work presented in this paper.

Various approaches have been reported to build ontologies [3], but few fully detail the steps needed to obtain and structure the taxonomies. This observation has led us to propose a methodology entailing a semantic commitment to normalize the meaning of the concepts. This methodology, detailed in [1], emphasizes the conceptualization step since the ontologist has to express in natural language differentiating principles for each concept and property, which justifies the taxonomies in the domain targeted by the ontology. The **DOE**[7] (*Differential Ontology Editor*) tool, an ontology editor implementing this methodology, has been used prior to the **Protégé2000** environment which has enabled the final formalization of the ontology in OWL. Once formalized, we have linked it to a part of the DOLCE foundational ontology [6], thus guaranteeing a well-founded audio-visual core ontology.

4 A Proposed Audio-Visual Description Core Ontology

In this section, we present our audio-visual description core ontology. First, we detail its content (section 4.1). Second, we show how this ontology is linked to the DOLCE foundational ontology, in order to give it a sound and consensual upper-level justification (section 4.2). Section 4.3 points to the problems that then emerged, and the technical tricks that are needed to reconcile both visions. The resulting ontology is available at `http://opales.ina.fr/public/ontologies/coront/`.

4.1 The Ontology Content

From a documentary standpoint, AV documents can be analyzed following two dimensions, production and broadcast, the two main activities that concern

[6] Jane Hunter's MPEG-7 ontology is available at `http://metadata.net/harmony/MPEG7/mpeg7.owl`.

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

them. In fact, the descriptions have to keep a trace of how a document was produced (who are its authors, what is its structural organization, etc.) and, subsequently, how it was broadcasted (when, how often, on which channel, etc.). However, there is no need for the common documentalist to know when and how long the shooting of a sequence included in a given program took, nor that its emission implied the participation of a specific telecommunication actor. Therefore, the ontology does not contain an exhaustive list of all specific AV-related activities, but rather the ones that concern the core of the document description task such as defined by the professional documentalists.

The main concept is the *AV production object*, which represents the core notion of an AV document. The first distinction occurs between a *program*, (a rather stand-alone entity from the points of view of production and broadcast), and a *sequence* (a part of a program or other sequences). These concepts are then specialized by means of form or content-linked differentiating features in order to obtain the classification scheme that is common to all needs: their *genre*. For example, programs are divided into *heterogeneous* and *homogeneous*: the first is characterized by a sequence of autonomous elements in form and in content, unlike the second. They are then classified according to their length, and to their general content (fiction, informative, entertainment). After some further specialization, one can find the usual TV genres: *sitcom*, *tv show*, *documentary*, etc. (see Figure 1).

The notions used to characterize the AV objects are also defined in the ontology. First, we have introduced a hierarchy of the roles that people can play in a program, whether as authors (*producer*, *director*) mentioned because of their importance in the program production, or participants (*host*, *actor*), being part of the description since they are visible or audible in the document. Then, we can find a large set of AV properties that mirror a given production or broadcast preoccupation or mode. They are organized according to their belonging to the production world (way of filming, such as *camera motion*, editing or post-producing, such as *text insertion*) or to the broadcasting one (*periodicity*, *intended audience*, etc.). A typology of the general *themes* that a document can refer to completes the ontology.

We have then to introduce relations to structure the domain knowledge and to enable reasoning. Observing the needs expressed in our sources, we could elicit some template structures – see figure 2 – that account for such articulations. To sum up, we need (i) relations amongst AV objects (mainly, the mereological relation *part-of* between sequences and programs), (ii) relations between these objects and their properties (whether broadcast properties of production ones, such as an *hasForAuthor* or *lasts*) and (iii) relations linking these objects and the entities "from the real world" which they represent, and that will be provided by domain-specific ontologies. These relations allow us to specify AV-related reasoning knowledge. For instance, the code below gives the formal definition in OWL of a *dialog sequence* which is "a spoken sequence that has at least two participants".

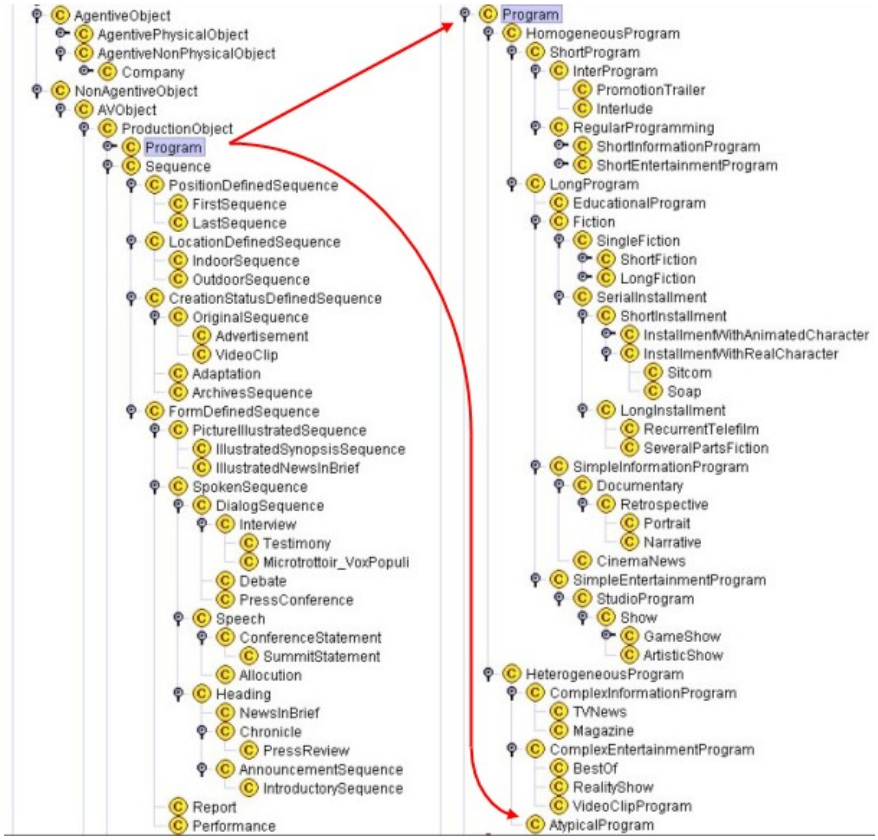

Fig. 1. Concepts taxonomy: the possible genres for characterizing the sequences (*on the left*) and the programs (*on the right*) in the AV description core ontology

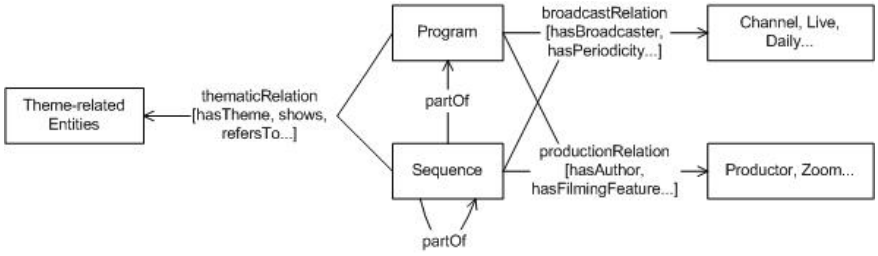

Fig. 2. Informal sketch of templates used for AV document description (general notions and examples)

```
<owl:Class rdf:ID="DialogSequence">
  <rdfs:subClassOf rdf:resource="#SpokenSequence"/>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty>
        <owl:ObjectProperty rdf:about="#hasParticipant"/>
      </owl:onProperty>
      <owl:minCardinality rdf:datatype="&xsd;#int">2</owl:minCardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
</owl:Class>
```

The knowledge can also be more complex and general inference rules allow us to complete the formalization of the ontology. For instance, the code below states in SWRL [4] that "if a program contains a sequence which is presented by a person, it has this person as a participant".

```
<ruleml:imp>
  <ruleml:_body>
    <swrlx:classAtom>
      <owlx:Class owlx:name="Program"/>
      <ruleml:var>prgm</ruleml:var>
    </swrlx:classAtom>
    <swrlx:classAtom>
      <owlx:Class owlx:name="Sequence"/>
      <ruleml:var>sqce</ruleml:var>
    </swrlx:classAtom>
    <swrlx:classAtom>
      <owlx:Class owlx:name="Person"/>
      <ruleml:var>presenter</ruleml:var>
    </swrlx:classAtom>
    <swrlx:individualPropertyAtom swrlx:property="hasPart">
      <ruleml:var>prgm</ruleml:var>
      <ruleml:var>sqce</ruleml:var>
    </swrlx:individualPropertyAtom>
    <swrlx:individualPropertyAtom swrlx:property="presentedBy">
      <ruleml:var>sqce</ruleml:var>
      <ruleml:var>presenter</ruleml:var>
    </swrlx:individualPropertyAtom>
  </ruleml:_body>
  <ruleml:_head>
    <swrlx:individualPropertyAtom swrlx:property="hasParticipant">
      <ruleml:var>prgm</ruleml:var>
      <ruleml:var>presenter</ruleml:var>
    </swrlx:individualPropertyAtom>
  </ruleml:_head>
</ruleml:imp>
```

4.2 The Ontology Foundations

Related to the Semantic Web initiative, some foundational ontologies have emerged, grounded on philosophy or linguistics, that can be used as references for more domain-specific ontology building. To benefit from such potentially consensual approaches, we have chosen to try and link our entities to the DOLCE upper-level ontology [6], and especially to its "Description and Situations" (D&S) module [2]. These works come with rather complete formal axiomatizations, and the *ontological patterns* provided are very interesting, since we could compare our own templates with them.

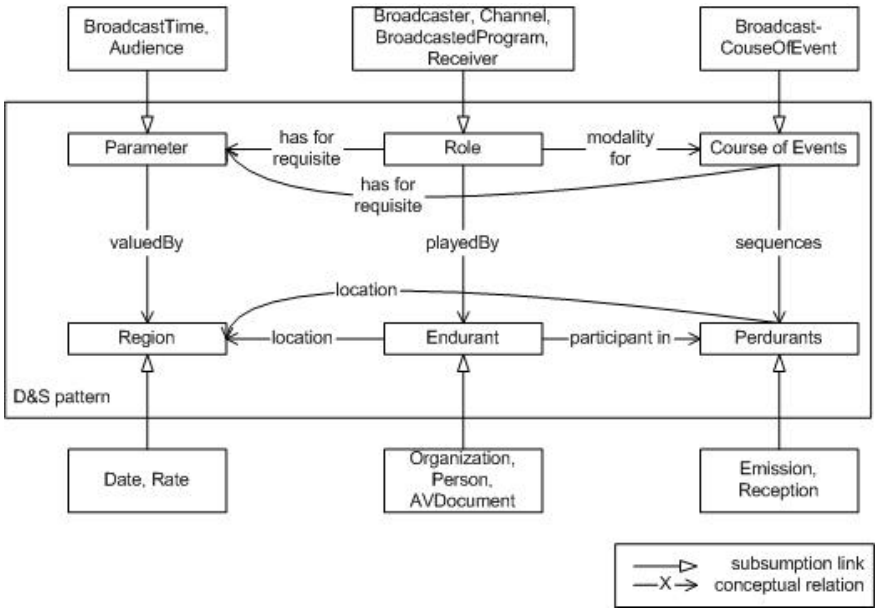

Fig. 3. Introduction of AV core concepts following a part of D&S pattern

The D&S module provides a pattern that can be used to describe situations. These descriptions are composed of *descriptions* for *courses of events*, *roles* that entities can *play* in these events, and *parameters* that are used to describe roles and events and that are valued in abstract *regions*. To adapt these notions to the AV domain, we have applied this pattern to the two main activities we have previously mentioned: production and broadcast. For instance, the description of the broadcast of an AV document involves a broadcast course of events, which *sequences* broadcast actions such as *sending* the program through a network channel. Broadcast roles, such as *broadcaster* and *receiver*, are played by entities such as social organizations (*companies*) or *persons*. In the description of these events, we find parameters like *broadcast time* (valued by *dates*) or *periodic-*

ity (*daily*, *weekly*, etc). Figure 3 shows an example of how these concepts are specialized in our audio-visual core ontology.

4.3 Reconciling Foundational Considerations With Core Domain View

Relating our core domain ontology to a foundational one has helped to clarify our ontological commitment. However, the complexity of the notions involved by this upper-level consideration may hide the original view on the domain core, and thus reduce the relevance of the ontology. For instance, the descriptions we really need here have to be centered on the documents, and we do not have to make all the information expressed according to the D&S pattern explicit. As an example, the knowledge that a given *program* plays a *broadcasted program* role in a broadcast course of events is quite useless in our case: we rather need a simple property pointing at the time when it was broadcasted, and its intended audience, and so on. Indeed, all that deals with the course of events/actions is far from being mandatory from a document-centered point of view.

We have therefore introduced relational shortcuts that could enable future systems to benefit from both legitimacies: the one that is given by the consensual, theory-grounded upper-level view, and the one that results from keeping a cognitively natural representation of the domain. For example, we can consider that a relation *wasBroadcastedAt* that exists between a program and a date is something useful to avoid stating that the program plays the message role in a broadcast course of events which had that date as a value for the broadcast time parameter it had for requisite. The point is that the use of this domain-relevant relation can now be grounded on a sound theoretical basis.

All that we need is to introduce formal reasoning knowledge (including definitions and inference rules) that will allow a knowledge-based system to deal simultaneously with the concepts and relations that represent both viewpoints. For example, the link between our relation and its equivalent upper-level relational path can be – partly – represented by the inference rule represented graphically in Figure 4.

5 Applications

We have seen how we have designed an audio-visual description core ontology and how its branches can be related to a foundational upper-level. This ontology can itself be specialized for specific needs. For instance, it is possible to add all the broadcast properties defined in the TV Anytime format (*e.g.* broadcast status and mode). Hence, these metadata will benefit from the OWL formalization of the ontology and will have a direct traduction in RDF.

Another application, encountered in the OPALES project, was the description of educational AV documents according to a semiotic point of view. A whole range of interpretation values was added to the core ontology, as well as the relations between these interpretations and AV objects and properties, allowing

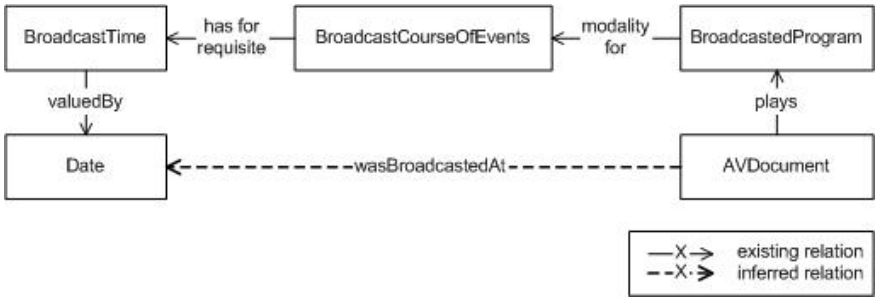

Fig. 4. Rule deducing the *wasBroadcastedAt* relationship from D&S-compliant information

for example to judge the scientific value of a given material. The core ontology was extended with hierarchies of concepts in order to create domain-specific descriptions (in our case, geography), and with relations explaining how the AV objects refer to domain-entities. We can notice that all the broadcast notions of the core ontology were useless for this application, unlike the production-related ones.

6 Conclusion

We have presented in this paper various scenarios that make different use of audio-visual document descriptions. However, all these applications need common concepts and relations that can be modeled into a core ontology. We have detailed the content of this ontology as well as its underlying sources of information and the methodology of construction we have followed. Moreover, we have shown how we can link this core part to the fundational ontology DOLCE, and how we can address the usability problems raised by such a commitment. Finally, we have given two examples of specific applications that may benefit from this ontology by specializing it easily.

The work made until now suffers from a limitation: the bias introduced by our information sources that are rather TV-centered. This is due to the fact that the traditional networks are the only ones to have the will and means to describe the programs they broadcast. A better articulation between the concepts and the relations presented and a more general diffusion (even communication) theory is however desirable. Anyway, the forthcoming implementation of a legal depository for web-broadcasted AV documents should give us a context for experimenting the robustness of the ontology proposed here.

We hope that anchoring our work both in practical observations and foundational considerations will prove efficient when such evolutions, including domain and application extensions, will show up. Until now, this framework has shown to nicely fit our needs. However, further experimentation is still necessary to show that such a methodology is cost-efficient. Especially, the merging of the

practical and foundational viewpoints on the core of the domain will be difficult to maintain since it needs an important conceptualization effort and the computational complexity that comes with it is likely to hamper the algorithms currently used in implemented knowledge-based systems.

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. A. Gangemi and P. Mika. Understanding the Semantic Web through Descriptions and Situations. In *International Conference on Ontologies, Databases and Applications of Semantics (ODBASE'03)*, Catania, Italy, 2003.
3. A. Gómez-Pérez, M. Fernández-López, and O. Corcho. Ontological Engineering. *Advanced Information and Knowledge Processing*, Springer Verlag, 2004.
4. I. Horrocks, P. F. Patel-Schneider, H. Boley, S. Tabet, B. Grosof and M. Dean. SWRL: A Semantic Web Rule Language Combining OWL and RuleML. W3C Member Submission, 21 May 2004. `http://www.w3.org/Submission/SWRL/`
5. 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, USA, 2001.
6. C. Masolo, S. Borgo, A. Gangemi, N. Guarino, A. Oltramari and L. Schneider. DOLCE : a Descriptive Ontology for Linguistic and Cognitive Engineering. WonderWeb Project, Deliverable D17 v2.1, 2003.
7. ISO. Information Technology - Multimedia Content Description Interface (MPEG-7). Standard No. ISO/IEC 15938:2001, International Organization for Standardization(ISO), 2001.
8. OWL, Web Ontology Language Reference Version 1.0. W3C Recommendation, 10 February 2004. `http://www.w3.org/TR/owl-ref/`
9. 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.
10. RDF, Ressource Description Framework Primer W3C Recommendation, 10 February 2004. `http://www.w3.org/TR/rdf-primer/`
11. R. Troncy. Integrating Structure and Semantics into Audio-visual Documents. In *Proc. of the 2nd International Semantic Web Conference (ISWC'03)*, LNCS 2870, p.566-581, Sanibel Island, Florida, USA, 2003.
12. 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.
13. C. Tsinaraki, P. Polydoros and S. Christodoulakis. Integration of OWL ontologies in MPEG-7 and TV-Anytime compliant Semantic Indexing. In *Proc. of the 16th International Conference on Advanced Information Systems Engineering (CAiSE 2004)*, Riga, Latvia, 2004.