Semantic description of multimedia documents: the MPEG-7 approach

Adami Nicola, Bugatti Alessandro, Leonardi Riccardo, Migliorati Pierangelo

DEA - University of Brescia
Via Branze 38, 25123, Brescia, Italy
{adami,bugatti,leon,pier}@ing.unibs.it

Abstract. This article addresses a possible approach on how to index a multimedia document in a semantic manner. At present, both automatic and semi-automatic tools allow to index a document using low level features, obtaining structural information like shot segmentation, speaker separation and so on. In spite of its utility, this information is too elementary to perform some tasks like querying, retrieval or authoring of other descriptions. Instead enabling powerful description schemes that allow to describe things like semantic time and place, concepts, relations among them and others, ensures an effective description of the semantic content of documents. In particular we analyze a description scheme, the Semantic DS, developed within the MPEG-7 standardization process[1], and we give an explanation of its components and examples for their usage.

Introduction

Since a few years, a huge amount of audio-video digital material from multiple sources (digital radio and television, Internet, digital archives and others) has been made available to each single user and this amount will increase dramatically in the next years. All this overflow of information brings some problems: how to extract only relevant (from the user's point of view) material? how to browse all the material in an efficient way? how to query all the "database" of available material by non trivial questions (for instance "Recover all videos containing material related to a wedding")? The starting point for answering all of these needs is to have a description of the whole content. This description should be rich and powerful enough in order to capture not only the syntactical structure of the document (for example the subdivision of a movie in scenes and shots) but also its semantic content. According to this requirement a lot of research activities have been started ([1–6] among others) and within the MPEG-7 standardization process many efforts have been made in order to create a Description Scheme (DS) which is able to represent the semantics of any document (audio, video, images, ...). The result of this latter effort is the definition of the Semantic DS as defined in MPEG-7 Final Committee Draft [7]. This contribution outlines concepts and ideas that have lead to this DS. In section 1 we give a short introduction to the MPEG standardization group with a focus on the MPEG-7 phase. In section 2 there is a description of the Semantic DS and the tools contained in it. In section 3 an example is provided in order to clarify the previously exposed concepts and the XML[2] syntax of the generated descriptions that use the Semantic DS. In section 4 we finally explain how to use the abstraction mechanism to create templates for descriptions reuse and to make inference about the world described in a description.

1 The MPEG-7 standard

The ISO/IEC JTC1/SC29/WG11, also called MPEG (Motion Picture Expert Group), has been set up by the ISO/IEC standardization body in 1988 to develop standards for the coded representation of moving pictures, associated audio and their combination. While the first two phases, MPEG-1 and MPEG-2, focused on coding and compression of audio and video signals and the third MPEG-4 added new functionalities, the MPEG-7 standard (for an introduction see [8]), also known as "Multimedia Content Description Interface", aims at providing standardized core technologies allowing description of audio-visual content in multimedia environments. This is a challenging task given the broad spectrum of requirements and targeted multimedia applications, and the broad number of audio-visual features of importance in such context. In order to achieve this broad goal, MPEG-7 standardizes:

[1] This work is a presentation of the work of the MPEG7 Multimedia Description Scheme Group inside the ISO/IEC JTC1/SC29/WG11 standardization process

[2] XML is a trademark of W3C group

- **Datatypes** that are description elements not specific to the audio-visual domain that corresponds to reusable basic types or structures employed by multiple Descriptors and Description Schemes.
- **Descriptors** (D) to represent Features. Descriptors define the syntax and the semantics of each feature representation. A Feature is a distinctive characteristic of the data, which means something to somebody. It is possible to have several descriptors representing a single feature, i.e. to address different relevant requirements. A Descriptor does not participate in many-to-one relationships with other description elements, i.e. it is not possible to have many instances of the same descriptors related to the same descripted element.
- **Description Schemes** (DS) to specify the structure and semantics of the relationships between their components, which may be both Ds and DSs. A Description Scheme shall have descriptive information and may participate in many-to-one relationships with other description elements.
- A **Description Definition Language** (DDL) to allow the creation of new DSs and, possibly, Ds and to allow the extension and modification of existing DSs.
- **Systems tools** to support multiplexing of descriptions or description and content, synchronization issues, transmission mechanisms, file format, etc.

The standard is subdivided into seven parts:

1. **Systems**: Architecture of the standard, tools that are needed to prepare MPEG-7 Descriptions for efficient transport and storage, and to allow synchronization between content and descriptions. Also tools related to managing and protecting intellectual property.
2. **Description Definition Language**: Language for specifying DSs and Ds and for defining new DSs and Ds.
3. **Visual**: Visual description tools (Ds and DSs).
4. **Audio**: Audio description tools (Ds and DSs).
5. **Multimedia Description Schemes**: Description tools (Ds and DSs) that are generic, i.e. neither purely visual nor purely audio.
6. **Reference Software**: Software implementation of relevant parts of the MPEG-7 Standard.
7. **Conformance**: Guidelines and procedures for testing conformance of MPEG-7 implementations.

At the moment MPEG-7 is not yet a standard but it is in the Final Committee Draft phase. MPEG-7 will be a standard in September 2001.

In this article we will address only a DS contained in the part 5 about Multimedia Description Scheme, the Semantic DS which is the tool that enables us to give a semantic description of the content.

2 The Semantic DS

At present we have the possibility to describe an AV document in terms of low level features like histogram, dominant color, audio loudness and so on, but these tools often are too elementary to provide useful information about the content of a multimedia document. Moreover very relevant information are implicit inside a semantic description and we need a powerful tool in order to describe this kind of information.

A typical information that we would describe can be represented by this example: "John and his wife Mary bought a book about New York in the shop at the corner of the 5-th street last Saturday during their holidays". In this case it is easy to see that there is a lot of information regarding people, times, places and relations among them. At the same time there are concepts like wife or shop that are not self-referenced in this example, but they need to have an explanation, a sort of definition in some elementary manner. For accounting of all these problems a Description Scheme has been proposed, the Semantic DS [7], which aims to represent the semantic content of the narrative world, where the narrative world can be the world depicted in the specific instance of AV content (e.g. in a movie which contains a spy story, the spy story is the narrative world) or in any abstracted description of the possible worlds described in some possible media occurrences (e.g. the city of New York in winter, independently from any media which can contain it).

As shown in figure 1 the Semantic DS is composed by many "tools" that can be divided in the following categories:

- **Semantic entity tools** : these tools describe semantic entities in the narrative world such as objects, events, places and times.
- **Semantic attribute tools** :These tools describe attributes of semantic entities related to abstraction levels and semantic measurements in time and space.

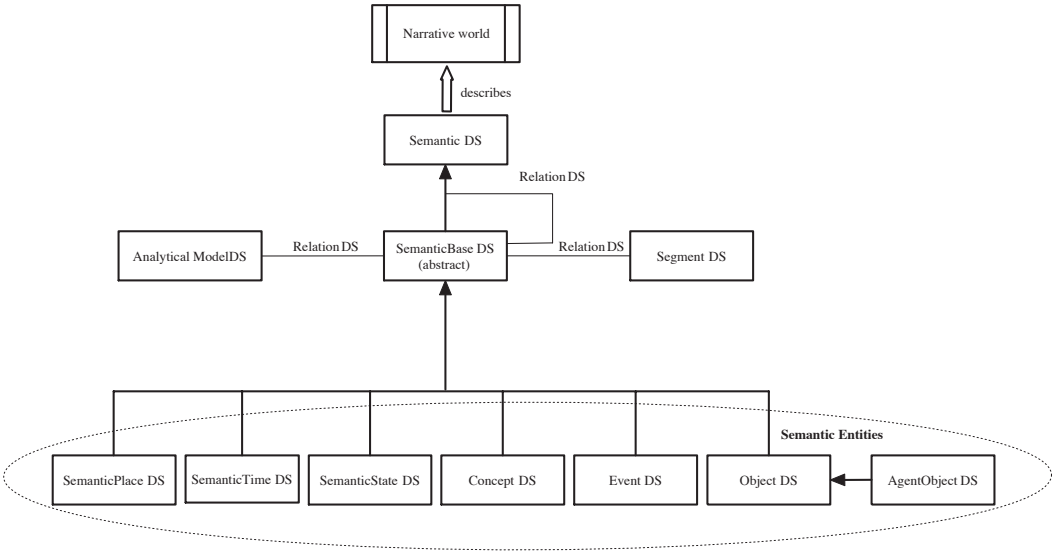

Fig. 1. Model of the tools for describing the semantics of AV content.

– **Semantic relation tools** : these tools describe relations among semantic entities, among segments and semantic entities, and analytical models and semantic entities.

Each of them is defined using a Description Definition Language (DDL) based on XML Schema standard, which seemed the more effective choice because it is an emerging standard for exchanging and encoding of information. In addition there are also many tools both for parsing and browsing XML descriptions.

Let us now take a further look to each DS, excluding the Semantic DS and Semantic Base DS, which have been created only to enable an inheritance mechanism.

2.1 Object DS

The Object DS describes a perceivable or abstract object in a narrative world. A perceivable object is an entity that exists, i.e. has temporal and spatial extent, in a narrative world (e.g. the book in the previous example). An abstract object is the result of applying abstraction to a perceivable object (e.g. any book). Essentially, this generates a template of the object in question. This DS has the possibility to include other objects in order to describe the decomposition of the object into sub-objects.

The AgentObject DS extends from the Object DS. The AgentObject DS is a specialized Object DS that encloses the Agent DS[3] within the Object DS in order to represent agents like people, organizations or groups of persons.

2.2 Event DS

The Event DS describes a perceivable or abstract event in a narrative world. A perceivable event is a dynamic relation involving one or more objects or events occurring in a region in time and space of a narrative world (e.g., John and Mary buy a book). An abstract event is the result of applying abstraction to a perceivable event (e.g., anyone buy a book). Essentially, this generates a template of the event in question. An event represents a change in the (combined) state for one or more objects.

2.3 Concept DS

The Concept DS describes an entity that is not the result of abstraction (or generalization) of any concrete entity (e.g., "harmony", "full-bodied", or "freedom"). A concept is usually represented by a group of objects and events, with metaphor and analogies, and blending several abstract concepts together. Concepts can formally be described as collections of properties.

[3] The Agent DS represents the abstract concept of an "agent": it can be a person (e.g., actor, director, character, dubbing actor), an organization (e.g., a company) or a group of persons (e.g. a musical ensemble)

2.4 SemanticPlace DS

The SemanticPlace DS describes a location in a narrative world. The SemanticPlace DS is a specialized SemanticBase DS that encapsulates the Place DS within the SemanticBase DS. The Place DS is currently defined by MPEG-7 as a tool to include information like name, country, postal address and so on. In the previous example a Place could be both the corner of the 5-th street or New York City. It is also possible to describe an extent of a spatial interval, like for example "within 4 miles around New York City".

2.5 SemanticTime DS

The SemanticTime DS describes a time in a narrative world. The SemanticTime DS is a specialized SemanticBase DS that encapsulates the Time DS and semantic relative time information within the SemanticBase DS. The Time DS is defined by MPEG-7 CD as a tool to include information like duration, time origin, time unit, time interval.

2.6 SemanticState DS

The SemanticState DS describes and parameterizes semantic properties of a semantic entity at a given time, in a given spatial location, or in a given media location (e.g., height and weight). It is a set of numerical and verbal attributes that can be attached to semantic entities such as objects and events and other semantic elements such as semantic relation graphs.

2.7 AbstractionLevel

The AbstractionLevel datatype describes the kind of abstraction that has been performed in a description. When it is not present in the description, then the description is concrete - it describes the world depicted by AV content and references the AV content (e.g., through a MediaOccurrence element in SemanticBase DS or a Segment DS). If the AbstractionLevel is present in the description, a formal abstraction is in place - the description contains variables or is a template for other descriptions. This tool is very relevant because it allows to create templates for entities and it can be used to make simple inferences about the world described in the scheme and to improve performance in tasks as retrieval and querying. We will explain further this tool in section 4.

2.8 Relation DS

The Relation DS is a tools set for describing relations among semantic entities, among segments and semantic entities, or among analytic models and semantic entities, as shown in figure 1. In document [7] it is possible to find all the list of the relations defined inside the standard. Every type of relation (object-to-object, object-to-event, event-to-event and so on) is composed of many elementary relations belonging to this class of relations. For instance in the object-to-object class the defined relations are *memberOf, partOf, componentOf, substanceOf.* Obviously in the Relation DS there are two fields, Source and Target, to identify the source and the target of the relation. The individual relations can be described via inline relations in the SemanticBase DS descriptions or it may also use a Graph DS descriptions in order to create very complex descriptions.

2.9 Other DSs related to Semantic DS

We describe hereafter other DSs or components used in the Semantic DS or related to it:

- **MediaOccurrence DS**: describes appearances of semantic entities in the media. This can be reached either using a MediaLocator (a "pointer" to a piece of video or an image or a sound) or using an instance of a descriptor (for example the histogram of a picture). The purpose of this description scheme is to provide access to the same media information as the Segment DS, but without the hierarchy and additional temporal and spatial information.
- **Segment DS**: is an abstract DS from which to derive tools for describing spatial, temporal, or spatio-temporal segments of AV content and hierarchical structural decompositions of segments.

- **Graph DS** : represents a graph of relations among MPEG-7 description schemes-e.g. a graph representing the spatio-temporal structure of a set of segments. Although hierarchical structures (trees) are adequate in many cases, (for example, for efficient access and retrieval), some relations cannot be expressed using such structures. The Graph DS can describe more general relations among MPEG-7 description tools.
- **AnalyticModel DS**: Analytic models describe the association of labels or semantics with collections of AV content. The collections may be specified by enumerating the members of the collection, such as by using ContentCollections or DescriptorCollection, or may be specified using parameterized representations of the collections in the form of ProbabilityModels. This allows the Analytic models to characterize the representation of different semantic concepts by models in terms of clusters of AV data, example collections of AV content descriptions, or probability models.

3 An example of istantiation

For a better understanding of the Semantic DS we show how to represent the example given in section 2: "John and his wife Mary bought a book about New York in the shop at the corner of the 5-th street last Saturday during their holidays". The first thing to note is that the description of a document is not unique, i.e. it depends from the author of the description and from his/her point of view. A possible representation can be the one shown in figure 2, in UML-like fashion.

The DSs described in the previous sections allow to create a description of this example by means of the XML Schema defined by the DDL language: the syntax is quite intuitive (see Table 1 at the end of this document), but for a further explanation see the reference document in the MPEG-7 Final Committee Draft [7].

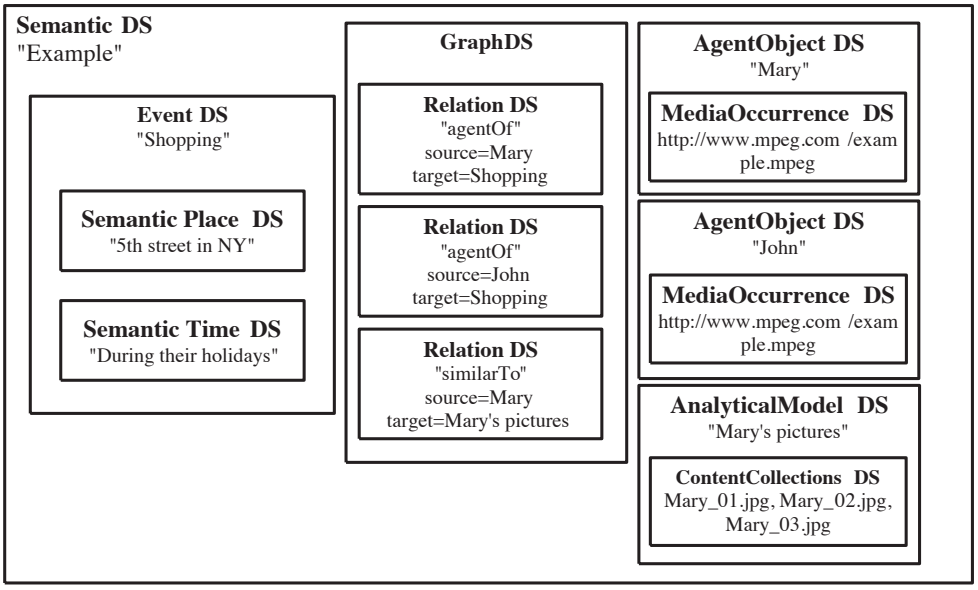

Fig. 2. Example of an instantiation of the Semantic DS.

4 Abstraction and templates

The representation of semantic content described in the previous example is a useful tool in order to establish a richer description than a low level description, but we need a mechanism to achieve abstraction, which can be used to link a real description (i.e. a description which is related to a concrete AV document) to general concepts. The idea is to create a sort of "template" for a particular event, object or concept and to link the real description with the corresponding entities from the template.

For the previous example a possible template can be seen in figure 3 (and its representation using XML is shown at the end of this document in Table 1). The first thing to note is the presence of the label

AbstractionLevel, which means that a kind of abstraction is present. In this case the AbstractionLevel value is set to 1, i.e. the description is not refered to a piece of multimedia document, but it represents an abstract description of the world, in particular of three events: the act of taking, buying and making a payment. In the upper part of the figure there is the event "To take", which involves the presence of a "Performer", i.e. the subject that makes the action, represented by an AgentObject DS and an "Object", i.e. the subject that receives the action, represented by an Object DS. There are two relations among them: the "Performer" is the *agentOf* the event "To take" and the "Object" is the *patientOf* the event "To take". Both these relations and the others used later are standarized in [7]. The event "To buy" is a *specializationOf* the event "To take": this means that also in this event there is a Perfomer and an Object, but there are other entities as a Customer (which is a *specializationOf* the AgentObject "Perfomer") and an AgentObject named "Seller", which is the *sourceOf* the event "To buy". Besides the event "To buy" entails another event, the event "Payment" which includes the entity "Price" in itself. We used for this model an Entity-Relationship representation , but normally this is described using the MPEG-7 DDL language. Here there is a very simple example of the possibilities of the abstraction mechanism. We defined the action "To buy" using relations between this event and its partecipants (the customer, the seller, the object which is bought). In addition, by using the relation *specializationOf* it is possible to derive the action "To buy" from the action "To take" in a sort of Object Oriented approach.

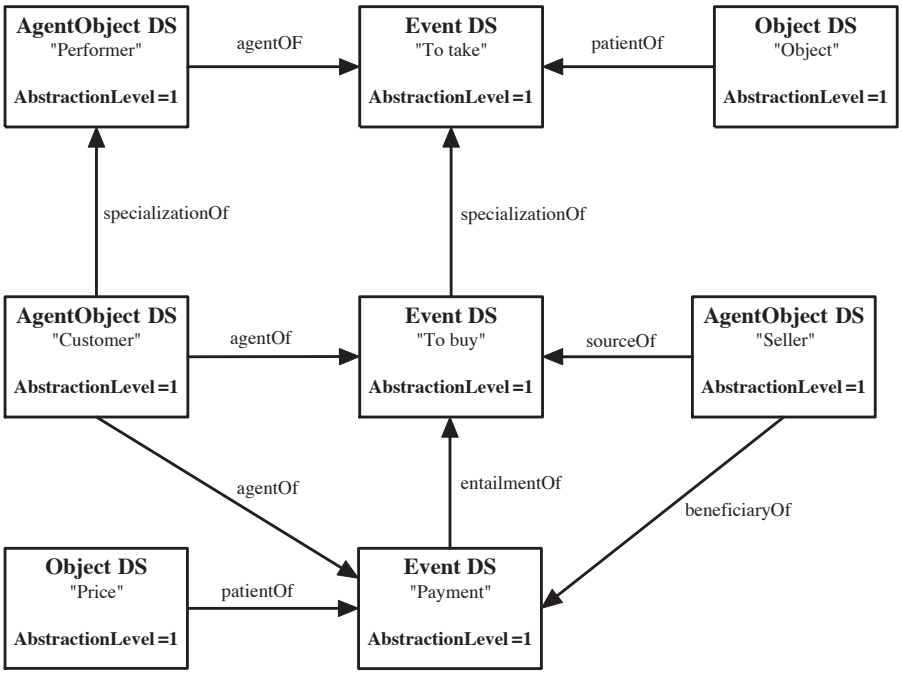

Fig. 3. Solution for a possible template according to the example.

What is the aim of this conceptual modeling? We have at least three main points supporting this mechanism:

- we can link a concrete description to an abstract one, in order to add semantics and to indicate the rules of each entity in that description. In the previous example John and Mary should be costumers, the book should be the object of the transaction, the event "Shopping in New York" should be the event "To buy". This way it is possible to have relations between entities in a concrete description and this "meta-information" could be used for an improved retrieval or to make some simple inferences. For instance having a huge database of semantic descriptions similar to the previous one, we could ask for queries like "Retrieve all the scenes where there is a purchasing and the buyed goods have a price higher than 100 dollars" or "Retrieve all the scenes where the seller is John Smith". These queries are not possible in a normal textual description.
- we can generate a sort of "template" for interesting concepts in every particular domain. For instance in a soccer match interesting concepts could be goals, faults, corners, player's roles and so on. Having

MPEG-7 descriptions of these concepts enables the people who create descriptions of real AV material to link each entity in the program with the corresponding templates. Different operators (content creators, broadcasters, ...) could share the same templates obtaining complete semantic descriptions.
- in a dual fashion, we can have authoring tools in order to create new descriptions according to a sort of knowledge base implicit in templates. Each template can be used as a kind of unfilled form, with empty fields for each concept related to it. For instance the template event "goal" could have an AgentObject who is the goal maker, an other AgentObject who is the goal keeper, a SemanticTime to contain the instant of the objective goal. The operator should only fill this "fields" with the real entities in the actual match that have to be described.

5 Conclusions

In this paper we focused on a new approach for the description of multimedia document. While descriptions in terms of scenes, shots, type of underlying audio and so on are very effective for browsing and other very simple tasks, we need another level of description in order to take into account the semantics of the AV content. Many efforts have been made in multimedia community about this item and, altough the research is in a preliminary stage, also inside the MPEG-7 standardization process it has been developed a Description Scheme with the aim to "encode" the semantics of any AV material. We briefly presented the MPEG-7 standardization process and then we focused on the Semantic DS, the tool to realize semantic descriptions of the content. After an explanation of its structure and its components, we showed the possibility to have an abstraction mechanism and to build templates for concepts. Both these mechanisms enable us to perform tasks like querying and retrieval in a more effective way and to realize authoring tools in order to create compliant descriptions shared between different operators.

6 Acknowledgements

The Semantic Description Scheme presented in this paper is the result of the contributions and collaborative efforts of many people. The authors are particularly grateful to the members of the MPEG Multimedia Description Scheme Group for their many contributions towards the development of the Semantic Description Scheme in the MPEG-7 standard. Among others we want in particoular to thank Ana B. Benitez [4], Hawley Rising [5], Corinne Jörgensen [6] , Koiti Hasida [7], Rajiv Mehrotra [8], A. Murat Tekalp and Ahmet Ekin [9], Toby Walker [10].

References

1. N. Adami, A. Bugatti, A. Corghi, R. Leonardi, P. Migliorati, L. A. Rossi, C. Saraceno, "ToCAI: a framework for Indexing and Retrieval of Multimedia Documents", In Proc. International Conference on Image Analysis and Processing (ICIAP´99), Venice, Italy, Sept. 1999.
2. R. Leonardi et al., "The ToCAI description scheme for indexing and retrieval of multimedia documents", In Proc. European Workshop on Content-Based Multimedia Indexing (CBMI´99), Toulouse, France, Oct. 1999.
3. A. Ferman, A. Tekalp and R. Mehrotra, "Effective content representation for video", In Proc. IEEE International Conference Image Processing, Chicago, IL, Oct. 1998.
4. Y. Rui, T. Huang and S. Mehrotra, "Browsing and retrieving video content in a unified framework", In Proc. IEEE Workshop on Multimedia Signal Processing, Dec. 1998.
5. C. Saraceno and R. Leonardi, "Indexing audio-visual databases through a joint audio and video processing", International Journal of Imaging Systems and Technology, 9(5):320-331, Oct. 1998.
6. C. Saraceno and R. Leonardi, "Identification of story units in audio-visual sequences by joint audio and video processing", In Proc. International Conference on Image Processing 1998, Chicago, IL, U.S.A., Oct. 1998.
7. Text of ISO/IEC 15938-5 Final Commitee Draft - Information Technology - Multimedia Content Description Interface - Part 5 Multimedia Description Schemes, ISO/IEC JTC1/SC29/WG11 MPEG00/N3966, Singapore, SG, May. 2001.
8. MPEG-7 Heads of Delegation, "An Introduction to MPEG and the Standardization Process", ISO/IEC JTC1/SC29/WG11 N3962, MPEG01, Singapore, SG, March 2001.

[4] Dept. of Electrical Engineering, Columbia University, New York, NY 10027, US.
[5] Sony, San Jose, US.
[6] University at Buffalo, State University of New York, US.
[7] Cyber Assist Research Center, Tokyo 135-0064, JP.
[8] Kodak, US.
[9] University of Rochester, US.
[10] Sony, Tokyo, JP.

```
<Semantic id="Description example">
    <! - - Formal abstraction: usage AbstractionLevel = 1 - ->
    <AbstractionLevel dimension="1"/>
    <Label> <Name>Template example</Name> </Label>
        <SemanticBase xsi:type="AgentObjectType" id="Performer">
            <AbstractionLevel dimension="1"/>
            <Label><Name>Performer</Name></Label>
            <Relation xsi:type="ObjectEventRelationType" name="agentOf" target="To take"/>
        </SemanticBase>
        <SemanticBase xsi:type="ObjectType" id="Object">
            <AbstractionLevel dimension="1"/>
            <Label><Name>Object</Name></Label>
            <Relation xsi:type="ObjectEventRelationType" name="patientOf" target="To take"/>
        </SemanticBase>
        <SemanticBase xsi:type="EventType" id="To take">
            <AbstractionLevel dimension="1"/>
            <Label><Name>To take an object</Name></Label>
        </SemanticBase>
        <SemanticBase xsi:type="AgentObjectType" id="Customer">
            <AbstractionLevel dimension="1"/>
            <Label><Name>Who buys something</Name></Label>
            <Relation xsi:type="ObjectObjectRelationType" name="specializationOf"
              target="Performer"/>
            <Relation xsi:type="ObjectEventRelationType" name="agentOf" target="To buy"/>
            <Relation xsi:type="ObjectEventRelationType" name="agentOf" target="Payment"/>
        </SemanticBase>
        <SemanticBase xsi:type="AgentObjectType" id="Seller">
            <AbstractionLevel dimension="1"/>
            <Label><Name>Who sells something</Name></Label>
            <Relation xsi:type="ObjectEventRelationType" name="sourceOf" target="To buy"/>
            <Relation xsi:type="ObjectEventRelationType" name="beneficiaryOf" target="Payment"/>
        </SemanticBase>
        <SemanticBase xsi:type="EventType" id="To buy">
            <AbstractionLevel dimension="1"/>
            <Label><Name>To buy something</Name></Label>
            <Relation xsi:type="EventEventRelationType" name="specializationOf" target="To take"/>
        </SemanticBase>
        <SemanticBase xsi:type="ObjectType" id="Price">
            <AbstractionLevel dimension="1"/>
            <Label><Name>Price</Name></Label>
            <Relation xsi:type="ObjectEventRelationType" name="patientOf" target="Payment"/>
        </SemanticBase>
        <SemanticBase xsi:type="EventType" id="Payment">
            <AbstractionLevel dimension="1"/>
            <Label><Name>Payment</Name></Label>
            <Relation xsi:type="EventEventRelationType" name="entailmentOf" target="To buy"/>
        </SemanticBase>
</Semantic>
```

Table 1. Syntax of the example shown in figure 3.