Introducing a Unified Framework for Content Object Description

P. Daras[1], A. Axenopoulos[1], V. Darlagiannis[1], D. Tzovaras[1], X. Le Bourdon[2], L. Joyeux[3], A. Verroust-Blondet[3], V. Croce[4], T. Steiner[5], A. Massari[6], A. Camurri[6], S. Morin[7], A.D. Mezaour[7], L. Sutton[8] and S. Spiller[9].

[1]Centre for Research and Technology Hellas (CERTH/ITI), Thessaloniki, Greece; [2]JCP-Consult, France; [3]Institut National de Recherche en Informatique (INRIA), France; [4]Engineering Ingegneria Informatica S.p.a, Italy; [5]Google Ireland Limited, Ireland; [6]University of Genoa (UNIGE), Italy; [7]Exalead, Paris, France; [8]Accademia Nazionale di Santa Cecilia (ANSC), Rome, Italy;
[9] EasternGraphics Gmbh (EGR), Germany;

E-mail: daras@iti.gr

Abstract: In this paper, a novel framework for description of rich media content is introduced. Firstly, the concept of "Content Objects" is provided. Content Objects are rich media presentations, enclosing different types of media, along with real-world information and user-related information. These highly complex presentations require a suitable description scheme in order to be searched and retrieved by end users. Therefore a novel Rich Unified Content Description is analysed, which provides a uniform descriptor for all types of Content Objects irrespective of the underlying media and accompanying information.

Keywords: multimedia description, multimodal search and retrieval, Content Object

1 Introduction

Multimedia content, which is available over the Internet, is increasing at a faster pace than respective increase in computational power. In 2006, digital content, produced by either professional or amateur users, reached the capacity of 161 exabytes, while it is expected that by 2010 it will reach the capacity of 998 exabytes, inducing a 6 fold increase [IDC 2007]. In 2011, the amount of digital information that will be produced will be almost 10 times the content produced in 2006 [IDC 2007]. Due to the

widespread availability of digital recording devices, improved modelling tools, advanced scanning mechanisms as well as display and rendering devices, even over mobile environments, users are getting more empowered to live an immersive and unforgettable experience with latest-generation digital media. This growth in popularity of media is not accompanied by similar rapid development of media search technologies. The most popular media services on the Web are typically limited to textual search. However, over the last years, significant efforts have been devoted, mainly by the European research community, to achieving content-based search of images, videos and 3D models.

In order for this ever growing media content to be easily searched and retrieved by next generation content-based search engines, a framework for describing media in a standard format to ensure interoperability is required. Towards this direction, the MPEG-7 standard [Martinez 2002] offers a comprehensive set of multimedia description tools, which can be used by applications that enable quality access to content. MPEG-7 gives a generic framework that can support various applications, facilitating exchange and reuse of multimedia content across different application domains. Similarly, JPSearch [Dufaux 2007] aims to provide a standard for interoperability for image (JPEG, JPEG2000) search and retrieval systems. More specifically, the goal of JPSearch is to define the interfaces and protocols for data exchange between devices and systems.

Currently, media content is often the result of an off-line, cumbersome and lengthy creation process. This media is delivered to end-users for consumption as a finalised complete media presentation in the form of bit streams, followed by a play-out at the end-user's device. In the context of the Future Internet (FI), on the other hand, the concept of Content Objects [Zahariadis 2010] has been introduced to describe rich media experiences created as just-in-time composition of content that is easily located, synchronised, reused and composed. The availability of the constituent Content Objects and their spatial and temporal relationships, rather than an opaque stream of pixels and audio samples, opens up new opportunities for content creation and consumption. In this new era, where the volume and quality of transferred content rise sharply and more users evolve from mere consumers to active creators, new approaches for describing, searching and retrieving this rich media content are required.

With the objective to address the increasing demands of the FI, the EU-funded project I-SEARCH [Axenopoulos 2010] aims to provide a novel unified framework for multimodal content indexing, search and retrieval. The I-SEARCH framework will be able to handle specific types of multimedia and multimodal content (text, 2D image, sketch, video, 3D objects and audio) along with real-world and user-related information, which can be used as queries and retrieve any available relevant content of

any of the aforementioned types. The search engine which I-SEARCH is proposing will be highly user-centric in the sense that only the content of interest will be delivered to the end-users, satisfying their information needs and preferences. Being able to deal with all the aforementioned types of media, dynamic content, real-world and user-related information, this novel framework fits perfectly in the nature of the Content Objects as described in [Zahariadis 2010].

Multimodal search and retrieval has been already addressed by numerous commercial applications. Several "closed" industry standards are available today, mainly for mobile devices like Apple's iPhone or smart phones based on Google's Android operating system. Due to the small hard- or software keyboards on mobile devices, alternative input types are desirable. In Android devices, a "search by voice recognition" functionality is available (sound is transformed into a textual query). The same functionality is available in several applications for iPhone, e.g. Bing's, Yahoo!'s, and Google's own search applications. Google Goggles [Google] goes one step further by allowing for images and GPS location to be search queries. This can be used to search for e.g. text contained in images, landmarks or attractions in images, or product search. The returned search results are of multimodal form (map results, image results, video results, and obviously text results). While none of these standards are publicly available outside the scope of the particular applications, I-SEARCH, on the other hand, proposes an open solution that goes beyond what is commercially available today.

In this paper, the novel framework for description of rich media content that is introduced by I-SEARCH is described in detail. The Rich Unified Content Description (RUCoD) consists of a multi-layered structure, which will integrate intrinsic properties of the content, dynamic properties, non-verbal expressive, emotional and real-world descriptors. RUCoD will serve as a formal representation of Content Objects, which will be also clearly defined in the paper. The relations of RUCoD with other well-known standards for multimedia description, such as MPEG-7, will be also presented. Special focus will be on the implementation of RUCoD to the three use cases of the I-SEARCH project, namely: search for music content, furniture model retrieval and 3D object/avatar retrieval for games.

The rest of the paper is organized as follows: In Section 2, the concept of Content Object is described in detail. The specification of the RUCoD, which is introduced for the Content Object description, follows in Section 3. In Section 4, the implementation of RUCoD in three different use cases, realized within the scope of I-SEARCH, is described. In Section 5, a brief review of the well-known standards for multimedia description and use (MPEG-7, JPSearch, MPEG-21) is given, followed by a comparison with RUCoD. Finally, conclusions are drawn in Section 6.

2 Content Object

The concept of Content Objects (COs) has been introduced in [Zahariadis 2010]. More specifically, the following definition was given:

"*A Content Object is a polymorphic/holistic container, which may consist of media, rules, behaviour, relations and characteristics or any combination of the above*".

The definition given above is rather generic. In order to obtain a more concrete idea of CO, the following definition has been developed in the context of the I-SEARCH project and is presented for the first time in this paper. This definition was inspired by the I-SEARCH requirements, in order to address a broad range of applications related to multimodal search and retrieval as well as multimodal interaction.

"A Content Object is the representation of a specific instance of either a physical object or a physical entity (an entity that has physical existence, e.g. an earthquake) or an abstraction (a general concept formed by extracting common features from specific examples), an event or a concept, which might have multiple views (many images, videos, audio files, text, real-world and user-related information)."

According to the definition above, it can be inferred that an integral part of the CO is multimedia. A CO cannot exist without the existence of media items inside it. A CO can span from very simple media items (e.g. a single image or an audio file) to highly complex multimedia collections (e.g. a 3D object together with multiple 2D images and audio files) along with accompanying information. When a user refers to a CO, s/he directly refers to all of its constituting parts.

From a Future Internet (FI) perspective, the adoption of COs is expected to revolutionize access to digital content, since instead of sharing, searching and retrieving single media items, novel FI architectures can be appropriately designed to support exchange of COs. A significant step towards this goal has been made through multimodal search and retrieval, which is addressed by the I-SEARCH project.

2.1 Comparison with Relevant Concepts

Multimodal search and retrieval is a quite new area of research and only few approaches have been reported so far that deal with this problem.

In [Yang 2009], a framework for cross-media retrieval is presented, where the query example and the retrieved results can be of different media types. In order to realize multimodal search, the concept of *Multimedia Document* (MMD) is introduced. A MMD is defined as a set of co-occurring multimedia objects (e.g. images, audio and text) that are of different modalities but carry the same semantics. If two multimedia objects are in the same MMD, they can be regarded as context of each other.

Another approach that addresses the problem of multimodal search is presented in [Zhang 2006], where both intra- and inter-media correlations are learnt among multi-modality feature spaces in order to construct a semantic subspace containing multimedia objects of different modalities. Here, the concept of *Multimedia Bag* is introduced, which defines a container including text instances, image instances and audio instances that share the shame semantic concepts.

Both of the abovementioned methods define a novel rich multimedia representation (either Multimedia Document or Multimedia Bag), a container that integrates multiple modalities of the same semantics into one single object that can be searched and retrieved as a whole. Using the terminology of this paper, both Multimedia Document and Multimedia Bag provide good approximations of the CO, since they represent rich collections of multimedia with the same semantics. Based on the existing state-of-the-art methods in multimodal search, the I-SEARCH project aims to provide a more formal definition of the CO plus a novel framework to support efficient search and retrieval of COs.

Based on its former definition a Content Object may consist of media, rules, behaviour, relations, characteristics or any combination of the above. In the sequel, we will identify which of the above features can be addressed by the CO definition introduced in this paper:

- *Media:* it is the digital representation of anything that a human can perceive/experience with his/her senses and can be captured (through a specific device, such as camera, microphone, etc.) or created (using an authoring tool). As previously mentioned, multimedia is an integral part of the CO. What should be highlighted here is the multimodal nature of COs. To be more specific, COs do not consist of a single media item but they can be highly complex multimedia collections along with accompanying information. This information will be searched and retrieved as a whole, irrespective of the media type that is used as query, even in cases where the query modality is absent from the CO.
- *Rules:* can refer to the way an object is treated and manipulated by other objects or the environment (discovered, retrieved, casted,

adapted, delivered, transformed, and presented). In this paper, a novel framework for unified CO description is introduced (to be analysed in the following sections). This description scheme provides the rules on how these COs will be searched, retrieved, adapted, delivered and presented, using the search, retrieval and adaptive presentation framework provided by I-SEARCH.

- *Behaviour* can refer to the way the object affects other objects or the environment. Currently, behaviour is not explicitly supported in the RUCoD format. However, implicitly there are some functions such as Relevance Feedback, where the selected COs (positive or negative) can affect the ranking position of other COs. In this case, user behaviour (that can be modelled as a RUCoD query, with U-descriptors) can affect the results of COs (with respect to their user-related parts)
- *Relations:* refer to relations between a CO and other COs. In I-SEARCH relations are addressed in two ways. Firstly, relations between the different media items, which are constituting parts of the same CO, are directly established, since they are included within the same CO description file. Secondly, each CO may contain links to other COs that are somehow related to each other.
- *Characteristics:* these meaningfully describe the CO and allow retrieval of its related COs. As will be described below, I-SEARCH introduces a novel framework for unified description of COs, supporting efficient multimodal search and retrieval of them.

2.2 Content Object Creation

Since no detailed description currently exists to address the multimodal nature of a CO, a framework for unified description of COs has been introduced within I-SEARCH. The Rich Unified Content Description (RUCoD) is a data representation of a CO consisting of descriptions (features/characteristics) of various multimedia types that are somehow associated to each other. A detailed description of the RUCoD specification is available at the following Section.

What has not yet been explained is how a CO is created. As previously defined, a CO may consist of several multimedia types, user-related information and real-world information. This information can be created, generated or captured by using a variety of hardware and software tools, input devices, sensors, etc. In order for this input to be mixed into a uniform representation, an Authoring Tool is required. A conceptual diagram of it is depicted in Figure 1.

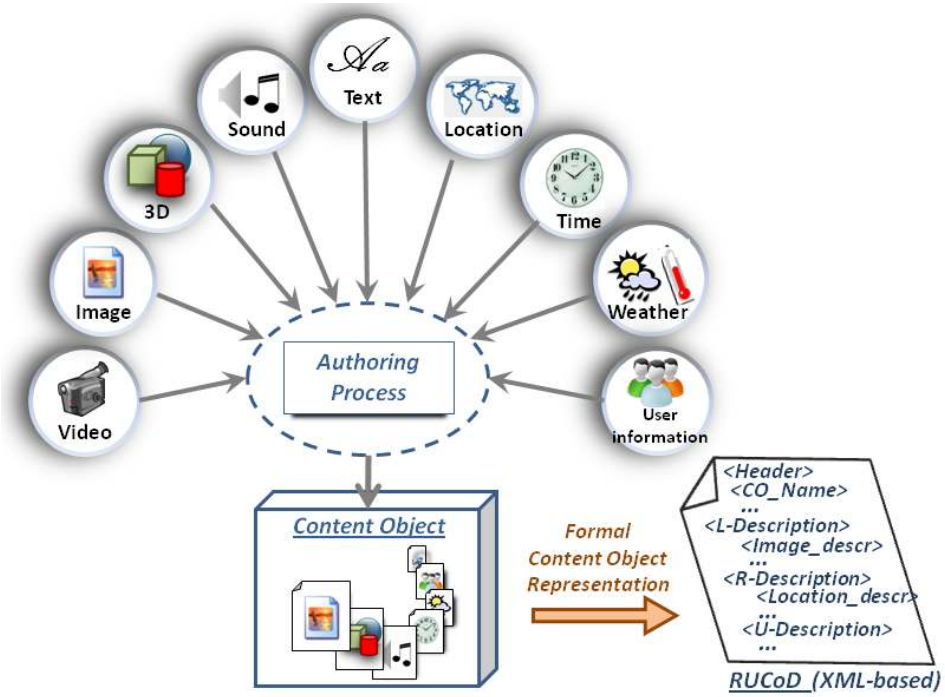

Figure 1: A conceptual diagram of an Authoring Tool for Content Objects

The Authoring Tool will take as input all different types of media items, real-world information (location, weather, time, etc.) and user-related information (emotional/expressive characteristics) and will produce a rich media representation, a Content Object. The formal description of a CO is its RUCoD file, which is an XML-based document specifying descriptors for all the above input types.

Through an appropriate user interface of the Authoring Tool, the user will be able to add manually all related information to create the CO. A special functionality of the tool is that it will assist users to easily add links to other relevant COs. Links are underlined as the representation of relations with other COs. These links gather the intention of the User to relate other COs to the one he is defining. They represent the relation among concepts that the User is expressing. With an appropriate search interface, user will be able to search for similar COs and create relations among COs.

3 RUCoD Specification

In this section, the specification of RUCoD is introduced. RUCoD will serve as a generic multimedia content descriptor, enhanced with real-world information, expressive and emotional descriptions, in order to facilitate the retrieval of different types of media irrespective of the query format.

The goal is to enable the development of very heterogeneous applications, ranging from pure search and retrieval to personalized, user interaction specific and/or context-aware search. Therefore, this unified approach is proposed, where the actual metadata and the real world and user-related parts reside in the same format.

The general form of the RUCoD structure is given in Figure 2 below:

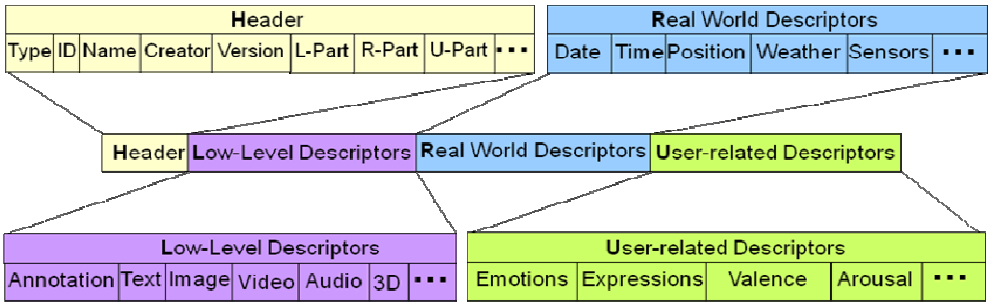

Figure 2: The RUCoD General format

The overall structure of a RUCoD description file, in XML, is given in Figure 3.

```
<?xml version="1.0" encoding="UTF-8"?>
<RUCoD xmlns="http://www.isearch-project.eu/isearch/RUCoD">
    <Header>
        <ContentObjectType>Physical Entity</ContentObjectType>
        <ContentObjectName xml:lang="en-US">My Bulldog Barking</ContentObjectName>
        <ContentObjectID>3577B5EF-523F-4946-9734-C974CEA6C646</ContentObjectID>
        <ContentObjectVersion xsi:type="int">2</ContentObjectVersion>
        <ContentObjectCreationInformation>...</ContentObjectCreationInformation>
        <ContentObjectTypes>
            <MultimediaContent xsi:type="Text">
                <FreeText>It is the image, video and 3D representation of my bulldog...
                </FreeText>
            </MultimediaContent>
            <MultimediaContent xsi:type="Object3D"> ...</MultimediaContent>
            ...
            <RealWorldInfo>...  </RealWorldInfo>
            <UserInfo> ...  </UserInfo>
        </ContentObjectTypes>
        <Links> ... </Links>
    </Header>
    <Description xsi:type="RUCoDDescriptor">
        <L_Descriptor xsi:type="TextType">
            <TextDescription>...</TextDescription>
        </L_Descriptor>
        <L_Descriptor xsi:type="Object3D">...   </L_Descriptor>
        <L_Descriptor xsi:type="SoundType">...  </L_Descriptor>

        <R_Descriptor>
            <RealWorldInfo xsi:type="ContextType">... </RealWorldInfo>
            ...
        </R_Descriptor>

        <U_Descriptor xsi:type="UserType">
            <MediaName>BulldogSound</MediaName>
            <UserDescription xsi:type="Valence">... </UserDescription>
        </U_Descriptor>
    </Description>
</RUCoD>
```

Figure 3: The overall structure of a RUCoD file

In the example above, the RUCoD description corresponds to the Content Object entitled "My Barking Bulldog". RUCoD consists of the following main parts:

- *Header*: includes general information about the Content Object, such as the type, name, ID and creation information. Moreover, the RUCoD Header encloses some general information about the different media (3D, images, sounds, videos, text) and accompanying information (real world data, user-related cues) that constitute the Content Object.
- *Description*: it is the core part of the RUCoD including detailed information about the corresponding media and contextual information (real world, user-related). It consists of a) the L-Descriptors part, where the low-level descriptors, extracted from each separate media (3D, images, sounds, videos, text), are presented and

b) the R-Descriptors part, which maintains descriptors extracted from real-world sensors, representing time, weather, location, etc. and c) the U-Descriptors part, where descriptors related to the user behaviour (emotions, expressions) are stored.

The interrelations among the several parts of the RUCoD are given in the following conceptual model (Figure 4). At the centre of the model is the CO, which has a one-to-one relation with the RUCoD. A CO has one-to-many relations with the constituting media items, the user-related information and the real-world information. All the above media and accompanying information are connected with their corresponding descriptors in a one-to-one relation. Finally, each media item may produce one or more artefacts that can be used for descriptor extraction.

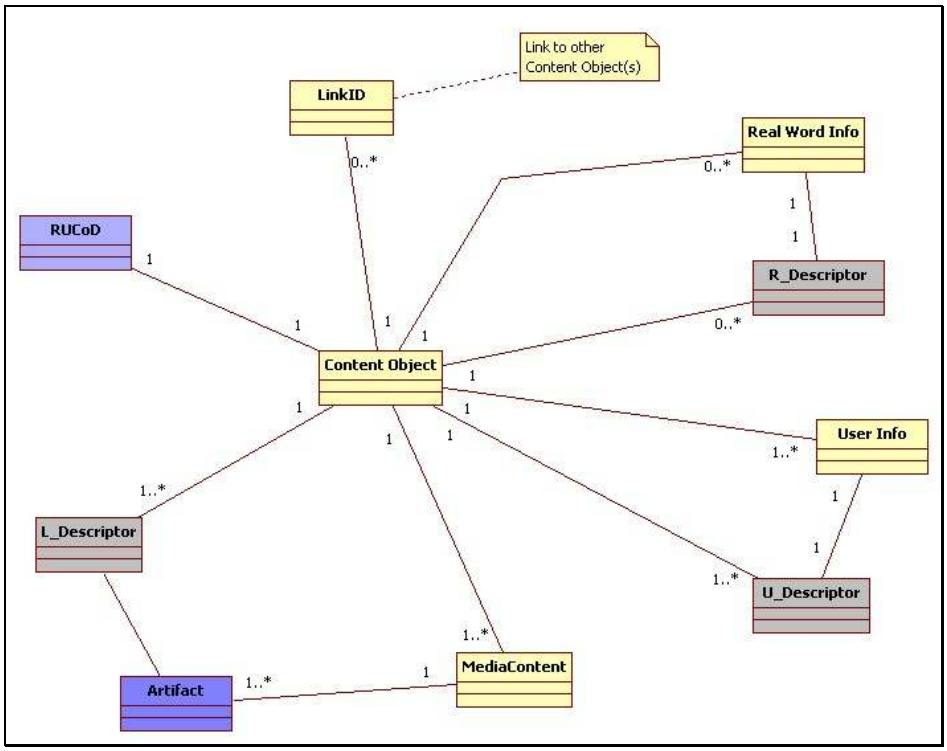

Figure 4: Conceptual Model presenting CO and the relations with its constituting elements.

3.1 RUCoD Header

In Figure 5, a more detailed view of the RUCoD Header is given. The first fields provide information about the Content Object, while the next part is

focused on the several constituting parts of the Content Object: different media, real-world information, user-related information.

```
<?xml version="1.0" encoding="UTF-8"?>
<RUCoD xmlns="http://www.isearch-project.eu/isearch/RUCoD" >
    <Header>
        <ContentObjectType>Physical Entity</ContentObjectType>
        <ContentObjectName xml:lang="en-US">My Bulldog Barking</ContentObjectName>
        <ContentObjectID>3577B5EF-523F-4946-9734-C974CEA6C646</ContentObjectID>
        <ContentObjectVersion xsi:type="int">2</ContentObjectVersion>
            <Creator>...</Creator>
        </ContentObjectCreationInformation>
        <ContentObjectTypes>
            <MultimediaContent xsi:type="Text">
                <FreeText>It is the image, video and 3D  ... </FreeText>
            </MultimediaContent>
            <MultimediaContent xsi:type="Object3D">
                <MediaName>Bulldog</MediaName>
                <FileFormat>x-world/x-vrml</FileFormat>
                <MediaLocator>
                    <MediaUri>http://3d-test.iti.gr:8080/3d-test/ContentObject/Bulldog.wrl</MediaUri>
                    <MediaPreview>http://3d-test.iti.gr:8080/3d-test/ContentObject/BulldogWRL.jpg
                    </MediaPreview>
                </MediaLocator>
            </MultimediaContent>
            <MultimediaContent xsi:type="SoundType">
                <MediaName>BulldogSound</MediaName>
                <FileFormat>audio/x-wav</FileFormat>...
            </MultimediaContent>
            <RealWorldInfo>
                <RWContextSliceName>BulldogContext</RWContextSliceName>
                <RWContextTypes>Time, Location, Temperature</RWContextTypes>
                <RWDescriptorsFormat>W3C/ISO 8601, GML, Celsius</RWDescriptorsFormat>
            </RealWorldInfo>
            <UserInfo> <!--  e.g. We could manually add some emotional cues, such as "angry".  -->
                <UserInfoName>BulldogEmotion</UserInfoName>
                <emotion>... </emotion>
            </UserInfo>
        </ContentObjectTypes>
        <Links>
            <linkID>3577B5EF-523F-4946-9734-C974CEA6C333</linkID> ...
        </Links>
    </Header>
```

Figure 5: The RUCoD Header

Content Object-related information comprises the following parts:

- *ContentObjectType*: it is the term used by wordnet [Wordnet] to determine the overall category the Content Object belongs to. This takes attributes, such as "*Physical Entity*", "*Abstraction*" or "*Abstract Entity*".
- *ContentObjectName*: it is the name given by the content object creator.
- *ContentObjectID*: an identifier, which is unique for a specific Content Object. In the example given the GUID format is used.
- *ContentObjectVersion*: an integer incremented on every change of the content object or RUCoD.

- *ContentObjectCreationInformation*: the creator (name) of the specific Content object.

The *ContentObjectTypes* field refers to the different media, real-world information or user-related information and includes the following:

- *MultimediaContent*: this container includes information of a specific multimedia object. It may also have the following subparts:
  - *MediaName:* A unique name within the RUCoD representing the specific media.
  - *FileFormat*: The MIME type of the file format of the stored media.
  - *MediaLocator*: The URI of the location where the specific media is stored, together with the optional location of the media item to be used as preview.
  - *MediaCreationInformation*: information about the creator of the media (if it is different from the Content Object creator).
- *RealWorldInfo*:
  - *RWContextSliceName:* A unique name within the RUCoD representing the specific ContextSlice.
  - *RWContextTypes:* A list of R-Descriptors for a particular ContextSlice that are used to describe the context of the media in this RUCoD record.
  - *RWDescriptorsFormat:* A list of formats that is used to describe each Real-World sensor descriptor for a particular ContextSlice.
- *UserInfo*: the emotional state associated with the content object, as specified by the author of the Content Object, stored as a fragment of EmotionML.
- *Links*: the IDs of Content Objects that are linked to the current Content Object (e.g. one linked Content Object could be the one representing the house of the "bulldog" Content Object).

3.2 L-Descriptor

A snapshot of a part of the RUCoD L-Descriptor is given in Figure 6. In the example, the low-level descriptors of one of the 3D objects of the "My barking bulldog" Content Object are specified. *L_Descriptor* may include the following fields:

- *MediaName*: this field should be identical with the *MediaName* field of the same media object defined at the RUCoD Header. It is essential in order to map the original media file with the corresponding low-level descriptors.
- *Shape3DDescription*: it is a container, enclosing descriptors of a specific 3D object. Here the type of the low-level descriptor as well as the matching method are defined.
- *GlobalShape*: defines a set of parameters of a 3D shape descriptor, such as the dimension of the descriptor vector, the type of the descriptor (text, numerical, integer) and the size in bytes of each descriptor. This information is required for parsing the descriptor file (which may be given in binary format).
- *ImageDescription:* it is a container, enclosing descriptors of a specific image. These may include Edge Orientation Histogram (Eoh_32), Probability weighted histogram (Probrgb_6_2), Laplacian weighted histogram (Laplrgb_6), HSV standard histogram (Histohsv_std) or SIFT local descriptors (sift_desc).
- *VideoDescription:* Consists of a set of visual objects present in several key-frames of the video and, for each object, a list of information on the key-frame images containing this visual object (time code of the image, position of the visual object inside the image).
- *AudioDescription*: defines a set of parameters for the audio signal, such as its fingerprint (for computing audio similarity), its rhythmic pattern and melodic profile. Examples of such low-level descriptors are the MPEG-7 descriptors (e.g. FundamentalFrequency, Power) as well as others more targeted at Music Information Retrieval tasks (e.g. Mel-Frequency Cepstral Coefficients, Inter Onset Intervals, Statistical Sound Description).
- *DescriptorLocator*: The URI of the location where the low-level descriptor of the specific media is stored, or, in alternative, the raw values expressed as space-separated numbers.

```
<Description xsi:type="RUCoDDescriptor">
    ...
    <L_Descriptor xsi:type="Object3D">
        <MediaName>Bulldog</MediaName>
        <Shape3DDescription xsi:type="CompactMultiViewDescriptor" matching="MultiViewL2">
            <GlobalShape totalNumOfViews="18" totalNumOfDescriptors="212" descriptorType="xsd:int" descriptorSize="4
                <Store xsi:type="Binary">
                    <DescriptorLocator>
                        <DescriptorUri>
                        http://3d-test.iti.gr:8080/3d-test/ContentObject/Bulldog_CMVD.descr
                        </DescriptorUri>
                    </DescriptorLocator>
                </Store>
            </GlobalShape>
        </Shape3DDescription>
        <Shape3DDescription xsi:type="DSR" matching="L1">
            <GlobalShape totalNumOfDescriptors="472" descriptorType="xsd:int" descriptorSize="4">
                <Store xsi:type="Binary">
                    <DescriptorLocator>
                        <DescriptorUri>
                        http://3d-test.iti.gr:8080/3d-test/ContentObject/Bulldog_DSR.descr
                        </DescriptorUri>
                    </DescriptorLocator>
                </Store>
            </GlobalShape>
        </Shape3DDescription>
    </L_Descriptor>
    ...
```

Figure 6: The L-Descriptor (3D descriptor) Part of RUCoD

- *TextDescription*: a container, enclosing descriptors of a specific text object.
- *RelatedSemanticConcepts*: a container for related semantic concepts for the current RUCoD data set identified by a URI like http://dbpedia.org/resource/Bulldog.
- *WordNetInheritedHypernym*: identifies the current RUCoD data set's inherited hypernym as defined in WordNet [Wordnet].
- *Language*: identifies the detected language of the text object.
- *NamedEntities*: a container, enclosing NamedEntity objects.
- *NamedEntity*: describes a matched named entity inside the text object. It has an attribute *type* which can be either *People*, *Organization* or *Event*.
- *RelatedTerms*: a container, enclosing RelatedTerm objects.
- *RelatedTerm*: describes a related term to the text object.
- *Sentiment*: describes the sentiment detected inside a text object. It has a *confidence* attribute that describes the confidence in sentiment analysis. Values range from *0* (very bad) to *5* (very good).
- *Category*: describes the detected category of the text, using for instance the DMOZ classification scheme [DMOZ].

- *Tokens*: a container, enclosing Token objects.
- *Token*: describes a token inside the text. It has an attribute *value* which is the value of the token.
- *Normalized*: describes the normalized version of the token.
- *Stem*: describes the stemmed version of the token.
- *PartOfSpeech*: describes the part of speech of the token in the context of the text segment
- *Expansion*: describes an expanded version of the token with an add-hoc ontology.

3.3 R-Descriptor

The real-world descriptor is based on the context as defined by Dey and Abowd [Dey 1999], "*information that can be used to characterize the situation of an entity"*, where an entity is a Content Object. However, the different dimensions of the context are not always orthogonal. In other words, semantic links can append between two dimensions of the context. For example, in a video, the position of the camera may change (semantic link between the location and the time), and the weather may change for different locations (semantic link between the weather and the location). To be able to define such complex context, we define the concept of Context Slices.

RUCoD *R-Descriptors* are grouped in ContextSlices where each descriptor provides information on a particular aspect of the Context (i.e., time, position, temperature, etc.). Each RUCoD record may contain multiple ContextSlices. Each ContextSlice is composed by a non-empty set of R-Descriptors. A ContextSlice may refer to multiple media and multiple ContextSlices may be used to set the context for a piece of media. The following picture describes the potential relationships between ContextSlices, media objects and R-Descriptors within a RUCoD record.

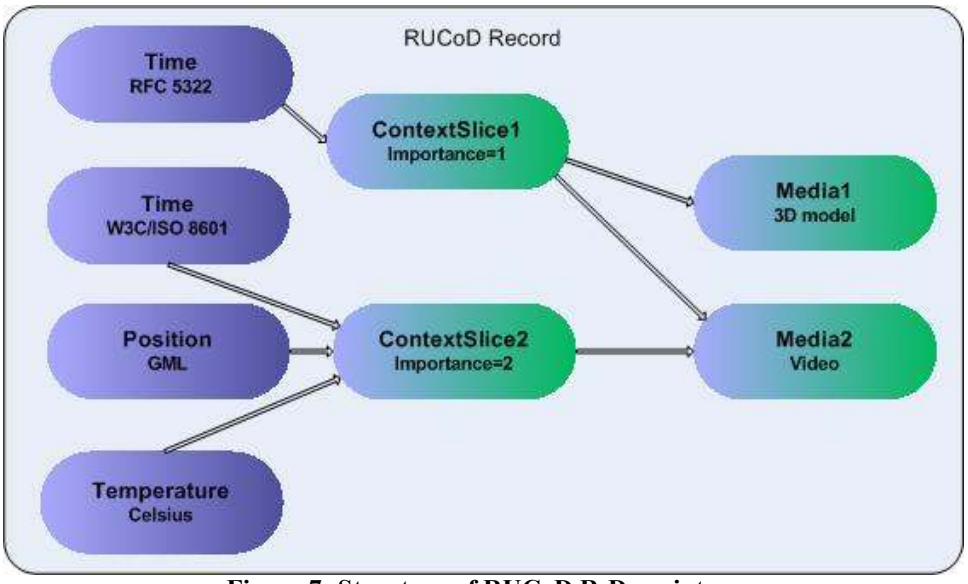

Figure 7: Structure of RUCoD R-Descriptors

In Figure 7, *ContextSlice1* describes the context for both video and 3D media item, while for the video item, *ContextSlice2* completes the relevant Real-World information. The importance of *ContextSlice2* is higher and it overrides any similar information that may be provided in *ContextSlice1*. For each *ContextSlice* the following information shall be provided to effectively match it with the relevant media objects and set their context.

- *MediaNamesList*: this field includes all the media items to which the specific R-Descriptors are mapped. The names used in *MediaNamesList* must be the same as the names defined in the *MediaName* fields of the corresponding media items, located in the RUCoD Header. It is essential in order to match the original media file to the corresponding real-world descriptors.
- *Importance*: this field should have a unique value that will allow dealing effectively with conflicts between different ContextSlices referring to the same media object.

One ContextSlice should also contain one or more context descriptors:

- *DateTime*: defines the start date of the CO, and its length if relevant.
- *SubjectPosition:* defines the location of the main subject of the CO. The format used in this descriptor is based on the GML standard, thus it may be a simple point or a complex shape.

- *ViewerPosition:* defines the location and the view (tilt, roll…) of the viewer. The format is based on the GML standard, thus it may also be a simple point or a complex shape.
- *Weather:* defines the weather (temperature, wind, condition…) relevant for the media.
- *PhysicalTags:* a list of the raw content of physical tags (RFID, 2D barcodes…)

For each sensor (aspect of the Context), the following information is required to effectively interpret the Real World data.

- *SensorFormat:* The format of the sensor information (e.g. string, integer, XML).
- *SensorAccuracy:* The accuracy of the specific sensor. It may be undefined if not applicable.
- *SampleUnits*: The units of the measured sensor samples if applicable.

3.4 U-Descriptor

U_Descriptor may include the following fields (Figure 8):

- *MediaName*: this field must be identical with the *MediaName* field of the same media object defined at the RUCoD Header. It is essential in order to map the original media file with the corresponding user-related descriptors.
- *UserDescription*: defines a set of parameters that encapsulate the emotional information extracted by the audio or video signal. It includes descriptors that define:
    - Position in the *Valence-Arousal 2D space*: it is an affective space whose regions are associated with the emotional states of “angry”, “calm”, “happy” and “sad”.
    - Position in the *Kinematics-Energy sensorial space*: it is useful to describe how gestures are performed (light/heavy vs. hard/soft). Both descriptors can be stored as the complete trajectory (i.e. one coordinate per analyzed frame) and/or an average value.

```
<U_Descriptor xsi:type="UserType">
    <MediaName>BulldogSound</MediaName>
    <UserDescription xsi:type="Valence">
        <LowLevelDescriptor totalNumOfDescriptors="20" descriptorType="xsd:float" descriptorSize="20">
        0.5 0.6 0.61 0.51 0.45 0.5 0.6 0.61 0.51 0.45 0.5 0.6 0.61 0.51 0.45 0.5 0.6 0.61 0.51 0.45
        </LowLevelDescriptor>
    </UserDescription>
    <UserDescription xsi:type="Arousal">
        <LowLevelDescriptor totalNumOfDescriptors="20" descriptorType="xsd:float" descriptorSize="20">
        0.51 0.45 0.5 0.6 0.61 0.51 0.45 0.5 0.6 0.61 0.51 0.45 0.5 0.6 0.61 0.51 0.45 0.5 0.6 0.61
        </LowLevelDescriptor>
    </UserDescription>
    <UserDescription xsi:type="AvgValenceArousal" matching="L2Distance">
        <LowLevelDescriptor totalNumOfDescriptors="2" descriptorType="xsd:float" descriptorSize="1 1">
        0.5 0.6
        </LowLevelDescriptor>
    </UserDescription>
</U_Descriptor>
<!-- computed from the video file -->
<U_Descriptor xsi:type="UserType">
    <MediaName>BulldogVideo</MediaName>
    <UserDescription xsi:type="AvgValenceArousal" matching="L2Distance">
        <LowLevelDescriptor totalNumOfDescriptors="2" descriptorType="xsd:float" descriptorSize="1 1">
        -0.5 0.6
        </LowLevelDescriptor>
    </UserDescription>
</U_Descriptor>

</Description>
</RUCoD>
```

Figure 8: The U-Descriptor Part of RUCoD

4 Application to I-SEARCH Use Cases

This section provides an overview of the three different scenarios supported by the I-SEARCH project. This is done in order to identify the specific needs of each use case, in terms of multimedia content description, and the subsequent modifications to their RUCoD descriptions.

4.1 Music Retrieval Scenario

This scenario involves search and retrieval of audio recordings, going beyond what the current search-by-example engines provide. The I-SEARCH platform will provide several usual queries (e.g. finding a song by providing a fragment, by singing it or by tapping its rhythm) but it will also allow taking into consideration the affective dimension.

The audio archive that will be exploited is the archive of ethnomusicology located at the Accademia Nazionale di Santa Cecilia in Rome, as well as recordings of classical concerts; each recording will be stored as a CO in the database, along with the following fields of the RUCoD specification:

- *Header*: almost all of the *Header* parts described in the previous section are present in this specific use case.

  - *MultimediaContent* will be used to point to the location of the *SoundType* object.
  - *Links* will be used to point to other Cos, e.g. other material (recordings, pictures, videos) gathered by the same researcher in the same place/region.
- *L_Descriptor*: three *AudioDescription* elements will be used to store the low-level information needed to perform the search-by-example queries (in a first iteration StatisticalSoundDescription, InterOnsetIntervals, FundamentalFrequency).
- *R-Descriptor*: Here the following RUCoD parts are identified: *DateTime*, *SubjectPosition*
- *U_Descriptor*: in this part of the RUCoD, the platform will store the UserDescription elements to support locating audio files sharing the same affective features expressed by the user performing the query (for instance, the location in the Valence-Arousal space).

4.2 Furniture Model Retrieval Scenario

This scenario involves search and retrieval of 3D Furniture Models. The scenario is applied when the end-user aims to search for a particular article of the office furniture industry, such as a chair, a desk or some container. The input for the search can be an image (e. g. photo, sketch, rendering of a 3D model), some text (e.g. words, phrases, letters) or a 3D object (e. g. CAD design of a piece of furniture).

For this specific use case, COs are basically pieces of Furniture. With respect to multimedia content, they consist of 3D models (in many available file formats) and/or 2D images (which represent several views of the 3D object). COs may also contain textual information, such as colour, style, material, cost, classification, manufacturer information, intended usage, as well as real-world information (manufacture date, e.g. old modern, not older than 10 years). Based on these facts, the following modifications to the initial RUCoD specification are foreseen:

- *Header*: almost all of the *Header* parts described in the previous section are present in this specific use case.
- *L-Descriptor*: Here the following RUCoD parts are identified: *MediaName, Shape3DDescription, GlobalShape, ImageDescription, DescriptorLocator, TextDescription, NamedEntities, RelatedTerms, Sentiment, Category, Tokens, Normalized* and *Stem.* More specifically:

- *NamedEntity*: the attribute *type* can take several values, such as *Cost, Components, Manufacturer, Dealers, Geometry, IntendedUsage*.

- *R-Descriptor*: Here the following RUCoD parts are identified: *DateTime*, *SubjectPosition*

A snapshot of the RUCoD XML file for a CO entitled "My Chair" is depicted in Figure 9. Focus is laid on the textual description part, where new items relevant to furniture retrieval scenario are introduced (material, geometry, colour, manufacturer, dealer, usage).

```
<Description xsi:type="RUCoDDescriptor">
    <L_Descriptor xsi:type=Text>
        <TextDescription>
            <Language confidence="0.895">en</Language>
            <NamedEntities>
                <NamedEntity Type="Colors">
                    <Color>black</Color>
                    <Color>white</Color>
                    <Color>multicolored</Color>
                </NamedEntity>
                <NamedEntity Type="Style">modern</NamedEntity>
                <NamedEntity Type="Materials">
                    <Combination>
                        <Material>leather</Material>
                        <Material>satin stainless steel</Material>
                    </Combination>
                </NamedEntity>
                <NamedEntity Type="Cost">standard</NamedEntity>
                <NamedEntity Type="Classification">chair</NamedEntity>
                <NamedEntity Type="Manufacturer">3577B5EF-523F-4946-9734-C974CEA6C333</NamedEntity>
                <NamedEntity Type="Dealers"> <Dealer>3577B5EF-523F-4946-9734-C974CEA6C444</Dealer>
                </NamedEntity>
                <NamedEntity Type="Geometry">
                    <Width>55</Width>
                    <Hight>100</Hight>
                    <Depth>60</Depth>
                </NamedEntity>
                <NamedEntity Type="IntendedUsage">Conference room</NamedEntity>
            </NamedEntities>
            <Sentiment confidence="0.754">5</Sentiment>
            <Category>Shopping/Furniture/Office/Chairs
            </Category>
            <Tokens>
                <Token value="chair">
                    <Normalized>chair</Normalized>
                    <Stem>chair</Stem>
                    <PartOfSpeech>NOUN</PartOfSpeech>
                    <Expansion source="wordnet">furniture</Expansion>
                </Token>
            </Tokens>
        </TextDescription>
    </L_Descriptor>
```

Figure 9: Snapshot of the <TextDescription> part of of RUCoD for the furniture retrieval use case.

4.3 Game Model Retrieval Scenario

This scenario consists of two use cases. In the first, game developers, game designers/modders will be able to search and retrieve 3D objects, which will assist them in modelling 3D worlds and objects. In the second, non-professional users (gamers) will be able to search for 3D game avatars.

Regarding the first use case, the searchable objects can be physical entities (animals, buildings, plants, vehicles, etc.) than can be used to create the 3D scene of a game. The RUCoD format in this case shares many similarities with the one presented in Section 3. Therefore, it can be easily a subset of the latter. The only difference is that in COs of this use case the 3D media item is a mandatory field.

For the second use case, a non-professional user may begin with a video sequence of a game, select a visual object extracted from this video representing an avatar and search a 3D object having a similar 2D view. In fact, the query will be a subset of an image that corresponds to a visual object extracted from this video. Here, the specification of the 3D and video content is also quite similar to the case described in Section 3, so it will result in a similar RUCoD, as well.

5 Comparison with Existing Standards

5.1 The MPEG-7 Standard

MPEG-7 has been standardized in 2001 (ISO/IEC 15398), aiming to facilitate exchange and reuse of multimedia content across different application domains and to ensure interoperability among them. It provides rich multimedia content description tools for applications ranging from content management, organization, navigation, and automated processing.

The four major MPEG-7 building blocks, which are required to build an MPEG-7 description and deploy it, are given below:

- *Descriptor*, which defines the syntax and the semantics of a Feature representation. A Feature is a distinctive characteristic of data.
- *Description Scheme,* which describes the structure and semantics of the relationships between Descriptors.
- *Description Definition Language,* which is a formal language allowing creation of new Description Schemes and Descriptors, as well as extension and modification of existing Description Schemes.

- *Systems Tools,* which support multiplexing of descriptions, synchronization of descriptions with content, delivery mechanisms, and coded representations for efficient storage and transmission.

To create MPEG-7 descriptions of any multimedia content, the first requirement is to build a wrapper for the description using the *Schema Tools*. The different description tools for the media content use generic *basic elements*, which are the XML-based MPEG-7 bricks.

RUCoD specification has borrowed several elements from the MPEG-7 standard, especially those related to description of multimedia items. Indicatively, we present the following MPEG-7 elements that are also exploited in RUCoD specification:

- *MediaLocator* and *MediaUri* are used to describe the link to a specific media item.
- *Creator* is used for description of the author of a media item.
- *Annotation* as a part of RUCoD represents textual information of a media item or CO.
- *Image/Video/Audio Descriptors* are used for the low-level descriptions of the separate media items within a CO.
- *Segment* is used to describe a temporal video segment.

Moreover, the following RUCoD elements stem from existing MPEG-7 elements, after appropriate modifications in order to fit to the nature of COs:

- *ContentObjectName, ContentObjectCreationInformation* instead of *name* and *CreationInformation* to represent the name and creators of COs.
- *TextDescription, Shape3DDescription, ImageDescription* and *VideoDescription,* similar to MPEG-7 *ContentDescription* to distinguish between the descriptors of different modalities inside the same RUCoD.

Finally, numerous new elements are introduced within the RUCoD framework to support the emerging needs of multimodal search and address several types of information:

- With respect to low-level descriptor extraction for media items, novel descriptors are introduced. As an example, for 3D content description, new state-of-the-art descriptors are introduced, which achieve higher

retrieval performance than those included in MPEG-7. Similarly, new descriptors are introduced for image, video and audio content.

- The low-level description of media items is also accompanied by specification of the matching scheme for each descriptor. In this case, the description scheme does not leave the responsibility for choosing the appropriate matching method to the search engine.
- New types of information describing the COs are introduced, such as real-world descriptors and user-related descriptors. These enrich the CO description and improve the retrieval performance, by introducing new querying capabilities.

5.2 The JPSearch Standard

The goal of JPSearch [Dufaux 2007] is to build a standard for interoperability among image search and retrieval systems. Search and retrieval functionality for images is provided by many systems in the web, however, in a way that tightly couples many components of the search process. JPSearch is designed in a way that decouples the components of image search and provides a standard interface between these components.

More specifically, JPSearch aims at defining the interfaces and protocols for data exchange between devices and systems, while restricting as little as possible how those devices, systems or components perform their task. This is achieved by the provision of a common query language to search easily across multiple repositories with the same search semantics, facilitating the use and reuse of profiles and ontologies.

Similar to JPSEARCH, RUCoD aims to ensure interoperability in media search, by proposing a standard description scheme. One of their differences is that RUCoD specification is focused on the description of COs, while JPSEARCH standardises the entire search and retrieval framework, including query format, schema and ontology, metadata embedded in image data, data interchange format between repositories. On the other hand, RUCoD addresses a broad range of media (apart from images), real-world and user-related information, which makes it appropriate for various application domains.

5.3 The MPEG-21 Standard

The MPEG-21 standard [Burnett 2004] was ratified in 2004. It aims to provide a framework for content creation and sharing which is usable by

all players of the digital content scenario from creators and distributors to consumers. The standard covers the entire content production and delivery "food-chain" with interoperability and automation in mind. In such vision the two main concepts in MPEG-21 are *Digital Items* and *Users*. Within the framework the latter continuously interact with- and manipulate the former.

A Digital Item (DI) is defined as the basic entity within the framework: it is a combination of resources, metadata and structure. Resources are the assets within the item (i.e. the actual content, possibly remote); metadata provides information about the Digital Item per se or the single resources within it; structure holds data about the relationships between the components of the Digital Item.

Users are actors who interact with the Digital Items, possibly in relationship with other Users. The framework is agnostic about Users' roles: this means that anyone who uses Digital Items (be it a content owner, provider, consumer, etc.) is a User. Interaction and manipulation by Users of the content are regulated within the framework by Rights Management mechanisms regarding what Users can and cannot do with the Digital Items and especially the resources they hold.

Technically the MPEG-21 *Digital Item* is defined by the Digital Item Declaration (DID). This is represented in the *Digital Item Declaration Language* (DIDL) which is an XML Schema defined by the standard. Describing all the components of the DID is out of the scope of the current work: in this context it is important to highlight the concepts of *container* and *item* within the *Digital Item*. The former is a structure which allows the grouping of both items and/or other containers. The latter is a grouping of items and/or components, which in turn are resources with a set of descriptors (i.e. metadata). Essentially this implies that the MPEG-21 model enables the possibility of nested items within items (defined as compilations) thus hierarchical content structures. A *Resource* (which could possibly be a physical object) is a uniquely identifiable asset within the item.

An interesting feature of MPEG-21 is the presence within the DID of *selections, asserts,* and *predicates* (all of which are in true, false or undecided states) which enable choices to be made on the objects. MPEG-21 is also concerned with Rights and Intellectual Property management issues which are not particularly relevant to our context.

The MPEG-21 *Digital Adaptation* tries to tackle the issue of accessing media "any time and anywhere" [Timmerer 2008]. This implies the adaptability of Digital Items and their resources to different devices, networks etc. Essentially the DI goes through an Adaptation Engine which transforms it according to target environment in terms of descriptions which can be (using the standards terminology): *terminal capabilities*,

network characteristics, user characteristics, natural environment characteristics. Digital Item Adaptation also takes care of metadata adaptation in terms of content change which reflects the metadata and metadata scaling and filtering.

The CO presented above is similar to the MPEG-21 DI. In fact both share a multimodal approach to media which can be of any type. Both MPEG-21 and RUCoD allow for rich metadata to be created and attached to objects. The *L-Descriptors* and *R-Descriptors* are a specific feature of the RUCoD. This doesn't mean that similar metadata couldn't be added to MPEG-21 DI, yet the RUCoD is particularly targeted at indexing, sharing, search and retrieval. In this direction the model doesn't directly address or enforce adaptation, although adaptation can be easily implemented thanks to the provision of elements such as the *FileFormat* (in the example above we have wave file for audio which could be converted to a lighter format for mobile appliances or low-bandwidth networks). Additionally the Real World descriptors allow to directly gather information about the *natural environment* of the user enabling rich, immersive user experiences [Zahariadis 2010].

One of the main differences between the MPEG-21 Digital Item model and the Content Object underlying to the RUCoD is that while MPEG-21 DI can in turn be hierarchical, a CO cannot contain another CO. This apparent 'limitation' in the model actually fits in the idea of overcoming the traditional strict nested object model allowing for a more open one where objects live in a smooth “universe” in which the user is the main actor. An example is the possibility of the *RelatedSemanticConcepts* field which enables connecting CO in a non-hierarchic manner thus enabling for more flexible in a more user-centric manner.

6 Conclusions

In this paper, an attempt to define the concept of Content Objects was made. Content Objects (COs) are rich media presentations, which enclose at the same time different types of media items related to the same physical entity, event or concept. Apart from media items, real-world and user-related information can also be identified within a CO. This definition was introduced to address the requirements of the EU funded project I-SEARCH with respect to multimodal search and retrieval and shares common features with the CO concept given by the User Centric Media Cluster.

Furthermore, a unified framework for a formal representation of COs has been introduced in this paper. The Rich Unified Content Description (RUCoD) provides a uniform descriptor for all types of Content Objects

irrespective of the underlying media and accompanying information. A specification of the RUCoD's constituting parts, using an example CO for illustration, was also provided.

Special focus was on the identification of the potential modifications of RUCoD in order to address the three different usage scenarios of I-SEARCH, since each use case has different requirements regarding media search and retrieval. As it was described in the relevant section, the generic nature of RUCoD is adequate to address all scenarios with minor modifications.

Finally, the relations of RUCoD with other well-known standards for multimedia description, such as MPEG-7, JPSearch and MPEG-21, were presented. RUCoD shares several common features with these standards but introduces also numerous innovative features, which are inline with the emerging demands of multimodal search in the Future Internet.

Acknowledgements

This work was supported by the EC project I-SEARCH (http://www.isearch-project.eu/).

References

Axenopoulos A., Daras P., Tzovaras D., "Towards the Creation of a Unified Framework for Multimodal Search and Retrieval" *2nd International ICST Conference on User Centric Media - UCMedia 2010*, Palma de Mallorca, September 1-3, 2010.

Burnett, I., Van de Walle, R., Hill, K., Bormans, J., Pereira, F., "MPEG-21: goals and achievements", *IEEE Multimedia,* Oct-Dec 2003,Vol. 10. Issue 4.

Dey A. K., Abowd G. D., "Towards a Better Understanding of Context and Context-Awareness", In HUC '99: Proceedings of the 1st international symposium on Handheld and Ubiquitous Computing, 1999.

DMOZ Open Directory Project, http://www.dmoz.org/

Dufaux F., Ansorge M., Ebrahimi T., "Overview of Jpsearch: A Standard for Image Search and Retrieval", *Content-Based Multimedia Indexing, CBMI'07*, London, June 2007.

Google Goggles, http://www.google.com/mobile/goggles/

IDC consultancy report,"The expanding Digital Universe", March 2007

Martinez JM., Koenen R., Pereira F. "MPEG-7: The generic multimedia content description standard, part 1". *IEEE Multimedia* 9(2):78–87 (2002)

Timmerer C., Vetrob A., Hellwagnera H, "Mpeg-21 Digital Item Adaptation", in *Encyclopedia of Multimedia 2nd ed.*, 2008.

Wordnet, A lexical database for English, http://wordnet.princeton.edu/

Yang Y., Xu D., Nie F., Luo J., Zhuang Y.. “Ranking with local regression and global alignment for cross-media retrieval”, *Proceedings of the seventeen ACM international conference on Multimedia*, Beijing, China, 2009.

Zahariadis T., Daras P., Bouwen J., Niebert N., Griffin D., Alvarez F., Camarillo G., "Towards a Content-Centric Internet", *Towards the Future Internet - Emerging Trends from European Research*, IOS Press, ISBN 978-1-60750-539-6, pp. 227-236, Apr 2010.

Zhang H., Weng J., ”Measuring Multi-modality Similarities Via Subspace Learning for Cross-Media Retrieval”, Lecture Notes in Computer Science, 2006, Volume 4261/2006.