Towards Integrated Authoring, Annotation, Retrieval, Adaptation, Personalization and Delivery of Multimedia Content*

Horst Eidenberger[1], Susane Boll[2], Stavros Christodoulakis[3], Doris Divotkey[1], Klaus Leopold[4] Alessandro Martin[5], Andrea Perego[6], Ansgar Scherp[7], and Chrisa Tsinaraki[3]

[1] Vienna University of Technology, Austria. Emails: eidenberger@ims.tuwien.ac.at, doris.divotkey@ims.tuwien.ac.at
[2] University of Oldenburg, Germany. Email: susanne.boll@informatik.uni-oldenburg.de
[3] Technical University of Crete. Emails: stavros@ced.tuc.gr, chrisa@ced.tuc.gr
[4] University of Klagenfurt, Austria. Email: klaus.leopold@itec.uni-klu.ac.at
[5] University of Milan, Italy. Email: martin@dico.unimi.it
[6] University of Insubria at Varese, Italy. Email: andrea.perego@uninsubria.it
[7] OFFIS Institute for Information Technology, Germany. Email: scherp@offis.de

Abstract. We describe the CoCoMA task of the DELOS II European Network of Excellence on Digital Libraries. CoCoMA aims at the unification of the most important aspects of multimedia management and multimedia presentation, i.e., the integration of authoring, annotation and presentation design with on-demand content adaptation, ad hoc media retrieval (semantics-based and content-based), and personalized delivery and visualization of presentations. The paramount goal of the CoCoMA activity is to maximize the added value from task and data integration by the identification and exploitations of connection points and inherent workflow similarities. The paper provides a brief description of the involved research fields, suggests a architecture for integrated multimedia consumption and presentation, and discusses the most prominent connection points (e.g., the reuse of content-based metadata for content adaptation and personalization). Problems and solutions are discussed jointly and illustrated by the components of the application prototype developed for the DELOS project.

1 Introduction

Multimedia presentations containing interactive media content go through a number of processing steps before they arrive at the user interface. Media streams are captured and manually or (semi-)automatically annotated on various levels of semantics. Single media items are spatio-temporally organized and authored to interactive presentations. During delivery, the content of media streams is adapted to technical requirements and personalized to user requirements.

In this paper we describe an approach to integrate these processing steps. We present the CoCoMA task of the DELOS II European Network of Excellence on Digital Libraries (www.delos.info). CoCoMA aims at a solution for the provision of content-

* This research was supported by the European Network of Excellence on Digital Libraries (DELOS) and by the Austrian Scientific Research Fund under grant number P16111.

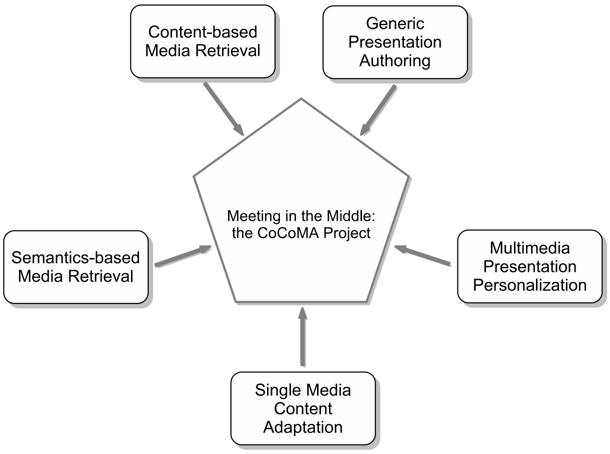
Content-based Media Retrieval
Generic Presentation Authoring
Meeting in the Middle: the CoCoMA Project
Semantics-based Media Retrieval
Multimedia Presentation Personalization
Single Media Content Adaptation

Fig. 1. The CoCoMA project for integrated multimedia presentation design

and context-aware rich interactive multimedia presentations by controlling data fusion and metadata reuse. Figure 1 sketches the basic idea. We focus on four major functions: generic presentation authoring, content-based and semantics-based annotation and retrieval, content adaptation and media personalization. For the sake of simplicity, the latter two areas are distinguished in this paper by the following attribution. Content adaptation summarizes all single and multi-media manipulation operations that are targeted towards technical requirements (e.g. network bandwidth). Media personalization denotes all manipulation operations targeted towards user requirements (e.g. special interests).

We believe that each of the four considered functions can benefit significantly from data collected and generated by the other functions. Content-based metadata (features like color histograms, camera motion, and cepstral audio features) may provide valuable control data for the content adaptation function (especially, in mobile environments). For example, such metadata can be employed to skip irrelevant information during encoding, or to select less restrictive quantization schemes for sensitive content (content with high motion activity). Semantically represented annotations (based on ontologies) and content-based metadata can jointly be used for interactive media querying in multimedia presentations. A common foundation could, e.g., be provided by the MPEG-7 standard [1]. The integration of the media annotation function with presentation authoring and presentation personalization allows for the seamless integration of the search functions. This way, personal experiences by non-linear movement through multimedia presentations become possible. Furthermore, search functions can be employed to provide novel means of presentation authoring and personalization. For example, we propose a novel presentation item that describes media objects by content-based queries instead of links to media files. Queries are executed at run-time (enriched with user knowledge) and the results are integrated with static presentation elements. This flexible way of personalization decides the presentation contents at view-time hand in hand with the user. Eventually, the application of constraint-based presentation authoring methods allows for flexible presentation personalization. The spatio-temporal relationships of CoCoMA presentations are not defined statically but by space and time operators. This paradigm reduces the burden of a layout on the presentation personalization to a minimum. Media objects can be added, exchanged and removed easily, if appropriate spatial and/or temporal rules for these operations are provided.

The CoCoMA project comprises of three steps. First, the exploitation of the fundamental idea and the identification of chances for integration. Second, the closer investigation of promising starting points and the design of solutions. In the second step, a complicating factor is that CoCoMA should possibly be based on earlier work by the involved research groups. Hence, the design process must consider architectural requirements and constraints of existing solutions. The last step is the implementation of the proposed design in a proof of concept prototype, and the construction of an application demonstrator. Suitable applications can, e.g., be identified in the sports domain such as summary presentations of soccer transmissions and various cultural domains like virtual tourist applications, preservation and cultural heritage applications. In summary, the CoCoMA activity aims at meeting in the middle, the integration of related functions of multimedia presentation design where promising starting points can be identified.

The paper is organized as follows. Section 2 gives background information on the involved research areas and illuminates starting points for integration. Section 3 discusses the architecture and building blocks of the proof of concept prototype. Section 4 focuses on challenging research questions. Solutions are illustrated by samples from the CoCoMA prototype.

2 Research areas and starting points for Integration

2.1 Content-Based Retrieval

Content-Based Retrieval means retrieval of media objects by their perceivable content: in case of image or video this means visual cues such as colors or shapes and in case of audio, retrieval is done by audible cues like sound or loudness. Multi-modal retrieval combines various media types such as text, image, audio and video.

State-of-the-art in content-based retrieval bases on the representation of audio and video by features. Meaningful features are extracted from the media objects and the actual retrieval step is performed by similarity measurement among media objects according to the extracted features. The difficulty to express high-level concepts with low-level features is called the semantic gap [2]. There exist various audio-visual features, similarity measures and retrieval models. Probabilistic models employ user relevance feedback information for retrieval (e.g., Binary Independence Retrieval). On the other hand, the most commonly applied approach is the Vector Space Model whereby media objects are represented by their feature vector and similarity is given as a distance measure (e.g., Euclidean distance) in the feature space [3].

The following paragraphs give a brief overview over common features for visual retrieval and for audio retrieval. Crucial work in the area of feature design has been performed by the MPEG group with the development of the MPEG-7 standard. The most important audio and visual descriptors have been investigated and standardized [1].

Visual Retrieval Commonly analyzed visual cues are color, texture, shape, and spatial localization and orientation of objects [2, 4, 5]. Histograms are the most frequently used technique to represent color distribution because of easy computation and efficient application. Transformation to frequency space (Cosine or Fourier Transform) is useful for texture analysis to determine characteristics like coarseness, direction and regularity. The usage of Gabor wavelets has proven to match human perception of texture. Shape analysis (e.g., by edge detection on the basis an edge histogram) is used for object detection and recognition, and—in case of video—also for object tracking. In terms of video, motion activity, camera and object motion are extracted from the video content.

Audio retrieval Audio retrieval comprises speech recognition, music analysis and environmental sound recognition. From the audio signal time domain and frequency domain features can be extracted. Frequency domain features are usually given by Discrete Cosine Transform (DCT) or Short Time Fourier Transform (STFT) [6]. Based on these coefficients more advanced features like Brightness and Total Spectrum Power can be built [7]. Examples for time domain features are Zero Crossing Rate (ZCR) and Short Time Energy (STE).

2.2 Semantics-Based Retrieval

Semantics-based retrieval for multimedia content rely on the metadata describing the content semantics. Semantics-based retrieval in multimedia is based on MPEG-7 [1], which is the dominant standard in multimedia content description. Although MPEG-7 allows, in the MPEG-7 MDS, the semantic description of the multimedia content using both keywords and structured semantic metadata, several systems follow the keyword-based approach [8–11]. The keyword-based approach is limiting, as it results in reduced precision of the multimedia content retrieval. As an example, consider a fan of the Formula-1 driver Alonso, who wishes to retrieve the audiovisual segments containing the overtakes that Alonso has performed against Hamilton. If the user relies on the keyword "overtake" and the names "Alonso" and "Hamilton", he will retrieve, in addition to the segments containing the overtakes of Alonso against Hamilton, the segments containing the overtakes of Hamilton against Alonso (which are also queried using the keyword "overtake" and the names "Alonso" and "Hamilton").

This problem may be solved, at least at some extent, if the structured semantic description capabilities provided by MPEG-7 are exploited. The major shortcoming of most of the systems based on the structured MPEG-7 semantic metadata is that the general-purpose constructs provided by MPEG-7 are used without a systematic effort for domain knowledge integration in MPEG-7 [12–14]. Domain knowledge, captured in domain ontologies expressed using MPEG-7 constructs, is systematically integrated in semantic MPEG-7 descriptions in [15]. In addition, a methodology for the integration of OWL domain ontologies in MPEG-7 has been developed in [16, 17], in order to allow the utilization of existing OWL domain ontologies, which make interoperability support within user communities easier.

Structured semantic content descriptions cannot be fully exploited by keyword-based user preferences; As the MPEG-7/21 user preferences allow only keyword-based descriptions of the desired content, the MPEG-7/21 based systems either utilize keyword-only metadata and ignore the structured MPEG-7 semantic metadata [9–11] or ignore the MPEG-7/21 user context model and follow proprietary user preference description approaches on top of the structured MPEG-7 semantic metadata [12]. A semantic user preference model for MPEG-7/21 has been proposed in [18] that allows the full exploitation of structured semantic multimedia content descriptions.

Another limitation is due to the lack of a transparent and unified multimedia content retrieval framework that allows exploiting all the aspects of the MPEG-7 multimedia content descriptions. A proposal for solving this problem is made in [18], where the MP7QL language is proposed. The MP7QL is a powerful query language for querying MPEG-7 descriptions, and also provides a user preference model that allows for expressing preferences about every aspect of an MPEG-7 multimedia content description. The MP7QL queries may utilize the user preferences as context, thus allowing for personalized multimedia content retrieval.

2.3 Presentation Modeling

A multimedia presentation may be considered as a graph, where each node corresponds to a set of heterogeneous multimedia objects (e.g., text, images, audio and video files), grouped depending on their content relationships and organized according to a given spatial and temporal disposition. By contrast, the edges connecting the nodes denote the execution flow of the presentationi.e., the sequence according to which the objects in each node are displayed to the user.

Multimedia presentation modeling then concerns two main issues: representing the presentation structure (i.e., the presentation graph) and representing the spatial and temporal disposition of objects in each node. The available approaches can be grouped into two main classes, operational and constraint-based, depending on how the spatial and temporal disposition of objects is represented.

In operational approaches, a presentation is specified by describing its final formi.e., the exact spatial and temporal location of objects inside each frame of the presentation is expressed by using (x, y) coordinates and timelines. Thanks to this, operational approaches have the advantage of being easy to implement. Nonetheless, they are not user-friendly and are not suitable when a presentation consists of a high number of objects. In fact, although authors have a complete control over the presentation, they are required to keep in mind the exact spatial and temporal position of each object in each node of the presentation.

In constraint-based approaches the final presentation is generated starting from a specification where constraints are used to represent the spatial and temporal relations existing among objects. This allows authors to provide a high-level specification of the presentation, whereas the task of deciding the exact spatial and temporal disposition of objects is in charge of the system. As a consequence, constraint-based systems are more flexible and user-friendly than operational ones, although they are more complex, due to the fact that they must carry out the presentation generation task.

Independently from their differences, both operational and constraint-based approaches are designed for building presentations using a fixed structure (usually modeled as a tree with one or more branches) and a fixed set of objects. Consequently, when alternative versions of the same presentation are required, varying in duration or using different subsets of objects, the author must specify them explicitly. This not only increases the complexity of the presentation specification task, but it also makes very difficult personalizing a presentation taking into account end users' interests and skill levels. In order to address this issue, a multimedia presentation authoring model has been developed, described in [19], where content relationships among objects are used to identify the objects associated with each node of the presentation and to build automatically different execution flows of the same presentation. This is obtained by supporting content constraints, allowing the author to specify a) the objects associated with the same "topic", b) the objects associated with different topics, and c) the objects associated with two consecutive topics. Such constraints can be specified explicitly or inferred from the content metadata possibly associated with multimedia objects. Thanks to these features, presentation specification becomes a task similar to object annotation, which results in making our approach suitable also for specifying presentations based on large repositories of multimedia object, such as digital libraries.

2.4 Content Adaptation

In [20] the architecture of an adaptive proxy for MPEG-4 visual streams is described which adapts MPEG-4 resources according to device capabilities and network characteristics. To this end, an adaptor chain concept has been introduced enabling the concatenation of several adaptation steps. The information of when an adaptor has to be invoked is hard coded in the proxy. Thus, this approach lacks extensibility in the sense that new adaptors can only be integrated into the existing system by re-compilation of the whole adaptation engine.

The MPEG-21 framework also supports tools for multimedia adaptation. This work is based on Bitstream Syntax Descriptions (BSD) [21, 22], i.e., an additional metadata layer which describes the high-level structure of a media bitstream. The main limitation of this approach is that one can only perform editing-style adaptation operations like removing, inserting, or updating parts of the bitstream. Another adaptation tool defined in the MPEG-21 framework is Adaptation QoS (AQoS) [23, 24] which enables users to describe the device and network quality of service (QoS). AQoS specifies the relationship between environmental constraints, media quality, and feasible adaptation operations. Adaptation engines can then perform look-ups in the AQoS table to ascertain adaptation operations for the multimedia content. Therefore, AQoS can provide hints for an adaptation decision taking engine. Only few projects are known at the moment that try to exploit the extended metadata annotation possibilities available with the new MPEG standards; examples are the ViTooKi Video Tool Kit project (`vitooki.sourceforge.net`) [25, 26] or the work described in [27].

2.5 Presentation Personalization

The personalization of multimedia presentations means the creation of multimedia content that meets a specific user's individual preference, interest, background and situational context—captured by a user profile. Even though one could prepare different documents for each targeted user or user group this would quickly become too laborious for many different users with their different user profiles. Hence, a dynamic creation of personalized content lies near at hand. Here, we find different approaches in the field. A research approach towards the dynamic generation of multimedia presentations based on constraints and logic programming is the Cuypers system [28, 29]. Within the Opéra project, a generic architecture for the automated construction of multimedia presentations based on transformation sheets and constraints is developed [30]. This work is continued within the WAM project with the focus on a negotiation and adaptation architecture for mobile multimedia services [31].

As indicated above, these and other existing research solutions typically use declarative description like rules, constraints, style sheets and the like to express the dynamic multimedia content creation. However, only those presentation adaptation problems can be solved that can be covered by such a declarative specification. Whenever a complex and application-specific personalization generation task is required, the systems find their limit and need additional programming. Approaches that base on XSLT would generally allow for a computationally complete transformation process—however, find their limitations in the manageability of large personalization applications. Consequently, we find with the MM4U framework a software engineering approach for the multimedia content adaptation and presentation [32, 33]. This framework provides application developers with a general, domain independent support for the creation of personalized multimedia content by exploiting the different approaches for multimedia content adaptation.

In order to be able to create multimedia content that is personalized for a certain user one needs multimedia content that can be used for the different users. Hence, retrieval of media content based on semantics, access to user profiles and the availability of adaptive content are prerequisites for building an integrated multimedia information system.

3 System design and building blocks

In this section, we present the component architecture of CoCoMA. The design falls in the two groups *presentation creation* and *presentation consumption*. Subsection 3.1 provides an overview over the project. The detailed discussions of design issues of presentation consumption and creation in subsections 3.2 and 3.3 are structured by the involved components, mostly stemming from our earlier work.

3.1 The component-based architecture of CoCoMA

Above we have sketched the overall goal of the CoCoMA activity, the integration of major aspects of multimedia presentation design. In detail, we endeavor to answer to the following major research questions:

1. How can the authoring process be enhanced by semantics-based and content-based media descriptions?
2. How can media metadata—in particular, content-based descriptions—be employed to achieve sophisticated content adaptation?
3. How can personalization and querying based on media metadata be integrated seamlessly? Is it possible to exploit the knowledge enveloped in the metadata for on-demand personalization?
4. How can semantics-based (e.g., ontology-based) and content-based metadata be merged and queried together?
5. Can we identify a constraint-based authoring process and presentation description scheme that simplifies personalization by offering the required degrees of freedom?

Obviously, all questions are targeted towards efficient metadata unification and reuse. Hence, design of metadata management is the central topic of the CoCoMA activity.

The building blocks that produce and consume the metadata are the same as named in the introduction: authoring, content-based annotation, semantics-based annotation, content adaptation and personalization. Figure 2 structures their relationships and their interfaces to the outside world. The presentation author and media annotator interacts with the authoring building block and the semantic annotation interface. A knowledge base of ontologies supports the annotation process. The presentation consumer interacts exclusively with the personalization and delivery component. Content-based annotation and content adaptation have no external interfaces. Content-based annotation is controlled by the authoring process. Content adaptation is triggered by the presentation engine.

Media data and media metadata are organized in two major databases. The media database holds the temporal (e.g., audio, video) and non-temporal (e.g., text, image) media content specific to the presentation context. The metadata database stores a variety of media-related metadata, user knowledge and system parameters. Metadata includes non-temporal data (e.g., textual media segment descriptions, domain ontologies, presentation constraints) and temporal data (e.g., motion descriptions, spectral audio descriptions). The metadata database is mostly filled by the two annotation building blocks and by the authoring process. Metadata is consumed by the content adaptation function (e.g., low-level color models, high-level relevance estimations) and by the personalization building block (e.g., merged with user knowledge for content-based media selection).

Apart from the research issues listed above, the integration process constitutes some engineering problems that have to be solved properly. An important prerequisite is that the CoCoMA design should—as far as possible—be based on existing solutions provided by the project participants. Hence, the integration process should—similarly to enterprise application integration—focus on appropriate metadata formats. In this context, a prominent role is played by the MPEG-7 standard for multimedia content description [1]. MPEG-7 provides structures for textual annotation and content-based annotation. These structures are employed for single media and multimedia descriptions. In addition, extensions are implemented where necessary (e.g., for ontology integration, MPEG-21-based media stream description, etc.). See Section 4 for details.

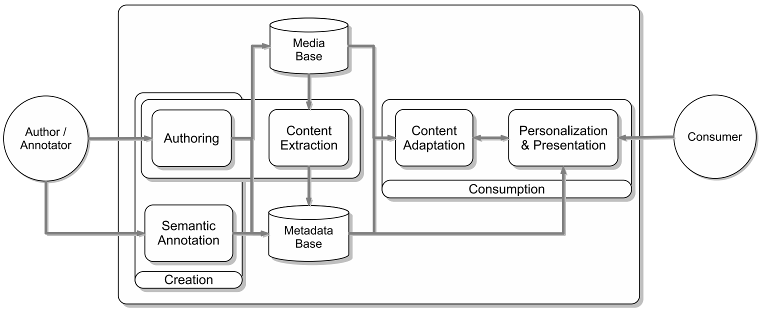

Fig. 2. CoCoMA building blocks and workflow

The remainder of this section discusses the building blocks of the CoCoMA design. Subsection 3.2 focuses on the components for presentation consumption. This user perspective comprises of a framework for the implementation of personalized mobile multimedia presentations and a framework for sophisticated content adaptation. Subsection 3.3 illuminates the presentation author's perspective. We introduce two frameworks for content-based and semantics-based annotation of multimedia content and our approach for the generation of constraint-based interactive multimedia presentations.

3.2 Presentation Consumption

MM4U for Personalized Multimedia Presentations The overall goal of the MM4U framework is to simplify and to improve the development process of personalized multimedia applications. For this, the MM4U framework provides application developers with an extensive support for creating personalized multimedia content. This support comprises assistance for the access to media data and associated metadata as well as user profile information and contextual information. The framework provides for the selection and composition and transformation of media elements into a coherent multimedia presentation.

For supporting the different tasks of the general multimedia personalization process, we developed a layered architecture. Each layer provides modular support for the different tasks of the multimedia personalization process. The access to user profile information and media data are realized by the layers (1) and (2), followed by the two layers (3) and (4) in the middle for composition of the multimedia presentation in an internal multimedia content representation model and its later transformation into the concrete presentation output formats. Finally, the top layer (5) realizes the rendering and display of the multimedia presentation on the end device. To be most flexible in regard of the different requirements of concrete personalized multimedia applications, the framework's layer allow extending the functionality of the MM4U framework.

The layered architecture of the MM4U framework allows being easily adapted to the particular requirements that can occur in the development of personalized multimedia applications. For example, special User Profile Connectors as well as Media Data

Connectors can be embedded into the MM4U framework to integrate the most diverse and individual solutions for storage, retrieval, and gathering for user profile information and media data information. The Multimedia Composition layer allows to be extended by complex and sophisticated composition operators. Thus, arbitrary personalization functionality can be added to the framework. The Presentation Format Generators layer allows integrating any output format into the framework to support the most different multimedia players that are available for the different end devices.

In this paper, we show a proof of concept of this openness and extensibility of the framework as it is embedded in a larger setting and integrated with an adaptive streaming technology (Section **??**) and content-based retrieval approach (Section **??**).

KoMMA for Multimedia Content Adaptation Intelligent adaptation of multimedia resources is becoming increasingly important and challenging for two reasons. First, the market continuously brings up new mobile end-user devices to which the content has to be adapted as these devices support different display formats and operate on various types of networks. On the other hand, with the help of metadata annotations, which are now available in the MPEG-7 and MPEG-21 standard, advanced forms of resource adaptations on the content level become possible. As none of the existing multimedia transformation tools and libraries can support all these different forms of basic and advanced adaptation operations, an intelligent multimedia adaptation node has to integrate such external tools and algorithms and perform an adequate sequence of adaptation operations on the original resource before sending it to the client [34].

In order to compute an adequate sequence of adaptation operations, we utilize a knowledge-based planning approach [35]. In general, a planner computes a plan by applying actions on an initial state to reach a goal state. In the context of multimedia adaptation, the initial state corresponds to the original multimedia resource, which can be described by means of MPEG-7 descriptions. The goal state is the adapted multimedia content according to the usage context which is, e.g., terminal capabilities or usage environment. The usage context can be expressed by means of MPEG-21 descriptions. Finally, actions are adaptation operations that have to be applied on the original multimedia content in order to meet the usage context.

In the implementation of the multimedia adaptation node, the described planner—referred to as the adaptation decision-taking engine—acts as preprocessing module for the adaptation engine. Upon a client request, the adaptation decision-taking engine computes an adaptation plan which is later executed by the adaptation engine [36].

3.3 Presentation Creation

VizIR for Content-Based Media Annotation and Retrieval VizIR is an open framework providing common ground functionalities and strategies for the development of multimedia applications that benefit from multimedia retrieval techniques [37]. VizIR is a workbench, designed to be easily adaptable and extensible. The creation of reusable assets has been a major design goal. This means that the components of VizIR are designed to facilitate changes and extensions. The interfaces between the components are clearly defined to assure interoperability and the coupling is loose to allow for easy exchange.

The framework provides a comprehensive collection of classes for all major multimedia retrieval tasks such as storage and management of media and annotated metadata [38]. It allows for the textual annotation of semantic information about the content as well as content-based metadata directly extracted from the media content. The core item of the VizIR system is the strong and flexible querying and retrieval component. It comprises algorithms for automatic feature extraction and similarity measurement among media objects based on the derived media descriptions. Furthermore, the VizIR framework contains a set of user interfaces for browsing the media databases, query formulation (by example or sketch) and query refinement and a couple visualization tools. The framework provides implementations of various content-based descriptors for image, audio and video data. Amongst them, most of the visual descriptors of the MPEG-7 standard (such as Dominant Color, Color Structure, Edge Histogram, Camera Motion, Motion Activity, etc.) have been implemented [1]. Furthermore, VizIR incorporates a set of state-of-the-art audio descriptors from various application domains [39].

VizIR allows for the usage of arbitrary features and querying models. To accomplish this, a generic querying language was developed [40]. Depending on the underlying querying model that is used the formulation of queries happens on different levels of abstraction. This means that it is either possible to state queries on a very low-level by defining explicitly the low-level features and the used querying scheme or to define queries on a semantically high level. Thereby, the querying component uses models to break down the high-level query and translates it to a lower level that can be solved. Moreover, the combination of features of different type (audio, video, text) is possible, which lays the foundation for multi-modal retrieval. The retrieval component in general and the querying language in particular may as well be adapted to take semantics-based annotations into account. For this purpose, the VizIR framework contains an implementation of the full MPEG-7 Multimedia Description Schemes to describe and annotate multimedia data [41].

The power and flexibility of the VizIR framework forms a solid basis for the envisioned integrated solution to provide content- and context-aware rich interactive multimedia presentations.

DS-MIRF for Semantics-Based Media Annotation and Retrieval The DS-MIRF (Domain-Specific Multimedia Indexing, Retrieval and Filtering) Framework [16, 15, 17] aims to facilitate the development of knowledge-based multimedia applications (including multimedia information retrieval, information filtering, user behavior description, multimedia content segmentation, multimedia information extraction and multimedia browsing and interaction) utilizing and extending the MPEG-7 and MPEG-21 standards.

The major components of the DS-MIRF framework are the following:

1. The *DS-MIRF Metadata Repository*, where domain ontologies and multimedia content descriptions are stored in MPEG-7 format. In addition to the current MPEG-7/21 metadata, the DS-MIRF Metadata Repository allows the management of semantic user preferences as described in [18]. Semantic queries are supported on top of the DS-MIRF metadata repository. The repository is accessed by the end-users

through appropriate application interfaces that utilize the expressive power of the MP7QL query language [18].

2. The *DS-MIRF Ontological Infrastructure* [16, 17], which includes: (1) An OWL Upper Ontology that fully captures the MPEG-7 MDS and the MPEG-21 DIA Architecture (the latter has been developed in the context of CoCoMA). (2) OWL Application Ontologies that provide additional functionality in OWL that either makes easier the use of the MPEG-7 MDS from the users (like a typed relationship ontology based on the MPEG-7 MDS textual description) or allows the provision of advanced multimedia content services (like a semantic user preference ontology that facilitates semantic-based filtering and retrieval). (3) OWL Domain Ontologies that extend both the Upper Ontology and the Application Ontologies with domain knowledge (e.g. sports ontologies, educational ontologies etc.).
3. The *GraphOnto Component* [42], which is an ontology-based semantic annotation component. GraphOnto facilitates both OWL ontology editing and OWL/RDF metadata definition and allows transforming both domain ontologies and metadata to MPEG-7 metadata descriptions. The MPEG-7 metadata may be stored in files or in the DS-MIRF Metadata Repository.

SyMPA for Content-Based Multimedia Presentation Authoring and Generation The presentation specification and generation component of the CoCoMA architecture, referred to as SyMPA, is based on the multimedia presentation model described in [19], which allows authors to group semantically related objects into independent sets representing each one a *topic*. A topic itself can be a presentation, since it is composed of a set of objects, played according to a given sequence. This is obtained by using a new class of constraints, called content constraints, that allow the author to define high-level, content-related semantic relations among objects, in order to build different presentation topics and the interconnections among them. In SyMPA, content constraints are not explicitly specified by the presentation author, but inferred from the annotations possibly associated with multimedia objects. Authors annotate objects using multiple metadata vocabularies (which may be plain sets of descriptors, conceptual hierarchies, and ontologies), concerning both high- and low-level features. Then they make use of content metadata in order to define the main topic of a presentation. Based on this, SyMPA retrieves the objects satisfying the query, andit groups them into nested subsets, determining both the nodes of the presentation and its structure. The author may then revise the presentation by modifying the presentation structure, the contents of each node of the presentation, and/or the spatio-temporal disposition of objects. This approach has the advantage of being easily applied to large repositories of multimedia objects (such as digital libraries), where multiple authors can annotate objects in a collaborative way and objects can be added and removed dynamically. Besides presentation specification and generation, the standalone version of SyMPA is designed also as a system for the management and annotation of multimedia objects stored in distributed repositories. For this purpose, the SyMPA architecture consists of three main components: a) a centralized database, storing object metadata and presentation specifications, b) a set of tools for performing object management and annotation, and presentation specification, and c) two modules in charge of, respectively, presentation generation and Object/Presentation

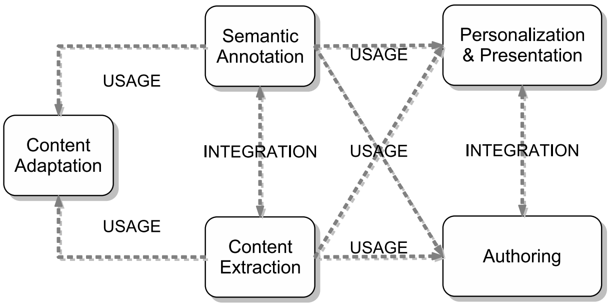

Fig. 3. CoCoMA integration challenges

retrieval. The tools for managing the system and authoring/displaying multimedia presentations are accessible by both end-users and authors through a Web interface. In CoCoMA, SyMPA has been extended with an interface allowing the exploitation of the features provided by the DS-MIRF framework (see Section 3.3).

4 Integration Challenges and Solutions

This section describes how the individual CoCoMA components are merged and how the components are employed for the benefit of others. Figure 3 sketches the components and the neuralgic connection points (mostly characterized by usage relationships). Subsections 4.1 to 4.5 explicate the five major points of integration.

4.1 Integration of Content-Based and Semantics-Based Retrieval

An important integration issue is the integration of CBR (Content-Based Retrieval—based on low-level features) with SBR (Semantic-Based Retrieval). A lot of independent research exists for both the approaches, but there are several real-life situations where none of the approaches can work by itself at a satisfactory level. Consider, as an example, a user interested in art who wants to get a drawing containing a girl who wears a hat. The user also remembers that the black colour dominates in the drawing. In this case, if the user uses SBR only, he will receive all the drawings containing a girl who wears a hat and the user has to browse the results in order to locate the one he has in mind. Using CBR only, the user would request the drawings where black dominates and the user has to browse the results in order to locate the one he has in mind. A more convenient approach would be to allow the user to pose queries having both semantic components and visual feature components and use one technique (e.g., SBR) for pre-filtering and the other (e.g., CBR) for refinement.

In order to support the above approach, we are working on the support of *Semantic and Content Based Retrieval* (SCBR), which allows for providing *Semantic and Content Based Queries* (SCBQs). These queries allow the specification of preference values (in the range $[-100, 100]$) for their constituents and may contain boolean operators. An

$$SCBQ = (SQ\ pv)\ (CQ\ pv)$$

Fig. 4. Regular Expression describing a Semantic and Content-based Query (*SCBQ*) without boolean operators. *SQ* is a Semantic Query component, *CQ* is a Content-based component and *pv* is a Preference Value.

$$SCBQ = ((SQ|CQ)\ pv)\ ((\mathrm{AND}|\mathrm{OR})\ (SQ|CQ)\ pv)*$$

Fig. 5. Regular Expression describing a Semantic and Content Based Query (*SCBQ*) with Boolean Operators. *SQ* is a Semantic Query component, *CQ* is a Content-based component and *pv* is a Preference Value.

SCBQ without boolean operators is described by the regular expression 4, while an SCBQ with boolean operators is described by the regular expression 5.

If boolean operators are not used, then:

1. A Semantic Query component is a list of desired Semantic Entity Specifications (SES), for each of which a user-defined preference value is specified. A semantic entity specification contains the desired semantic entity attribute and element values (for example the desired values of a person's name, age, relationships with other persons etc.). An SQ without boolean operators is described by the regular expression 6.
2. A Content-based Query component is a list of Visual Feature Specifications (VFS), for each of which a user-defined preference value is specified. A visual feature specification contains the name of the visual feature and the desired value (or the range of desired values) (for example, dominant colour, pattern etc.). A CQ without boolean operators is described by the regular expression 7.

If boolean operators are used, then:

1. A Semantic Query component is the combination, using the boolean operators AND and OR, of desired Semantic Entity Specifications (SES), for each of which a user-defined preference value is specified. An SQ with boolean operators is described by the regular expression 8.
2. A Content-based Query component is the combination, using the boolean operators AND and OR, of desired Visual Feature Specifications (VFS), for each of which a user-defined preference value is specified. A CQ with boolean operators is described by the regular expression 9.

The SCBR queries can be expressed using the MP7QL syntax [18]. As an example, the query requesting a drawing where the black colour dominates and a girl who wears a hat is depicted, would be described by the formal expression 10. The query components of the example query are supposed to be of equal preference value.

4.2 Content-Based and Semantics-Based Metadata for Content Adaptation

The knowledge-based multimedia adaptation node within the CoCOMA architecture is designed as a proxy server, which forwards incoming client requests to the multimedia

$$SQ = (SES\ pv)*$$

Fig. 6. Regular Expression describing a Semantic Query component without boolean operators.

$$CQ = (VFS\ pv)*$$

Fig. 7. Regular Expression describing a Content-based Query component without boolean operators.

server. The server transfers the requested content together with the MPEG-7 description to the adaptation proxy which then adapts the content based on the client's usage context which is described by means of MPEG-21 metadata. Finally, the adapted content is streamed to the client together with the adapted MPEG-7 descriptions.

The main concept of the adaptation decision-taking engine is to describe multimedia adaptation operations semantically by their inputs, outputs, preconditions, and effects (IOPE). The operation's preconditions express the conditions that must hold before an operation can be applied. Typical preconditions for an image grey scale operation are, e.g., "the multimedia content is a JPEG image" and "the image is colored." Effects express the changes after applying an adaptation operation. The effect of a grey scale operation might be "the output image is grey." The semantics of an adaptation operation like, for instance, "JPEG image" is introduced by referencing MPEG-7 and MPEG-21 metadata which enables content-aware media adaptation.

The content-aware approach offers a wide range of possibilities for media adaptation, starting with general adaptation to meet the client's usage context and up to new sophisticated CBR-based content adaptation. Content adaptation based on CBR means that content-based features (as extracted for retrieval) are employed in the adaptation process. For example, it would be thinkable to use motion features to judge the visual complexity of scenes and adapt the encoding accordingly. On a semantically higher level, content-based features could be used to summarize scenes, skip irrelevant scenes and save bandwidth. Audio features could be employed to estimate the type of content (speech, music, etc.) and choose the encoding appropriately.

4.3 Authoring Based on Content-Based and Semantics-Based Metadata

The availability of content-based and semantic metadata, associated with multimedia objects, can be used to automatically carry out the presentation specification task. In particular, such metadata can be used in our authoring model to infer the content relationships existing among objects, which will then determine the set objects associated with each node of the presentation and the presentation structure itself. Author's intervention is required for carrying out two main tasks: (a) determining the spatial and temporal disposition of objects in each node of the presentation, and (b) selecting the objects to be actually used, based on their relevance.

So, if content-based and semantic metadata cannot make completely automatic the presentation specification procedure, they can be used for improving its efficiency, especially when dealing with large collections of objects, where finding objects may be a

$$SQ = (SES\ pv)\ ((\mathrm{AND}|\mathrm{OR})\ SES\ pv)*$$

Fig. 8. Regular Expression describing a Semantic Query component with boolean operators.

$$CQ = (VFS\ pv)\ ((\mathrm{AND}|\mathrm{OR})\ VFS\ pv)*$$

Fig. 9. Regular Expression describing a Content-based Query component with boolean operators.

difficult and time-consuming task. Authors may specify a presentation by defining a set of topics in terms of the semantic metadata available in the system. Based on this, the system returns the set of objects belonging to each node; then the author decides which objects should be used and their spatial and temporal disposition. Finally, the possible execution flows of the presentation are obtained evaluating the semantic relationships existing among the selected objects.

In the context of CoCoMA, the MPEG-7/21 metadata stored in the DS-MIRF Metadata Repository, which are associated with multimedia objects, are utilized to automatically carry out the presentation specification task in SyMPA. In addition, the DS-MIRF Metadata Repository is used both for storing multimedia presentations and object annotations defined using SyMPA and for locating multimedia objects that will be utilized in the presentations (using the semantic retrieval capabilities of the DS-MIRF framework). In addition, the DS-MIRF Metadata Repository is used for retrieving the metadata associated with the multimedia objects, which are used by the authoring model to infer the content relationships that exist among objects. The user preferences stored in the DS-MIRF Metadata Repository will be utilized systematically in order to allow presentation personalization so as to take into account the user likes and dislikes and to meet duration and/or space (implicit or explicit) constraints. Finally, the DS-MIRF ontological infrastructure is utilized by SyMPA as a set of metadata vocabularies for the selection of knowledge domain and topic of interest and will be extended with an ontology about ‘art’. The integration actvities involving SyMPA end the DS-MIRF framework are described in detail in [43].

4.4 Adaptive streaming in personalized multimedia presentations

The framework MM4U for the dynamic composition of personalized multimedia presentation has been integrated with the KoMMA framework for adaptive streaming of video content. During the presentation of a personalized multimedia document the adaptive streaming framework delivers the continuous media stream that best meets the current presentation situation. During the composition phase of the multimedia document, the media access layer of the framework searches the underlying media store and includes a reference to the content into the document.

The profile that drives the composition of the personalized presentation provides the parameters that apply to the adaptive video streaming. During the personalized composition of the multimedia document, those parameters from the user profile, that affect the presentation such as device configuration, bandwidth, etc., are included as an MPEG-21

$$SCBQ1 = ((AGirl, AgentObjectType)\ (exemplifies, Girl)\ (component, hat))\ 100)$$
$$((dominant - color\ black)\ 100)$$

Fig. 10. Formal Expression describing a query requesting a drawing where the black colour dominates and a girl who wears a hat is depicted.

description into the document. For the later access to the media content the media locator also includes the location of the proxy that finally creates and delivers the adapted stream. Hence, the parameters relevant for the adaptive are selected during the composition process and inserted in the presentation. After the composition of the multimedia content, a transformation into the final presentation format such as SMIL, SVG or Flash is executed. When the presentation is delivered to the user and is rendered the player actually accesses the media content. The reference to the media content is resolved and the player accesses the adaptation proxy and receives the adapted media content. This request includes the presentation context parameters which are then used by the adaptation proxy to dynamically adapt the streaming media to the client's presentation requirements.

4.5 Personalized presentation composition based on content-based retrieval

Besides the integration with adaptive streaming framework for video content, the MM4U framework has also been integrated with the visual information retrieval framework VizIR. The VizIR framework provides a generic architecture for developing visual information retrieval systems. It can be applied for any querying model that bases on the extraction of visual information of media elements and the computation of similarity of media elements by distance measurement in a feature space. A generic querying language for the VizIR framework has been developed with concrete instances for particular models. For the integrated solution of the MM4U and VizIR frameworks, the k-nearest neighbor model is used. The interfaces between the two frameworks are kept generic to allow for future adaptation of the model and the extension to different models.

For the integration of both frameworks, a new media connector for the MM4U framework has been developed. This VizIRMediaConnector uses the QueryObject provided by the MM4U framework for specifying requests to the underlying media connector. For the integration with the VizIR framework, this query object has been extended to support the parameters of a content-based retrieval query.

The querying object is passed from the MM4U framework via the VizIRMediaConnector to the VizIR framework where it is actually executed. The VizIR framework determines a ranked list of the most suitable media elements according to the given query and returns the query result back to the media connector. The retrieved query result is then converted within the VizIRMediaConnector to the MM4U compliant representation of the media elements.

With the integration of the VizIR framework, we enhanced the MM4U framework with capabilities for content-based meta data extraction and content-based retrieval techniques. The MM4U framework can use the content-based retrieval functionality the VizIR framework offers in different usage scenarios, acting on three different abstraction levels. These usage scenarios are:

1. The MM4U framework uses CBR querying internally (without knowledge of the user) to retrieve media elements similar to an example media element, e.g., an image.
2. The user chooses a media element that is used as input for a content-based query. However, no further knowledge about content-based querying is required from the user; especially no feature-selection needs to be conducted by the user. The MM4U framework uses a set of predefined or default values for the querying parameter, e.g., by appropriate descriptors identifying the type of the query like “sunset” or “landscape”. Here, the MM4U framework exploits the profile information about the user to optimize the content-based querying.
3. The user specifies a content-based retrieval query including additional query options. For this advanced mode, a user interface for the specification of content-based queries needs to be developed.

5 Conclusions and future work

The CoCoMA task of the DELOS II European Network of Excellence on Digital Libraries endeavors to integrate traditionally independent components of multimedia systems. The integration of content-based retrieval and semantics-based retrieval results in more precise retrieval results. Employing content-based and semantics-based retrieval methods for multimedia authoring, content adaptation, and personalization provides additional degrees of freedom for the media designer and leads to richer multimedia applications with higher flexibility. Eventually, the consideration of personalization issues in the multimedia authoring process refines it to a user-centered activity expressed in presentation-specific constraints.

In this work, we described the vision of CoCoMA, briefly sketch the involved research areas, state the major integration problems and illustrate novel paths to solve them. CoCoMA is work in progress with a clear focus on methodological integration. Currently, we are designing a service-oriented architecture where the individual components act as services and are integrated by a workflow management system. Following this scheme, our future work will be the implementation and user-based evaluation of a full-featured CoCoMA infrastructure.

Acknowledgements The authors would like to thank Christian Breiteneder and Hermann Hellwagner for their support and valuable suggestions for improvement.

References

1. Manjunath, B.S., Salembier, P., Sikora, T., eds.: Introduction to MPEG-7: Multimedia Content Description Interface. Wiley (2002)
2. Del Bimbo, A.: Visual Information Retrieval. Morgan Kaufmann (1999)

3. Fuhr, N.: Information retrieval methods for multimedia objects. In Veltkamp, R.C., Burkhardt, H., Kriegel, H.P., eds.: State-of-the-Art in Content-Based Image and Video Retrieval. Kluwer (2001) 191–212
4. Lew, M.S., ed.: Principles of Visual Information Retrieval. Springer (2001)
5. Marques, O.E., Furht, B.: Content-Based Image and Video Retrieval. Kluwer (2002)
6. Choi, Y.K., Kim, K.M., Jung, J.W., Chun, S.Y., Park, K.S.: Acoustic intruder detection system for home security. IEEE Transactions on Consumer Electronics **51**(1) (2005) 130–138
7. Guo, G., Li, S.Z.: Content-based audio classification and retrieval by support vector machines. IEEE Transactions on Neural Networks **14**(1) (2003) 209–215
8. Graves, A., Lalmas, M.: Video retrieval using an MPEG-7 based inference network. In: SIGIR 2002 Proceedings. (2002) 339–346
9. Rogers, D., Hunter, J., Kosovic, D.: The TV-Trawler project. International Journal of Imaging Systems and Technology **13**(5) (2004) 289–296
10. Tseng, B.L., Lin, C.Y., Smith, J.: Using MPEG-7 and MPEG-21 for personalizing videos. IEEE Multimedia **11**(1) (2004) 42–52
11. Wang, Q., Balke, W.T., Kießling, W., Huhn, A.: P-News: Deeply personalized news dissemination for MPEG-7 based digital libraries. In: ECDL 2004 Proceedings. (2004) 256–268
12. Agius, H., Angelides, M.: Modelling and filtering of MPEG-7-compliant meta-data for digital video. In: ACM SAC 2004 Proceedings. (2004) 1248–1252
13. Hammiche, S., Benbernou, S., Hacid, M.S., Vakali, A.: Semantic retrieval of multimedia data. In: MMDB 2004 Proceedings. (2004) 36–44
14. Lux, M., Granitzer, M.: Retrieval of MPEG-7 based semantic descriptions. In: BTW-Workshop "WebDB Meets IR" Proceedings. (2004)
15. Tsinaraki, C., Polydoros, P., Kazasis, F., Christodoulakis, S.: Ontology-based semantic indexing for MPEG-7 and TV-Anytime audiovisual content. Multimedia Tools and Application Journal **26** (2005) 299–325
16. Tsinaraki, C., Polydoros, P., Christodoulakis, S.: Interoperability support for ontology-based video retrieval applications. In: CIVR 2004 Proceedings. (2004) 582–591
17. Tsinaraki, C., Polydoros, P., Christodoulakis, S.: Interoperability support between MPEG-7/21 and OWL in DS-MIRF. IEEE Transactions on Knowledge and Data Engineering **19**(2) (2007) 219–232
18. Tsinaraki, C., Christodoulakis, S.: A user preference model and a query language that allow semantic retrieval and filtering of multimedia content. In: SMAP 2006 Workshop Proceedings. (2006) 121–128
19. Bertino, E., Ferrari, E., Perego, A., Santi, D.: A constraint-based approach for the authoring of multi-topic multimedia presentations. In: ICME 2005 Proceedings. (2005) 578–581
20. Schojer, P., Böszörményi, L., Hellwagner, H.: QBIX-G – A transcoding multimedia proxy. In: MMCN 2006 Proceedings. (2006)
21. Devillers, S., Timmerer, C., Heuer, J., Hellwagner, H.: Bitstream syntax description-based adaptation in streaming and constrained environments. IEEE Transaction on Multimedia **7**(3) (2005) 463–470
22. Vetro, A., Timmerer, C.: Digital item adaptation: Overview of standardization and research activities. IEEE Transaction on Multimedia **7**(3) (2005) 418–426
23. Vetro, A., Timmerer, C., eds.: Text of ISO/IEC 21000-7 FCD Part 7: Digital Item Adaptation. (2003)
24. Mukherjee, D., Delfosse, E., Kim, J.G., Wang, Y.: Optimal adaptation decision-taking for terminal and network quality-of-service. IEEE Transactions on Multimedia **7**(3) (2005) 454–462

25. Böszörményi, L., Hellwagner, H., Kosch, H., Libsie, M., Podlipnig, S.: Metadata driven adaptation in the ADMITS project. EURASIP Signal Processing and Image Communication Journal **18** (2003) 749–766
26. Schojer, P., Böszörményi, L., Hellwagner, H., Penz, B., Podlipnig, S.: Architecture of a quality based intelligent proxy (QBIX) for MPEG-4 videos. In: ACM WWW 2003 Proceedings. (2003) 394402
27. Steiger, O., Sanjuan, D.M., Ebrahimi, T.: MPEG-based personalized content delivery. In: IEEE ICIP 2003 Proceedings. (2003) 14–16
28. Geurts, J., van Ossenbruggen, J., Hardman, L.: Application-specific constraints for multimedia presentation generation. In: MMM 2001 Proceedings. (2001) 339–346
29. Cuypers: The Cuypers multimedia transformation engine website. Project Web site, University of Amsterdam (2006) `http://homepages.cwi.nl/~media/cuypers/`.
30. Bes, F., Jourdan, M., Khantache, F.: A generic architecture for automated construction of multimedia presentations. In: MMM 2001 Proceedings. (2001) 229–246
31. Lemlouma, T., Layaïda, N.: Context-aware adaptation for mobile devices. In: IEEE MDM 2004 Proceedings. (2004) 106–111
32. Scherp, A., Boll, S.: MM4U: A framework for creating personalized multimedia content. In Srinivasan, U., Nepal, S., eds.: Managing Multimedia Semantics. IRM Press (2005) 246–287
33. Scherp, A., Boll, S.: Paving the last mile for multi-channel multimedia presentation generation. In: MMM 2005 Proceedings. (2005) 190–197
34. Leopold, K., Jannach, D., Hellwagner, H.: A knowledge and component based multimedia adaptation framework. In: IEEE MSE 2004 Proceedings. (2004) 10–17
35. Jannach, D., Leopold, K., Timmerer, C., Hellwagner, H.: A knowledge-based framework for multimedia adaptation. Applied Intelligence **24**(2) (2006)
36. Jannach, D., Leopold, K.: Knowledge-based multimedia adaptation for ubiquitous multimedia consumption. Journal of Network and Computer Applications **30**(3) (2007) 958–982
37. Eidenberger, H., Breiteneder, C.: VizIR – A framework for visual information retrieval. Journal of Visual Languages and Computing **14**(5) (2003) 443–469
38. Eidenberger, H., Divotkey, R.: A data management layer for visual information retrieval. In: ACM MDM Workshop Proceedings. (2004)
39. Mitrovic, D., Zeppelzauer, M., Eidenberger, H.: Analysis of the data quality of audio descriptors of environmental sounds. In: WMS 2006 Proceedings. (2006) 70–79
40. Divotkey, D., Eidenberger, H., Divotkey, R.: Artificial intelligence and query execution methods in the VizIR framework. Journal of the Austrian Artificial Intelligence Society **24**(2) (2005) 17–27
41. Salembier, P.: MPEG-7 multimedia description schemes. IEEE Transactions on Circuits and Systems for Video Technology **11**(6) (2001) 748–759
42. Polydoros, P., Tsinaraki, C., Christodoulakis, S.: GraphOnto: OWL-based ontology management and multimedia annotation in the DS-MIRF framework. In: WMS 2006 Proceedings. (2006)
43. Tsinaraki, C., Perego, A., Polydoros, P., Syntzanaki, A., Martin, A., Christodoulakis, S.: Semantic, constraint & preference based multimedia presentation authoring. Journal of Digital Information Management **4**(4) (2006) 207–213