Background

We introduce the problems of multimedia generation as identified in the research literature. We also demonstrate how the trend has been to increasingly use knowledge-based solutions, led by the introduction of such techniques to the Web in the vision known as the "Semantic Web". These techniques have also been taken up by the multimedia community in their aim to implement Intelligent Multimedia Presentation Systems (IMMPS).

The previous chapter has observed technological trends and consequently suggested important requirements for future multimedia presentation systems. These requirements specify a need for a system which can, to as great an extent as possible automatically, select, adapt and integrate content to produce meaningful multimedia presentations. Traditional multimedia presentation research has produced systems which do not meet these requirements adequately. This lies in some accepted problems identified in the tasks which make up multimedia presentation generation.

This chapter looks at those problems (section 2.1), and then turns to knowledge representation theory (section 2.2). This field has become increasingly relevant to the information technology and Web communities (section 2.3), and been proposed as the panacea to a broad range of common problems.

It has also been picked up by the multimedia community and the work overlapping the two fields is reviewed (section 2.4). One particular proposal, an Intelligent Multimedia Presentation System (IMMPS), is considered in some more depth (section 2.5) as it will be referenced again in future chapters as an important theoretical basis for the implementation of this research.

2.1 Traditional multimedia generation

From considering multimedia presentation systems for commonalities, it can be determined that multimedia presentation generation can be divided into three tasks which exhibit varying degrees of interdependence in implemented systems: annotation/retrieval, adaptation and presentation.

These tasks are not independent from one another. Rather than being executed in a purely linear fashion, systems tend to support the interchange of data between the tasks. This is a result of multimedia presentation results not being determinable at execution due to the possible variation in the content being presented, its flexibility in being adapted to the presentation context and the constraints in the final presentation that must be respected. As a consequence later stages in the generation process may determine that a particular step can not be realized and must backtrack and repeat earlier steps under new conditions.

Each task does however involve different implementation and realization issues which contribute to the known difficulties in the multimedia generation field. These tasks and their known problems are introduced here.

2.1.1 Annotation and Retrieval

Multimedia content selection is a very different and difficult problem in comparison with textual retrieval where the query is usually also textual and is realized by string matching in the content store, aided by devices such as stemming and synonyms [Gauch,1991].

The key problems in the case of multimedia retrieval are that the form of query does not generally match the form of media being queried and with queries that are of the same form (e.g. user whistling to search an audio database) matching techniques are more complex than with text.

MMIR (multimedia information retrieval) systems have tended to focus on the latter case in order to perfect algorithms for non-textual media matching. However such low-level matching has the restriction of requiring the query to be in the same form as the stored media, and conversely, that the stored media is all of a single form. Hence mixed media stores are excluded from this approach, and queries are often not intuitive to the general user (e.g. much depends on the user's skill for drawing or whistling).

General cross-media queries are textual in nature, as text is considered the easiest media for a computer system to handle. In order that queries are then matched to media, the media objects are manually textually annotated. Then established text-matching algorithms are applicable to the multimedia retrieval. This additional annotation of data is often referred to as 'metadata', which means 'data about data'.

The metadata approach is the most common approach to multimedia retrieval. In annotated systems how the user forms the query can be very significant in determining the success of the retrieval, both in terms of the ambiguity of natural language and that the user may be unaware of how the media has been annotated. Annotated systems are also not aware of the broader meaning of the terms used in their metadata vocabulary e.g. that the keyword "Ford Orion" is a specific instance of a "car", which is a "vehicle". Hence retrieval is rather coarse e.g. only media with the exact annotation searched for is returned, rather than with other, similar, media.

The InfoHarness system [Shklar,1995] is an early prototype example of the use of metadata to aid in the retrieval of data from heterogeneous sources. In particular, the growth of content access as a result of the Web led to the need for a single tool to retrieve and handle heterogeneous information. The metadata is derived from analysis of the provided data. It includes *content-based* metadata such as the document vectors from semantic indexing, and *content-descriptive* metadata such as the location and size of a document or domain-specific attempts to capture the semantic meaning of a document. The system used the technique of "Latent Semantic Indexing" [Deerwater, 1990] to analyse natural language and associate related words.

A concept-based retrieval method was presented by Di Nubila et al [Di Nubila,1994], where textual descriptions of radiological slides were used to generate conceptual annotations for use in a multimedia retrieval system. Similar techniques have been used in later systems to improve retrieval through annotations generated from text

associated with non-textual media (e.g. WebSEEk [Smith,1996], AMORE [Mukherjea,1997]). Latent Semantic Indexing has also been applied to text associated to media items [Sclaroff,1999]. This type of approach is still very dominant on the Web, e.g. Google Image Search (possibly the most used image retrieval system on the Web at the time of writing) associates images with the text closest to them on the HTML page[16]. In all these cases the metadata are determinable only as a result of there already being natural language text associated with the media.

Chen et al [Chen,1994] discuss automated methods for metadata extraction from images and speech for use in mixed media retrieval. Much of their work is based around keyword spotting, which means that the keywords likely to occur in the media must be known in advance, and there is no allowance for linguistic nuances (e.g. use of synonyms or alternative grammatical forms). Their image analysis is limited to the recognition of text in an image.

Without methodologies to enable more automated metadata extraction, annotation requires a major, often prohibitive, authoring effort. Research done into automating the annotation of media has demonstrated some success in determining subjects from the analysis of low-level features of audio, images or video but tends to function best when the entire media set belongs to a narrow subject domain. In general, the task of extracting subjects from media is considered to be too complex for an entirely automated process [Smeaton, 2000].

Regardless, research on semi-automatic media annotation continues to date [Wang,2001; Li,2003; Hove,2004]. The general approach is to feed systems with *prototypical* occurrences of an object in the media and that further instances of the same object can be identified by the system through *low-level feature similarity*.

Current levels of accuracy in this work require that there is still subsequently a manual examination of the resulting annotation. However, progress does indicate that the task of creating meaningful annotations of media is gradually being better supported by information extraction systems.

Presently the aceMedia project is working on a knowledge-assisted analysis (KAA) platform which aims to advance the state of the art in this field [Athanasiadis,2005]. The interaction between the analysis algorithms and the knowledge is continuous and tightly integrated, instead of being just a pre- or post-processing step in the overall architecture (see Figure 2.1).

Multimedia documents are segmented (spatially and temporally) into regions and a set of concepts are assigned to each region together with a measurement of certainty. This assignment is based on low-level feature comparison between the regions and prototypical descriptors for a set of concepts.

[16] http://images.google.com From the FAQ "How does image search work?": "Google analyzes the text on the page adjacent to the image, the image caption and dozens of other factors to determine the image content…"

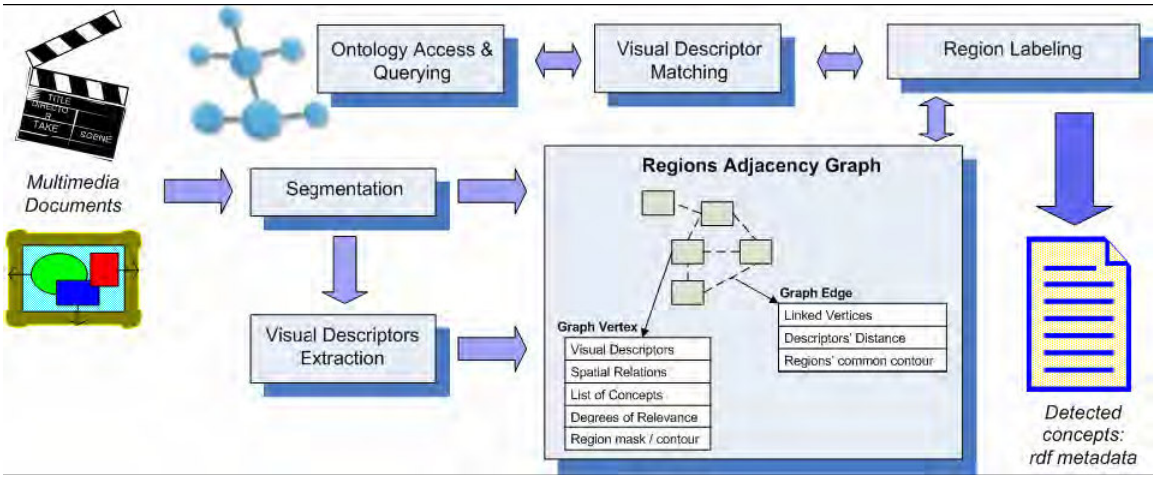

Figure 2.1 KAA platform architecture from aceMedia

There is also a person detection and identification module which represents the current state-of-the-art in person detection. The detector performs as much as 50 times better than previous state of the art detectors in evaluation experiments. The module uses a new paradigm by mapping images in a very high dimensional feature space – a feature space specially designed to reliably detect people irrespective of their clothing, poses, appearance, image background, and image illumination. Figure 2.2 illustrates the different aspects of detection.

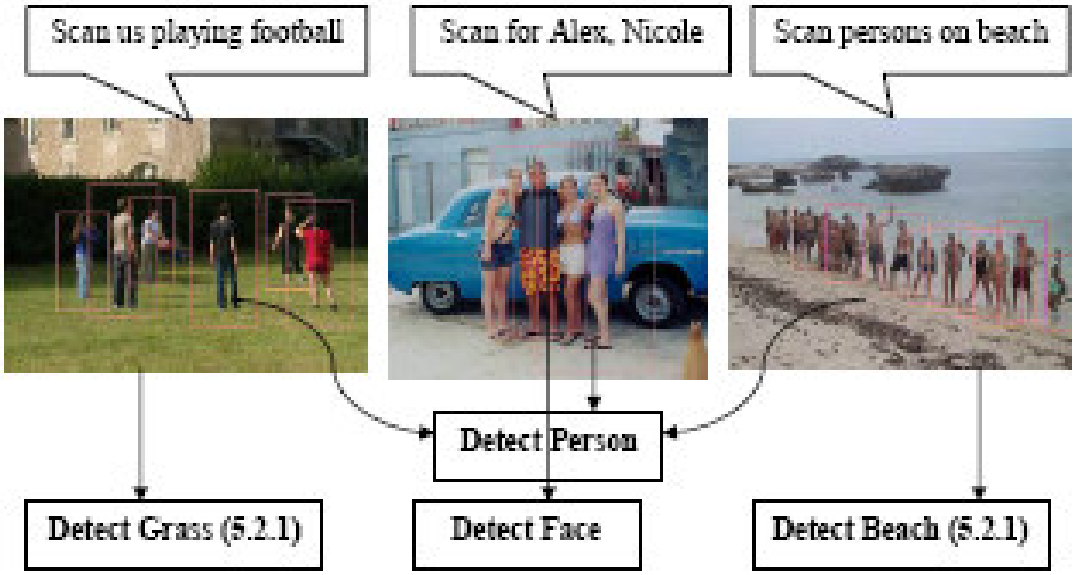

Figure 2.2 Example of object and person detection

Cutting edge research such as this represent the viability for the automated generation of annotations of persons, objects and events in visual media which could potentially support future scenarios such as the family tree generator mentioned in this thesis. Other emerging approaches support combining low level and concept-based multimedia search to improve access to this annotated media [Aurnhammer, 2006].

2.1.2 Adaptation

The issue of multimedia content adaptation means ensuring that a multimedia presentation fits to the context of its consumption while continuing to communicate its intended message. This differs from the adaptation of single media objects in that the message of a multimedia presentation is also carried in the relationships of the different media to one another, whether in terms of their spatial and temporal organisation, logical representation or interactive characteristics.

Before adaptation is possible, a computer system must be able to be aware of the context to which the content is to be adapted and the relationships through which the message of the multimedia presentation is communicated. Research issues focus on relational models for representing the multimedia presentation, as well as for defining the adaptation process in terms of its interaction with available contextual information to modify that model.

Unlike composite multimedia formats which do not allow a system to differentiate between the different media items (e.g. MPEG-2 with audio, video and text), XML [W3C, 2000] could be used as the basis for the representation of a multimedia presentation in a standardized declarative interchange format, with URL references to individual media items irrespective of their individual content encodings.

The SMIL [W3C,1998; W3C,2001] standard arose out of the W3C's work effort to declaratively model multimedia presentations. SMIL represents a multimedia presentation within two sub-trees: one defines the spatial organization through the definition of regions and the other the temporal organization through the ordering of the media objects within temporal elements (parallel, sequential, and exclusive). The elements for the media objects associate those objects with a spatial region as well as specify logical and interactive characteristics. SMIL's advantage is that it is relatively simple and hence can be quickly learnt and applied (as benefited HTML and hence the development of the Web), but this is also its disadvantage. Its structured XML model is not very flexible and the only allowance for adaptability is through a 'switch' element which selects a single child based on a specified attribute (the specification defines a set of standard attributes, for example system language and network bit rate).

An alternative declarative approach is ISO HyTime (Hypermedia/Time-based Structuring Language) [ISO,1997]. It defines a syntax for enriching SGML document type definitions (SGML being the more complex predecessor to XML) with well-defined multimedia semantics. The main primitives in HyTime are "Architectural Forms", which provide a structural description of multimedia content. A SGML element can be associated to an architectural form through the DTD, and hence inherit the associated multimedia semantics. HyTime offers a tightly defined co-ordinate space for spatio-temporal modelling and a location scheme for referencing HyTime documents at any level of granularity. However it is weak in interactivity (as it requires all spatio-temporal positioning to be known in advance) and dynamic adaptability (as adaptation parameters can only be available as concrete instances within the HyTime document).

ISO's MHEG-5 standard [Joseph,1995] is also a multimedia document model, but takes as its basis an object-oriented approach. It defines a hierarchy of classes which represent the media items, their properties (e.g. spatial and temporal position), actions and events. Class instances are linked together by event-condition-action rules which realize the behaviour of the multimedia presentation. This model, like SMIL, is at the level of the final form presentation, hence restricting its flexibility in realizing alternatives for different contexts. Media items are organized into groups, which are tied to a set of properties, and scenes which may also reference global objects (declared in a MHEG-5 document and re-usable across scenes). There is no standardized means of identifying and selecting groups, and no provision for associating metadata to MHEG-5 objects. Adaptation is supported through classes which represent variables whose value can be tested. Through this testing, different branches of the presentation model can be selected. As variables can only be set from within the document, a new document would be required for each context. This is overcome by MHEG-6 which defines an interface between a MHEG-5 presentation engine and Java applets, so that e.g., an applet could be called at execution to determine the actual value of the contextual variables. It is still the case on the other hand that all the variables for adaptation must be pre-instantiated in the document.

	Requirements	HTML	DHTML	SMIL	MHEG-5	HyTime
Basic Requirements	**Temporal Model**	-	script	interval-based	event-based	point-based
	Spatial Model	absolute positioning	absolute positioning	absolute positioning	absolute positioning	absolute positioning
	Interaction					
	Navigational	+	+	+	+	-
	Design	-	+	-	+	-
Advanced Requirements	**Reusability**					
	Granularity					
	Media Elements	+	+	+	+	+
	Fragments	-	-	-	-	+
	Documents	+	+	+	+	+
	Kind of Reusage					
	Identical	+	+	+	+	-
	Structural	-	-	-	-	+
	Identification/Selection	+	+	+	-	+
	Adaptation					
	Parameters of Adaptability					
	User Interest	-	+	-	MHEG-6	-
	Technical Infrastructure	-	+	+	MHEG-6	-
	Definition of Alternatives					
	Static	-	+	+	MHEG-6	-
	Dynamic	-	-	-	-	-
	Presentation-neutral Representation					
	Multimedia Functionality	very low	high	medium	very high	low
	Semantic Level	medium	very low	medium	low	very high

Figure 2.3 Comparison of Web multimedia models [Boll,2001]

[Boll,1999] examines these multimedia document models and evaluates them on the basis of a set of requirements derived from foreseen future multimedia system needs. These requirements are reusability, adaptability and widespread usability. The conclusion is that neither SMIL, MHEG-5 nor HyTime offer sufficient support for all three requirements. Figure 2.3 shows an overview of the study.

In the context of the project “Gallery of Cardiac Surgery” the ZyX model [Boll,1999b] is developed to specifically meet these future requirements. ZyX models a multimedia presentation in a SMIL-like tree structure with the difference that leaves can be left unbound. Connections can be dynamically made by a processing system in response to a specific need. A specific adaptation approach is also introduced in connection with this model [Boll,1999c]. It is noted that multimedia adaptation often concerns itself with the “lowest common denominator” i.e. single media elements. Either same media is offered in different forms (e.g. different quality according to network bandwidth or different language according to user preference) or media selection concerns itself with a single “match” (as in SMILs “switch”). The authors introduce *cross-media adaptation* which allows alternatives to be different media types or groups of media. This introduces the challenge of maintaining the message of the presentation.

Two formal models are specified to ensure that media alternatives preserve the presentation semantics and a correct flow of information. *Augmentation models* verify the semantic equivalence of automatically determined media alternatives. *Substitution models* control the degree of adaptation in terms of “closeness” to the original presentation flow. The media augmentation process queries an underlying database exploiting inherent technical data and metadata annotations to determine potential alternatives. Media equivalence is defined by structural tests and subject matches where subjects are annotated to individual media objects. Subject-temporal relationships are also modelled as the subject may only relate to a temporal sub-section of the media. The substitution process retrieves the current presentation state and the user context. It comprises a set of constraints that determine which potential alternatives are permissible for presentation.

The approach is aimed to relieve authors from explicitly, comprehensively specifying all content alternatives in a presentation. The ZyX model makes it possible to modify the presentation in a deeper way than the SMIL switch element. For example, groups of media objects can be handled as single composite objects. However the approach is ambiguous in its implementation. Subject equivalence will only succeed through exact matching of textual annotations, without allowance for any flexibility in the rules application (e.g. similarity relations, or differentiation between classes and instances). Media from heterogeneous sources may need to be re-annotated according to a centralized terminology. It also seems unlikely that structural tests alone (e.g. having the same duration) can offer a guarantee of media equivalence.

While ZyX is intended to meet the stated requirements of re-usability, adaptability and widespread usability, the inherent flexibility of its model is insufficient. Means must also be defined of how the process of adapting and realizing a multimedia presentation using that model is governed. In the context of a ZyX-based multimedia system, the system implementations specification of adaptation rules is fixed to a certain metadata vocabulary and is internally expressed in the system logic, unlike e.g. XSLT-based approaches. The system can not be flexibly changed to work with other content which is annotated differently or in other contexts where different adaptation rules are required.

Like ZyX, the MATN [Chen,1999] work is based on a more flexible multimedia presentation model, in this case based on an abstract semantic model called

augmented transition networks [Woods,1970]. Augmented transition networks are a procedural approach to representing syntactic facts about natural language.

It claims to offer simplicity and ease of modification by representing the multimedia presentation through *state transition graphs* and regular expression like grammars. It also claims to be scalable as it can model user interactions in a single framework unlike timeline models that require several timelines to model the same situation [Chen,2000].

The input to the multimedia augmented transition network (MATN) shall be a "multimedia input string" representing the multimedia presentation sequence. It consists of a set of states and directed arcs. States represent changes in a finite set of nodes, representing media streams. This change can be the ending of an active stream or beginning of a new one. Arcs represent permissible transitions from one state to another, which are labelled by the transition function. A condition/action table is defined externally from the input string to specify the rules for transversing from start to end. Sub-networks model media stream information, either the spatio-temporal relations of semantic objects occurring in image or video media or keywords identifying semantic objects in a text media stream.

Hence input strings have two levels of representation:

- Coarse-grained level - modelling the media streams
- Fine-grained level - modelling the semantic objects occurring in the media streams and their attributes.

Assessing this approach, we note that multimedia input strings and sub-networks are created by the designer in advance for a class of applications. The semantics expressed within the network can be used for multimedia database searching on semantics, and for multimedia browsing in response to user actions. Hence the model is a 'closed' one, that is, there is no means to dynamically integrate additional resources into the model. This restricts the possible adaptability to that of retracing arcs in the model – alternative representations must already be modelled and can not take into account external factors acquired dynamically at runtime such as information about the user or device. Unlike other multimedia models, which have dealt purely with syntactic representation, MATN also seeks to support a semantic representation of the content. Semantic objects can be modelled with descriptive keywords to aid the user search and associated to their spatial and temporal attributes in media streams. However, it seems that object identification is proprietary to individual implementations, meaning that semantically-based applications must be designed specifically to their MATN-based models. Augmented transition networks do not have a declarative format, restricting interchange and interoperability, and are not formally well defined. This suggests that MATN-based presentations may not adapt well to working with distributed content on the Web, though work has been done to model between MATN and SMIL [Chen, 2002].

Adaptation plays an important role for a system which may be required to present multimodal information in a wide variety of contexts. In the scenarios we have presented, for example, there have been differences identified in device (mobile, television), user (interests, location) and usage (background information, immediate need).

2.1.3 Presentation

Computer systems have been long employed in the representation of text, hence models for textual layout are well developed. Such models are widely implemented, for example, in e-publishing or Web browsers. These models define the formatting of text within specified regions, including in relation to non-textual objects within the same region. On the basis of this definition, systems can handle the re-formatting of text when textual or non-textual media is inserted, moved or removed.

Approaches to automatically integrate media into a common presentation document are already common on the Web. Using HTML as the ubiquitous base format, templates are filled with content from a database back-end or style sheets (CSS, XSLT) are applied to presentation-neutral content to define the final layout and style. However, the result HTML document is presented according to textual layout models. While mixed media layout issues can be handled in cases where the characteristics of the media is known in advance, for the on-the-fly integration of multiple media objects of different types in a synchronized, interactive presentation, the models developed for text can not be readily applied [van Ossenbruggen, 2001]. This is because multimedia differs fundamentally from text. Van Ossenbruggen et al. give a good summary of these fundamental differences:

1. Multimedia uses different document and presentation abstractions : while text has a widely accepted set of abstractions on both levels (e.g. document chapters, sections, headings, titles and presentation pages, columns, tables, lists), multimedia does not and is not likely to develop one.
2. Multimedia document formatting is not based on text flow : the linear flow of text on a page is flexible enough to allow for well established rules for its alteration upon changes in layout requirements (e.g. line and page breaking) but spatio-temporal positioning of objects in multimedia presentations can not be so flexible.
3. Multimedia transformations need feedback from the formatting back-end : the flexibility of text flow means that generally reformatting can always be handled, but the success of a reformatting operation in multimedia is not as certain as multimedia presentations have tighter constraints on layout.
4. Multimedia transformations are hard to describe in a functional language : transformation languages such as XSLT process the source document in a linear manner yet alterations to multimedia will often require backtracking to reprocess earlier elements.

These differences illustrate the need for research into systems which utilise a presentation model tailored to the needs of multimedia generation, as opposed to the well established textual layout model or single media-based solutions.

The multimedia layout manager LayLab [Graf,1996] treats the layout issue as a multi-dimensional constraint problem and is based around a dedicated constraint solver. This approach led to the proposal of a generic model for an IMMLM (Intelligent Multimedia Layout Manager) [Graf,1996b] which would be integrated within an IMMPS (Intelligent Multimedia Presentation System) [see Chapter 2.5]. The LayLab layout manager was incorporated into the task-specific multimedia systems PPP [Andrè,1996] and WIP [Andrè,1993]. Layout results in these systems are determined by the internal algorithms of the layout manager and are not accessible to or modifiable by an implementer. The manager also operates according to a general set

of constraints which, despite being applied to systems generating multimedia for specific tasks, can not be extended to support making layout decisions appropriate to those particular tasks.

The Delauney system [Cruz,1997] was intended as a framework for non-professional users to not only select heterogeneous multimedia content but also to specify its presentation in a domain-independent manner. It was implemented as an extension to an interactive, constraint-based system for visualising object-oriented databases. A virtual document is defined as a set of user-specified style sheets, each of which is tied to a query. These queries can be applied to a database or to Web documents, and must draw upon the vocabulary of the content source. In the case of the Web, basic low-level metadata attributes are added to documents in the search to improve the query results.

Clearly the system is an attempt to involve users in the specification of the desired multimedia layout. By abstracting layouts into external templates, different templates could be provided according to user and task. These templates contain placeholders for the media retrieved on the basis of an associated query and constraint specifications such as minimum/maximum size and overlap. The final layout is realized through the media retrieval and a constraint solver [Cruz, 1998].

However, as a multimedia system, the layout model does not address the temporal aspect. It also has a limited and proprietary specification. While simple alterations to a specified Delaunay-based application is straightforward enough with the graphical tool for layout modelling, the task of appropriate layout specification will depend much on the users understanding of the target content sources, both in terms of an appropriate query formulation and an appropriate layout for the quantity and form of the media that will be returned. The need for user specification of queries and layout limits the ‘automated’ aspect of the system to repeating the same content retrieval on the same layout, and restricts adaptability to the manual alteration of templates for changes in user or task.

Our scenarios raise the requirement to present information spatially (family tree) as well as temporally (during a television program). Furthermore, interactivity needs to be supported within the presentation to allow users to choose between options.

2.1.4 Summary

This overview of the multimedia generation process has considered three primary, interdependent tasks which realise the process, noting their recognised problems identified in research and from prototype implementations. These problems must be solved if the scenarios that have been given are to be realised by a multimedia presentation system.

Multimedia retrieval is realised by textually annotating media to be retrieved, yet this approach involves a large non-scalable manual effort and is generally not co-ordinated, resulting in non-re-usable application-specific annotations which still exhibit the ambiguities of natural language. Tools are emerging for semi-automatic annotation of media based on overcoming the “semantic gap” (from low level features to high level concepts).

Multimedia adaptation is facilitated by flexible multimedia presentation models, but these models do not in themselves support a high level of expressiveness for contextual representation and implementations must still resolve how the contextual representation is produced.

Multimedia presentation is found to be not as trivial as text, as multimedia has different layout requirements which must be resolved over a linear time-flow and respect the underlying meaning of the presentation.

More recent research has referred to the knowledge modelling field as a potential solution to these problems (we will introduce this research in chapter 3). To understand the application of knowledge modelling to multimedia presentation, we introduce knowledge representation theory and its' proposed application on the Web, called the "Semantic Web", and then consider uses of conceptual models and knowledge representations for multimedia complementary to the Semantic Web.

2.2 Knowledge representation theory

The purpose of knowledge representation (KR) theory is to enable the storage and manipulation of knowledge about the world in a formal, machine-processable model [Davis,1993]. Such knowledge models could then be leveraged by a computer application to make 'intelligent' decisions. Knowledge representation is a central problem of Artificial Intelligent research.

An early representational method for knowledge was the **semantic net(work)**. The "semantic net" was introduced as a method of modelling the structure and storage of human knowledge in the shape of a graph [Quillian,1968]. It takes the form of a directed graph, consisting of vertices (concepts) and arcs (relations). The most important relations between concepts are *subclass* relations between classes and subclasses and *instance* relations between particular concepts and their parent class (the 'IS-A' relation). These relations formed the basis for the rule of *inheritance* - that relations are inherited by subclasses from their superclass, and by instances from their parent. It was popular as it had an accessible graphical notation, but it lacked formal semantics which were necessary to support reasoning.

J F Sowa developed **conceptual graphs** [Sowa, 1976] as a means of graphically representing logical propositions in a semantic network-like form. Unlike semantic networks, conceptual graphs are underpinned by a formal logic and hence can be expressed e.g. by predicate calculus. Figure 2.4 shows the conceptual graph for the proposition 'A cat is on the mat':

Figure 2.4 Conceptual graph for 'A cat is on the mat'[17]

[17] Diagram and CGIF/KIF syntax courtesy J.F. Sowa, "Conceptual Graph Examples" http://www.jfsowa.com/cg/cgexamp.htm

In order to support the interchange of conceptual graphs between computer systems, a machine syntax is specified known as CGIF (conceptual graph interchange format)[18]. To provide interoperability with other internal knowledge representations, there is also a logic-based formalism called KIF (knowledge interchange format)[19]. The above proposition is expressed in these two formats so:

CGIF : `[Cat: *x] [Mat: *y] (On ?x ?y)`
KIF : `(exists ((?x Cat) (?y Mat)) (On ?x ?y))`

First-order logics are typically used as the mathematical basis to avoid excessive complexity and to assure the exactness of assertions. These interpret statements as predicates and constants e.g.

John and Mary are siblings → `Siblings(John,Mary)`

Another early method was **knowledge frames** [Minsky,1975]. Frames represent concepts and consist of slots which represent attributes and take values. In other words, all knowledge about a concept is stored within its frame. Frames look like typical database entries, but also support inheritance. Additionally, frames were not just declarative like semantic nets but could also contain procedurally expressed knowledge so that frame-based systems could call on procedures to decide how to use the frame, fill in values etc. While their logical basis offered no new expressiveness over first-order logic-based systems, they offered a concise way to express knowledge in an *object-oriented* way and used only a fragment of first order logic, offering greater decidability and more efficient means for reasoning.

Description logics [Nardi, 2003] formalize this declarative part of frame-based systems. The first Description Logic based system was KL-ONE [Brachman, 1985]. They draw upon object-oriented approaches and are expressed as a fragment of first-order logics. *Classes*, or sets of individuals, are modelled as unary predicates and *roles*, or relationships between individuals, are modelled as binary predicates. Roles can also carry value restrictions (a limit of the range of values that can be applied to that role) and cardinality constraints (constraints on how many values may be applied to a role). A description logic-based knowledge representation is considered as consisting of two components:

- The “Tbox” which specifies the general knowledge of the domain in the form of a terminology
- The “Abox” which specifies the knowledge specific to the individuals of the domain

The “Tbox” is considered to be static (unchanging over time) while the “Abox” is expected to be contingent to a particular set of circumstances, and hence subject to regular modification.

As an alternative to the frame-based approach, languages are proposed for representing knowledge for computer systems. These languages focus more on machine processability than on human readability. The low-level syntax is most

[18] ISO/JTC1/SC 32/WG2 document ‘Conceptual Graphs’ by J.F. Sowa http://www.jfsowa.com/cg/cgstand.htm
[19] ‘Knowledge Interchange Format’ – draft proposed American National Standard http://logic.stanford.edu/kif/dpans.html

commonly expressed in XML to support parsing and document interchange. They can be based on different formalisms but primarily draw on the concepts of the object-oriented model, and hence use description logics.

Some of these languages are designed to model the "Tbox" of description logic-based models and some the "Abox". The "Tbox" is often referred to in computer science as an **ontology,** and the modelling languages are called **ontology languages**. An instance of an "Abox" is commonly referred to as a **knowledge base** and it references and depends upon one or more ontologies from which classes and roles are drawn.

Ontologies are generally hierarchical data structures which organise all relevant concepts and their roles and rules in a domain. They are *an explicit specification of a conceptualization* [Gruber,1993]. They can be used to define the permissible set of concepts for an application („controlled vocabulary"), to classify concepts ("taxonomy"), and to relate concepts to associated and related terms ("thesaurus"), though description logic-based ontologies are more expressive than any of these.

Examples of early ontology languages are CycL[20] [Lenat,1988], a first-order logic-based language originally used with the CyC project and LOOM [MacGregor,1987] CyC attempts to build an ontology of everyday common-sense knowledge with the goal of enabling AI applications to perform human-like reasoning. Ontologies can also be represented graphically through the Entity-Relationship model [Chen,1976]. Other research has also considered the applicability of UML Class diagrams[22] [Cranefield,1999].

As ontologies can express rich semantic notions, they are often thought of as a knowledge representation model in and of themselves - *a computational model of some portion of the world* [Huhns,1997]. An ontology, being machine readable and intended to be static, allows applications to be standardized while the instance-specific information (the "Abox") is modified over time. Through declarative representations, they facilitate moving the complexity of a system from the application that processes the information into how the information is organized.

2.3 Knowledge representation for the Web

The "Semantic Web" was first envisioned by the Web's founder, Tim Berners-Lee [Berners-Lee,2001]. It is promoted as an extension of the current Web that will allow human users and software agents to find, share and combine information more easily. It does this by relying on machine-processable metadata and applications designed to utilise that metadata. The Semantic Web layer cake proposed by the W3C is shown in Figure 2.5.

[20] See also http://www.opencyc.org
[22] see Unified Modeling Language, Version 1.5 http://www.omg.org/technology/documents/formal/uml.htm

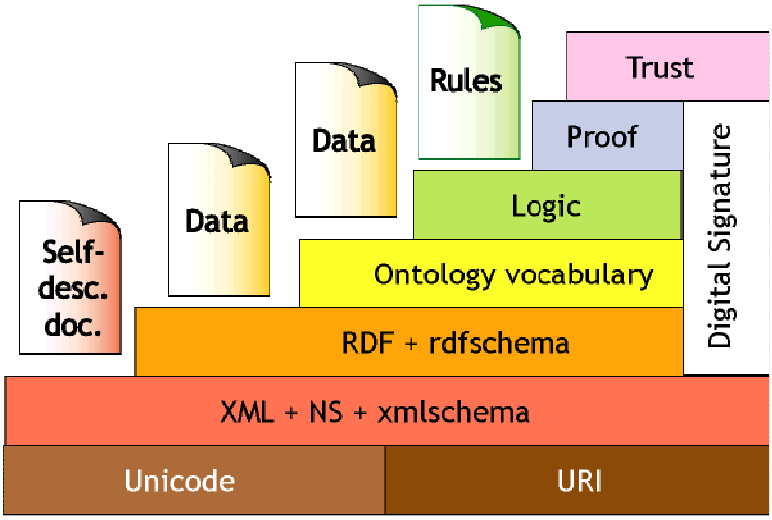

Figure 2.5 The Semantic Web layer cake[23]

Metadata in the form of the Resource Description Framework (RDF) [W3C,1999] is built upon the existing infrastructure of the Web, particularly XML as an extensible and interchangeable syntax for specifying the metadata and the URI scheme as an extensible and standardized means to unambiguously address the entities and relationships contained by the metadata.

The metadata is itself built upon by an ontological layer which can further qualify the metadata in terms of the known concepts in the metadata's domain, their relationships and the constraints upon them. Two independently developed ontology languages for the Web, DAML (inspired by object-oriented approaches and frame-based systems) and OIL (closer to description logics) [Fensel,2001], were merged into DAML+OIL [McGuinness,2002] and now form the basis for the W3Cs Web Ontology Language, OWL [W3C,2004].

RDF developed out of earlier work on the Meta Content Framework (MCF) [Guha,1996]. MCF is modelled by directed labelled graphs consisting of arcs where each arc is a triple consisting of two nodes and a label. Nodes are intended to represent anything that a human may wish to conceptualise within a computer system, and arcs represent their characteristics and relationships with other nodes. The label is the property-type, which is also a node. Nodes can be primitive data types or 'objects' (much like objects in an OO language). The model also includes some standard KR concepts such as object instantiation and typing (class-instance, subclass-superclass).

The RDF model is centred around resources, which are equivalent to MCF's nodes. Resources are identified in RDF by URIs. In the Semantic Web vision, anything can have an URI, regardless or not of whether that thing is a resource existing on the Web. So in principle a RDF resource can represent anything as long as it is associated to an URI. However, given its Web-oriented background, RDF is often understood to be aimed at describing addressable resources - the only form of resource whose URI can unambiguously identify it.

[23] Copyright World Wide Web Consortium. All Rights Reserved. Legal notice: http://www.w3.org/Consortium/Legal/2002/copyright-documents-20021231

RDF descriptions are expressed in the form of statements known as triples, as they consist of a subject, a predicate (the 'verb' in the statement) and an object. These are all represented by resources with the exception that the object can alternatively be a literal (i.e. a data type instance, such as a string). A sample RDF graph is shown below (Fig. 2.6), showing a subject resource with the local id 'me', four predicates and their values, which are two object resources and two literals. The top-most predicate is defined by the RDF specification for expressing class-instance relationships.

Figure 2.6 A sample RDF graph[24]

To support more complex statements, the RDF model also incorporates containers (so that a group of resources can play the role of a single object in a statement) and empty nodes (which represents an abstract resource in a statement). Statements can also be reified so that they can be the subject of another statement (reify = to regard something abstract as a concrete thing).

Knowledge expressed in the RDF model can also be serialised as XML, which is quite verbose but supports data interchange and integration. As a result, the W3C also promotes the use of RDF within XML resources such as XHTML and SMIL documents.

The ontological layer on top of RDF plays two roles in enabling and validating RDF-based metadata. Firstly, it implements a controlled vocabulary, defining the class hierarchy existing within a certain domain. Secondly, it specifies constraints on this domain, such as restrictions on property values. The RDF specification included a

[24] From the RDF primer http://www.w3.org/TR/rdf-primer/ See previous footnote for the required legal notice.

simple ontology language called RDF Schema (RDFS). However, while this supported simplified ontology building, RDFS was not very powerful. To meet needs for more powerful ontological expressability, OWL was developed.

A key task of ontologies in the RDF model is to provide a basis for defining what URIs identify. The meaning of URIs has grown into an on-going debate in the RDF community, as the specification does not provide for differentiation between different uses of URIs in identification. Four distinct uses of an URI have been identified: as a name, as an abstract concept, as a Web location and as a document instance [Booth,2003].

RDF can provide human-interpretable identity to URIs ("label" and "isDefinedBy" properties) but machine-interpretable identity can only be defined in terms of permitted properties and relation to other resources (e.g. equivalence, disjunction). As a result, RDF has the problem that processing applications need to be aware of the vocabulary being used by the RDF metadata, or have access to an ontology which can relate the vocabulary to another vocabulary that is known by the application. In a decentralised network such as the Web, a lot of weight is being attached to RDF vocabulary sharing and interoperability if intelligent Web-based applications are to be able to manage heterogeneous RDF metadata as an unified whole.

The RDF approach is close to that of semantic graphs. Work on formalizing its semantics was standardized by the W3C[25]. OWL(-Full) is specified as an extension of RDF, which means it can support the loose formal semantics of RDF (e.g. a class may be an instance of another class), though there are two constrained versions with more formal semantics: one which conforms to description logics, OWL-DL, and an even more constrained version for applications which do not need the extra expressiveness, OWL-Light.

2.4 Knowledge representation and multimedia

"Second generation" multimedia has been examined in the past sections (2.1): multimedia presentation generation on the basis of static templates and rules which process syntactic multimedia models (such as the second generation web is underpinned by back-end data storage and its dynamic inclusion into HTML templates through technologies like XSLT and JavaServerPages). Knowledge representation techniques which facilitate the generation of data that is machine-processable were then described (2.2). This is the underpinning of the third generation Web, the "Semantic Web" (2.3). Van Ossenbruggen describes "third generation multimedia" [van Ossenbruggen, 2001] as the next development in multimedia technologies. This is a requisite for "intelligent" systems that will provide content in a more dynamic and meaningful way to consumers by building knowledge representations into multimedia systems to support an "intelligent" processing of multimedia content for presentation. This section will consider the work that has been done in the field of knowledge representation and multimedia, before the chapter

[25] 'RDF Semantics' was published as a W3C Recommendation, 10 February 2004 http://www.w3.org/TR/rdf-mt/

closes with a deeper examination of a particular proposal for an intelligent multimedia presentation system.

2.4.1 Application-internal representations

Earlier multimedia systems have used models and conceptual representations for the participating multimedia content prior to the popularisation of knowledge representation techniques through the Semantic Web. We introduce a few of these systems.

Semantic modelling with images was used in VIMSYS [Gupta,1991]. It used a four layer model for picture description, which includes an image representation layer, an image object layer, an semantic object layer and a semantic event layer. Semantic objects are expressed as subclasses of image objects (which are low-level feature descriptions), and are associated to attributes and methods for retrieving those attributes. Semantic events can express spatial and temporal relations of semantic objects. This knowledge is used primarily to refine user queries on an image database, and improve upon similarity-based retrieval.

The SEQUOIA 2000 Project [Anderson,1994] drew up metadata specific for use with describing satellite images, defining means to associate spatial and temporal characteristics to them. The metadata basis was a relational database queried by SQL, not a KR system.

ViMod [Jain,1994] is a metadata model specifically for video databases. The authors identify four design goals for this model:

- Support for the temporal nature of video
- Accommodate all types of video data application
- Allow for reuse of video data in a different context
- Support both exact match and similarity based queries

Video is modelled as a sequence of temporal intervals, where for each a number of features can be represented or highlighted. The defined feature classes are mostly low level though a few high level features do occur (featured objects, object properties, content timeframe, content classification). For these, the authors acknowledged the need for terminologies to standardize feature values.

A Content-Based Hypermedia system [Grosky,1994] proposes a more KR-like object-oriented metadata schema, with *is-a, is-part-of* and association relationships. A distinction is made between a set of semantic objects, representing real world concepts, and media objects, representing digitized content (in this system, either images or video). Semantic objects could have an *appearing-in* association to media objects or sub-regions of those media objects, with the equivalent inverse association *represents*. This metadata forms the basis for browsing a 'hypermedia Web', where media objects (regions) are hyperlinked to other objects (regions) representing the same semantic object or related semantic objects. While a model vocabulary is introduced which expresses the association of media and concepts extended to support media segmentation, the model lacks a formal basis for defining its instances. It is unclear how media and concept shall be referenced, and if those references could also be validly used by other systems outside of the instantiating system.

The Multimedia Objects Server SOMm [Vieira,1999] is an object-oriented DBMS based on the structure of the multimedia document model MHEG-5 with extensions to support content-based search.

For each media item in the MHEG-5 model, classes are added for semantic information relating to that item. The model supports the provision of semantic terms with characteristics and relationships to other terms (a thesaurus). Terms also have identifying names and identification of the media in which the term is represented. This model demonstrates adding semantic information to a multimedia representation, but it is tightly integrated with the MHEG-5 format and is hence unsuitable for modelling multimedia knowledge independent of its final format realisation.

Such use of proprietary and solitary models to represent multimedia content within systems has been replaced in recent years by the growth of standards in the area of multimedia content representation, to which we now turn.

2.4.2 Standards for multimedia content representation

Bohm & Rakow specified requirements for metadata intended specifically for multimedia documents [Bohm,1994]. They list six categories:

- Metadata for the representation of media types e.g. format, encoding, compression technique
- Content-descriptive metadata e.g. objects identifiable in or concepts represented by media
- Content classification metadata e.g. subject area, level of detail, accessibility guidelines
- Document composition metadata e.g. relationships between document components, “role” of media
- Metadata for document history e.g. usage, previous changes
- Metadata for document location e.g. media reference

They also note the possible use of SGML for the representation of metadata. They do not, however, define any metadata vocabulary themselves.

Hunter & Iannella summarize the results of a Resource Discovery Workshop on Moving Image Resources which took place in 1997 in Bath, England [Hunter,1998]. The workshop reviewed the potential of the Dublin Core vocabulary [DCMI,1998] for describing moving image resources. While Dublin Core as a metadata vocabulary is specifically designed for describing textual documents, the workshop proposes a set of sub elements and schemes specific to audiovisual data as an extension of the vocabulary. The Dublin Core descriptions can then be expressed in the RDF model. Applying Dublin Core has the benefits that the base vocabulary is simple (15 elements) and already established in the Digital Library domain. However, the proposed extensions lead to greater complexity and a loss of information in data interchange (as other Dublin Core systems will not recognise the extensions). It is also noted that expressing multimedia structure in RDF while maintaining the Dublin Core semantics is problematic, as RDF syntax is not designed to reflect the spatio-temporal model of complex multimedia content. There are problems in terms of expressing collections of elements, pointers to media segments and specifying media synchronisation. One proposal is the use of RDF for the metadata description with SMIL for expressing the media structure.

The most comprehensive attempt at modelling multimedia descriptions is the ISO MPEG-7 standard, formally known as 'Multimedia Content Description Interface' [ISO,2001]. It specifies a standardised set of tools to describe various types of multimedia information. This is intended to be independent of the media content itself. Valid descriptions can be serialized as XML or binary format.

It does this using a set of key concepts:
* Descriptors (D) – each descriptor represents a particular feature of multimedia information.
* Description Schemes (DS) – a schematic definition of the structure and semantics of the relationships between its elements, which can be descriptors or other description schemes.
* Description Definition Language (DDL) – the schematic language for formally specifying descriptors and description schemes. The ISO decided upon XML Schema with a few extensions.
* System tools – to support the efficient binary encoding, multiplexing, synchronization and transmission of the descriptions.

Content is described from five viewpoints by the standard [ISO,2000]:

- Creation & Production
- Media
- Usage
- Structural aspects
- Conceptual aspects

The first three viewpoints can be grouped under “Content Management” while the remaining two are “Content Description”. These can further be divided into low-level (structural) and high-level features (conceptual). The structural aspect description scheme supports the spatial, temporal and hierarchical decomposition of media, the expression of relationships between those decompositions and the precise description of their low-level features.

MPEG-7 is a large standard, in terms of the amount of descriptors and description schemes that are defined. We mention only the conceptual aspect description scheme in more depth, as it is the closest in intention to traditional KR usage. The standard describes the Semantics DS as a tool that can be used “to describe narrative worlds depicted in, or related to, multimedia content by describing objects, events, concepts, states, places and times in those narrative worlds”. A ‘narrative world’ is defined as “a context for a ... description ... that is, ... the ‘reality’ in which the description makes sense”. In other words, it seeks to model knowledge about a specific domain, which in this case is the ‘narrative world’ of the relevant multimedia content.

The semantic description permits the specification of the semantic entities of the narrative world and the semantic relations which associate those entities to one another and to segments of the digital media being described. The entities can be classified as objects (perceivable entities that exist in the narrative world), agent objects (objects which are persons, groups of persons or organisations), events (perceivable entities that take place in the narrative world), concepts (non-perceivable or generalisations of perceivable entities), semantic states (parametric

attributes of entities which change in space and time), semantic places (spatial location) and semantic times (temporal location). Entities can be described by a label, a textual annotation, a set of identifying properties and by reference to an occurrence in the digital media. They can also exist in binary relations with other entity classes, relations or media segments. These relations are defined in the standard, e.g. *hasAgentOf* expresses the relation of initiator between an object and an event, *hasMediaPerceptionOf* expresses the relation of depiction between an entity and a media segment. Each relation has an inverse, but the model does not consider relations as bidirectional, nor can they be extended or new relations be defined.

The standard also defines a description scheme which can model the set of entities for a particular domain. The Classification Scheme DS provides the basis for defining terms (the entities) and organising them in a classification scheme. The terms have unique identifiers by which they can be referred to from other descriptions. The scheme models relations between the terms, which can be 'preferred', 'broader', 'narrower' or generically 'related'. Hence a Classification Scheme is closest to the definition of thesaurus in the knowledge representation community.

A major criticism is that the MPEG-7 model uses a syntactic, rather than a semantic data model [Nack,2002]. It is formally expressed using an extended version of the XML Schema language. As MPEG-7 data is processable only in a syntactic (XML) manner, it is a serious issue that there can be variations in (valid, equivalent) XML serialisations, and that XML processing can not manipulate information (e.g. inferencing) as a semantic model could do. Arguably, the choice of XML was a pragmatic one, particularly as the work on Web-based ontology languages was in an early and changing phase at the time of MPEG-7's publication.

Its major contribution to the knowledge representation community has been to provide a comprehensive vocabulary for multimedia descriptions. However, that MPEG-7 has established itself in the multimedia community as the standard for representing multimedia content has had the effect of discouraging a quicker uptake of semantic approaches to model multimedia.

2.4.3 MPEG-7 and Knowledge Representation

The MPEG group recognised the need for a semantic basis to MPEG-7 as early as their 2001 Sydney meeting, establishing an Ad hoc Group for MPEG-7 Semantic Interoperability and mandating them to produce a MPEG-7 ontology [Hunter,2001]. RDF Schema was chosen for the ontological model and DAML+OIL was introduced to express some semantic constraints beyond what was possible with RDFS. Only a small subset of the MPEG-7 vocabulary was modelled in this work. The modelling is also by the authors' own admission subjective, in that they have had to make their own decisions in terms of interpreting the intended MPEG-7 semantics from its syntax and description.

The Harmony project has incorporated this work into part of a larger aim to make metadata from different communities interoperable [Hunter,2003]. Using the ABC Vocabulary [Lagoze,2001] as the core ontology, a 'super-ontology' has been produced, MetaNet [Hunter,2001b], which expresses semantic relationships (e.g. equivalence, narrower, broader) between the metadata terms from different domains. As a result, semantic mappings and merging between metadata from different

domains, including that expressed in MPEG-7, should be possible, as well as single searches across heterogeneous metadata. A number of projects are described which demonstrate the use of the multimedia ontologies in the tasks of query mediation, ontology harmonization and multimedia aggregation.

A formal semantic model is a requirement if MPEG-7 data is to be more rigidly expressed and usable in knowledge-based multimedia systems. Other researchers, acting independently of the ISO, have also created MPEG-7 based ontologies in their own work:

- The full Multimedia Description Scheme (MDS) as an OWL-DL ontology [Tsinaraki,2004]
- DMAG MPEG-7 ontology generated automatically by a XSD2OWL mapping which is in OWL Full [Garcia,2005]
- INA core ontology for audio-visual description is inspired by several terminologies such as MPEG-7, TV Anytime and ProgramGuideML [Isaac,2004]

However, none of this work alters the fact that MPEG-7 is itself a syntactic standard, that MPEG-7 tools work on this syntactic level and that MPEG-7 semantic models that are generated from existing syntactic MPEG-7 inherit the disadvantages of the syntactic approach, e.g. that sets of MPEG-7 data describing an equivalent feature, due to variation in their underlying XML syntax and its interpretation in the ontology, could have different semantic representations. Each effort has made its own modelling decisions in representing MPEG-7 within OWL, so these efforts in themselves do not replace the need for standardization in this area.

Also, the present focus on semantic interoperability addresses the metadata vocabulary at the “TBox” level – the classes and their properties - and not the handling of the “ABox” i.e. metadata instances – it may also be that in an equivalent description the metadata classes can be interpreted as equivalent and yet the described instances are not due to variation in the feature identification (textual vs. URI, or use of different controlled vocabularies). There is a need now to integrate the various approaches to ontologizing MPEG-7 so that a standard can arise which ensures data interoperability and tool production.

An alternative to modelling MPEG-7 through a knowledge representation approach is to integrate MPEG-7 descriptions with existing knowledge bases. The use of URIs in MPEG-7 descriptions is significant to this approach as the URI scheme is used as the identification scheme for concepts in the Semantic Web. Some MPEG-7 descriptors are of type *TermUseType* or *ControlledTermUseType* (in the DDL Schema), which means they take as a value an URI referring to a term in a Classification Scheme. The standard itself appears to suggest that the referenced scheme does not have to be MPEG-7, hence any controlled vocabulary may be considered valid. URI references using descriptors of these types (Label, Property, Structured Annotation) from the Semantic DS could be used to integrate MPEG-7 semantic entities with knowledge representation models. However, these references can not indicate the knowledge model which provides identification through this URI, or how the model shall be processed to determine that identification, and hence the interpretation of those references would be application-dependant. This is a major problem if this approach is to enable interoperability with knowledge-based systems.

2.4.4 Multimedia annotation and the Semantic Web

As noted in the previous section, some researchers have begun to map MPEG-7, as the most comprehensive and widest used standard for multimedia annotation presently, into OWL ontologies in order to be able to additionally reason with the derived annotations. However, this approach is very much an intermediate solution to the lack of any other vocabulary from which an ontology could be built. The limitations of MPEG-7 has a base model for multimedia annotation and the additional requirements for semantic multimedia description have come to be accepted in the Semantic Web multimedia community [van Ossenbruggen,2004; Nack,2005].

At the time of writing, first efforts are being made towards a Semantic Web-tractable multimedia annotation scheme. A W3C Task Force for Multimedia Annotation on the Semantic Web was founded in 2005 [Stamou,2006] and has produced a first draft on Image Annotation [van Ossenbruggen,2006]. We have published an overview of existing multimedia annotation schemes and argued that the SWeMPs vocabularies developed in this research (section 5.8) can cover the minimal required characteristics of media description as a common vocabulary [Nixon,2006]. Another initiative has collected position statements from researchers in the field with the aim of determining a common Multimedia Ontology Framework[26], to which results of this research has also been submitted.

In this section we looked over the history of multimedia annotation, and focused especially on MPEG-7 as the main description scheme for media objects in use at present. We noted how as soon as MPEG-7 was standardized, the recognition of the value of representing it in a knowledge model led to initial efforts at developing a MPEG-7 ontology and that in the past couple of years a number of researchers have developed MPEG-7 based ontologies in their research. To close, the limitations of basing a future semantic scheme for media annotation on MPEG-7 have been recognized and the first steps are being taken to bring researchers in the field together to agree on a suitable vocabulary for semantic multimedia annotation. The research work in this dissertation is providing input to the community on the needs as regards to multimedia presentation generation, an aspect which is commonly overlooked in discussion on media annotation schemes.

The requirements for media annotations for use in multimedia presentation generation can only be determined within some defined framework, and thus we turn to the main proposal that has been made in the literature for such a “multimedia presentation generation” framework.

2.5 An Intelligent Multimedia Presentation System (IMMPS)

Bordegoni et al assessed contemporary research activities related to automated multimedia presentation generation, concluding that there is a need for intelligent multimedia presentation systems and that no generic model had emerged [Bordegoni,1997].

[26] http://www.acemedia.org/aceMedia/reference/multimedia_ontology/index.html

In their work, an intelligent multimedia presentation system means:

"able to make appropriated design decisions based on presentation- and contextual knowledge, and to manage the various interdependencies between choices"

Another statement in their work defines such a system so:

"able to flexibly generate various presentations for one and the same information content in order to meet individual requirements of users and situations, resource limitations of the computing system, and so forth"

In response, a **Standard Reference Model** is presented. The reference architecture is shown (Fig. 2.7). It is intended by the authors to be a generic reference architecture which reflects an implementation independent view of the processes required for the generation of multimedia presentations.

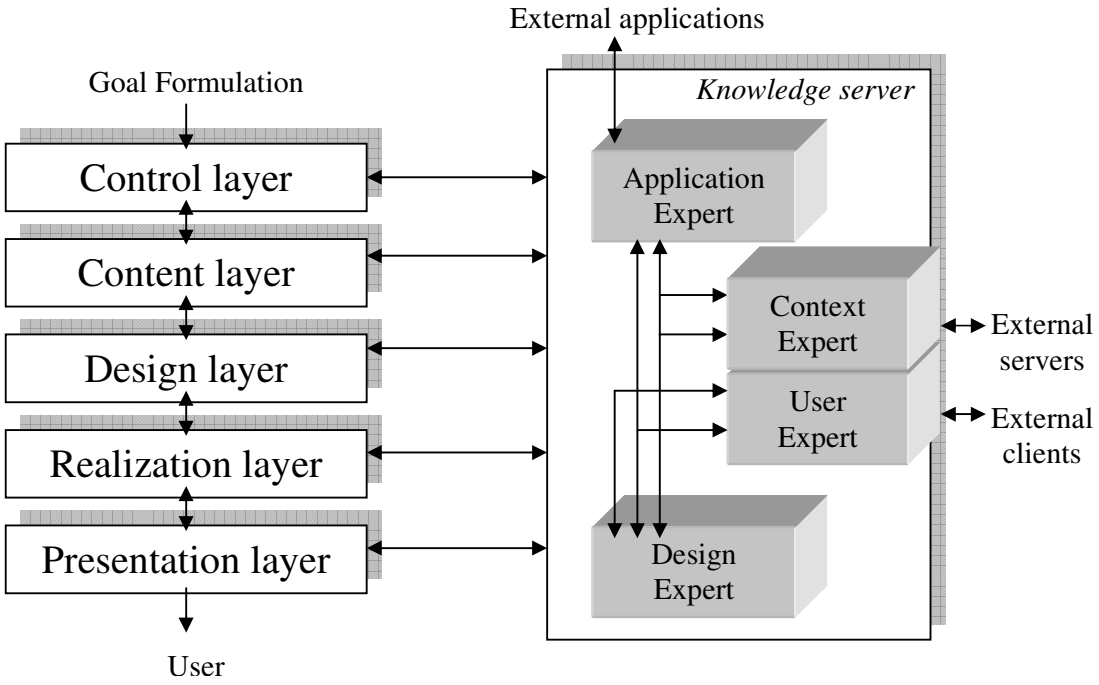

Figure 2.7 Reference architecture for IMMPSs (Reproduction from [Bordegoni,1997])

It consists of "layers" - abstract locations for tasks, processes, or other "components" - an objectification of a task, function or computing process. Components are characterized by their input/output behaviour. "Connectors" enable interchange of information between components. The five layers are:

- *Control* layer – organizes and filters presentation goals
- *Content* layer – selects appropriate content, maps it to appropriate media and chooses the appropriate order of communication
- *Design* layer – transforms the internal media decisions to specifications of media objects and their arrangement in the presentation
- *Realization* layer – realizes the media and layout design in concrete terms
- *Presentation Display* layer – renders the concrete presentation in a form perceivable by the user

The knowledge server is proposed to provide the layers of the Reference Model with several types of knowledge, contained in “expert models”. These are knowledge bases with an inference engine, and interfaces to the layers, other expert models and external knowledge sources. The four expert models specified in the Reference Model and the knowledge they are intended to model are:

- User Expert: users goals, preferences, knowledge, beliefs, abilities
- Application Expert: content provision, content characterization, content conversion, interfacing with application systems
- Content Expert: generation context (what has been generated so far), presentation context (what has been presented so far) e.g. the relationships between media objects and their semantic concepts, the ways in which the user has interacted with the presentation
- Design Expert: design constraints, device model

However, the authors also make explicit that the Reference Model does not seek to define internal or final formats for representing content, presentation or knowledge. Equally there is no formal definition of any protocol or format for the exchange of information or data within the system.

There have not been concrete implementations based directly on the Reference Model. An identified failing, which contributed to this, was the lack of any specifications for the data model of content, presentation or knowledge in an IMMPS. Some work has, however, sought to relate research prototypes back to the model. The research work which has taken on the task of realising intelligent multimedia presentation systems is reviewed in the following chapter.

2.6 Summary

In this chapter we have introduced the fields of multimedia presentation generation and knowledge representation. We outlined well-known research problems in the multimedia presentation field, and given the application of knowledge-based approaches to solve them (which will be discussed in chapter 3) we provided an explanation of knowledge representation with a focus on the Semantic Web (RDF,OWL) and discussed past and present work in representing multimedia through knowledge models. We noted that MPEG-7 has established itself as a standard for multimedia content description, yet it uses a syntactic (XML) data model. Recent work on building MPEG-7 ontologies was described, and it can be hoped that this will form a bridge for the multimedia community to cross over to using knowledge modelling techniques. Researchers in the Semantic Web and multimedia area are taking the first steps towards semantic vocabularies for multimedia description. Finally, given the need for knowledge-based systems that would make use of such knowledge models in “intelligent” multimedia applications, the Standard Reference Model for such an intelligent multimedia presentation system was presented. This Reference Model will serve as a useful guide for our system conceptualisation and implementation.