# Integrating Audiovisual and Semantic Metadata for Applications in Broadcast Archiving

Werner Bailer, Peter Schallauer

Institute of Information Systems
and Information Management
JOANNEUM RESEARCH
Steyrergasse 17
8010 Graz, Austria
{firstname.lastname}@joanneum.at

Alberto Messina, Laurent Boch

Centre for Research
and Technology Innovation
RAI
Corso Giambone 68
10135 Turin, Italy
{a.messina, l.boch}@rai.it

Roberto Basili, Marco Cammisa

University of Roma, Tor Vergata
Via del Politecnico 1
00133 Roma, Italy
{basili, cammisa}@info.uniroma2.it

Borislav Popov

Ontotext Lab, Sirma Group Corp.
Tsarigradsko Chausse 135
Sofia 1784, Bulgaria
borislav.popov@ontotext.com

**Abstract:** This paper describes the work done in the PrestoSpace project dealing with audiovisual and semantic content analysis, content description and retrieval in a broadcast archive environment. We present the metadata model that is used in the PrestoSpace system and discuss its design. A key issue is the integration of audiovisual content description and semantic metadata in the model as well as the use of a knowledge base that handles semantic information across the content set. We describe the audiovisual and semantic content analysis tools that are used to extract the metadata represented in the model and discuss the joint use of this information for content segmentation and retrieval.

## 1 Introduction

The PrestoSpace Integrated Project was launched in 2004 under the Information Society Technologies priority of the Sixth Framework Programme of the European Union. The consortium includes several European broadcasters and audiovisual archive owners, universities and research centres and industry representatives for a total of 35 active partners.

The objective of the project is to provide technical devices and systems for digital preservation of all types of audio-visual collections. The aim is to build-up preservation factories providing affordable services to all kinds of collection owners to digitise, restore, document and distribute their assets. This objective will be achieved through new technologies and processes, and will be implemented as facilities that will provide all holders of audiovisual material with an integrated affordable preservation service.

The project is organised in four main areas: preservation, where specific digitisation techniques are studied, optimised for different types and conservation conditions of media, restoration, that deals with defect analysis and correction of degraded material, storage, that performs surveys on archive management technologies and products and metadata access and delivery (MAD), concerned with content description and retrieval. The architecture of the MAD software platform is shown in Figure 1.

The exploitation of archived audiovisual contents has become an important contribution in the trails of today's growth of interest towards the preservation of cultural heritage. It is in this context, that modern broadcasters have been rediscovering the value of their audiovisual archives in the last decade. In addition, some forerunning works have already shown that approaches meant to the recovery and availability of archived materials may produce consistent cost savings in the overall programme production processes [DDS99]. In order to ensure the feasibility of this, metadata play a central role. Metadata are traditionally defined as "data about data", i.e. those pieces of information allowing a class of users of a system to utilise other pieces of information stored in the system for a determined purpose.

The rest of this paper is organised as follows: Section 2 presents the MAD metadata model and format. Sections 3 and 4 describe the technologies used to produce metadata respectively from audiovisual content analysis and from semantic analysis of text. Section 5 summarises the results obtained from joint use of audiovisual and semantic analysis. Section 6 draws conclusions and points at future developments.

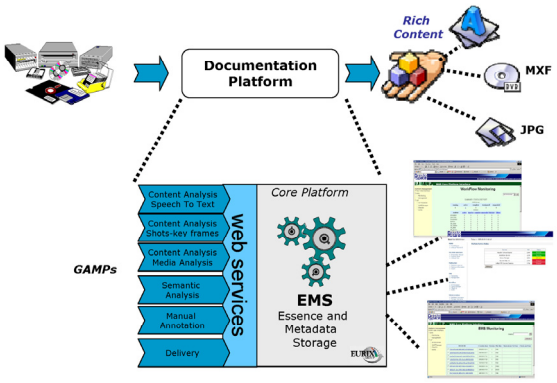


Figure 1. Architecture of the MAD platform.

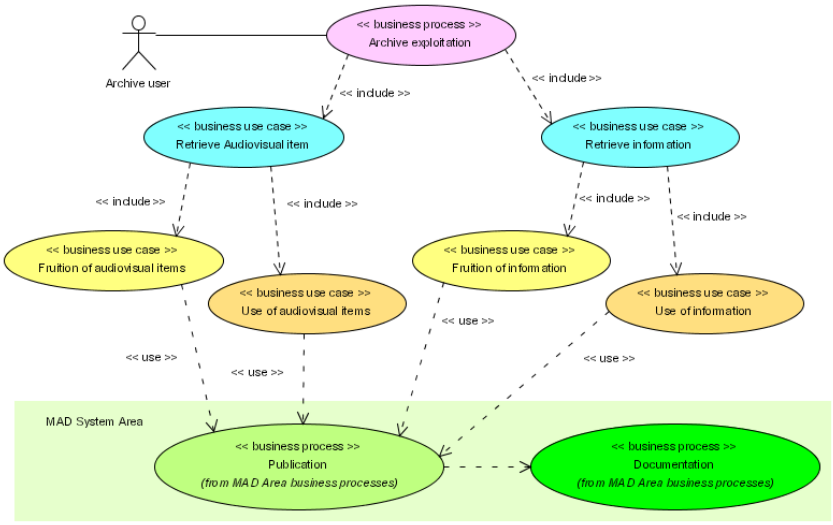


Figure 2. The MAD system business processes. Archive exploitation is fulfilled by retrieving information and audiovisual items from the archive. The MAD Publication process is designed to support to the archive exploitation process, and it depends on a Documentation process.

## 2 The MAD Metadata Model

Exploiting archived material with satisfactory levels of accuracy is possible only if an exhaustive metadata model is adopted for the documentation of archived audiovisual content. A metadata model is to be intended as a set of entities and their relationships with the intent to represent the information contained in the material instances managed by the system.

The PrestoSpace consortium, and particularly the MAD work area, has pursued a rigorous and thorough analysis in order to reach a set of logical data structures giving a complete coverage of the information requirements involved by the business case of the project.

The analysis started from the formulation of a business process model (illustrated by Figure 2), i.e. the organised collection of the core business processes which the PrestoSpace MAD Platform is in charge of realising: Documentation and Publication of archive assets in order to support their efficient exploitation. As a consequence, the data model must serve as a support for the representation and preservation of all the information produced during the actuation of the two core processes of the MAD Platform.

The principal value of the PrestoSpace MAD data model, consists in the establishment of the set of business entities and the set of relationships among the business entities that have to be managed by the MAD Documentation Platform [Mes06]. The main identified business entity is the *Editorial Object*. The Editorial Object plays the pivotal role with respect to all the other entities in the model. It can be defined as the abstract *concept* behind an audiovisual work, regardless of the physical realisations (materials) of the work. An Editorial Object (e.g. a news program) can be divided on its temporal timeline in its constituent parts (e.g. news stories), which are themselves considered Editorial Objects.

To achieve this central concept the data model developers took into account the existing documentation practises and information models, discussed in [Bau05], as well as the most relevant standard publications in the area (e.g. [Mpeg7]). Furthermore, a relevant amount of data processing requirements come from the experiences of PrestoSpace partners who directly operate as broadcasters or archive owners, and from the community of users of multimedia content and systems.

More in detail, the result of this analysis is that the knowledge related to an Editorial Object can be classified into: (a) language and identification information, (b) publication and production information, (c) realising material instances information, (d) editorial partition information, (e) detailed content-related information, (f) external enriching information and (g) ancillary information.

The MAD area data model collects and concretises all these requirements in a single logical structure.

### 2.1 Data Format

To concretely live in the context of a real system, the data model needs to be embodied in a technological framework, i.e. in a data format expressed in a well-defined data representation technology. The natural choice in PrestoSpace MAD has been to use a document-oriented approach, using XML as the data formatting technology.

Considerable efforts in the past years have yielded to the issue of several standards in the area of multimedia content description and in general in the area of metadata. Usually these efforts have ended in the production of data schemas (and more and more often XML Schemas), but rarely they provided a complete specification of the underlying information model, although it's obvious that such models have to be somehow implied. Therefore, the MAD area partners have finally oriented themselves to the evaluation of the existing state-of-the-art in this field with the intention to pick the best functionalities from each available source and to spend integration efforts in bridging the syntactic and semantic gaps among the different contributions.

The result of the outlined approach is the specification of a composite XML Schema document, in which each individual component is expressed using the syntactic tools defined in the external standard that is the most appropriate for the particular task ([Mpeg7] and [PMeta] were the two used standards), together with specific structures developed ad-hoc by the consortium. Summarising this means that (see also Figure 3):

- The root element is an expressly defined structure directly connected to the Editorial Object business entity.
- The top level global information, including identification information, production, publication and genre information, are expressed using P_META sets.
- The material realisation information is an expressly defined structure
- The editorial partition and editorial views of the Editorial Object are expressed using an MPEG-7 profile [Bai06]
- All content related information is expressed using the same MPEG-7 profile
- Enrichment information is expressed using ad hoc structures
- Ancillary data are realised using the same basic mechanism used for material realisation.

For the unique identification of the instances of audiovisual material the SMPTE UMID [Umid] standard is used throughout the whole MAD platform.

During execution on a particular instance of material representing and realising a definite Editorial Object, the MAD Documentation Platform processors (also known as GAMPs – Generic Activity MAD Processors) enrich the parts of the metadata document instance which fall under their respective competence in a collaborative fashion [Mes06].

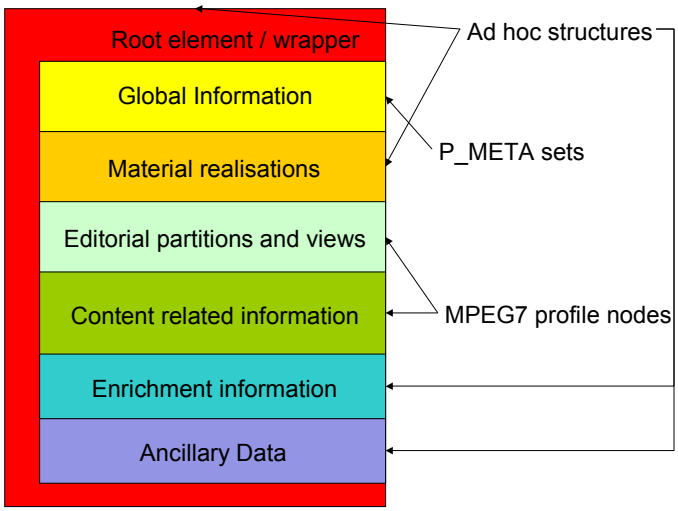


Figure 3. Schematisation of the MAD area data format structure.

## 2.2 Structural Description

The description of the structure of an Editorial Object and its decomposition into sub-segments is one of the most important parts of the content description. The reason is that apart from the global metadata of an audiovisual content all descriptions are related to some type of parts of the Editorial Object.

In the PrestoSpace MAD data model, we discriminate two main types of content decompositions: *editorial parts* and *editorial views*. Both editorial parts and views represent structural decompositions of the content along the time line. Parts and views may contain sub-segments, so that the editorial part and view decompositions may be recursive.

*Editorial parts* are segments of the content, in which all modalities are present and which represent a coherent editorial item, for example a scene or a news story. An editorial parts decomposition is a temporal decomposition of an audiovisual segment. As a consequence, the segments in an editorial parts decomposition must have exactly the same modalities and number of tracks per modality as the segment being decomposed. An editorial parts decomposition links the segments involved in a container/containee relation. The parts have to be included within the range of the segment being decomposed and may not overlap. The editorial parts structure is recursive and of arbitrary depth.

*Editorial views* are any other decompositions of the content, which need not contain all modalities of the original content, for example the decomposition into visual shots. The segments in an editorial views decomposition usually only have a subset of the tracks and modalities of the segment being decomposed. The depth of the hierarchy and the types of sub-segments depends on the view decomposition and on the types of editorial parts which are described.

The two most important types of editorial views in the MAD data model are the decompositions into visual shots and speech segments. *Visual shots* represent a decomposition of the visual modality into segments delineated by hard cuts or gradual transitions. Further descriptive information is attached to the segments representing shots: one or more key frames representing the visual content of the shot, and optionally a description of the visual features of each key frame and a description of the camera motion in the shot. *Speech segments* are segments of detected speech as defined by the automatic speech recognition (ASR) engine.

To describe the structure of the description, the MPEG-7 structure description tools are used. The distinction between the different editorial part and view descriptions follows the approach proposed in [Bai06] in order not to create any interdependencies between decompositions based on different features or feature combinations. The MPEG-7 Detailed Audiovisual Profile requires the identification of the type of each decomposition and segment by the mandatory use of criteria attributes and StructuralUnit elements respectively. While the profile proposal specifies their values for a number of common decompositions, the MAD data format specification extends this list for all the editorial parts and views supported in the MAD data model.

The segments representing Editorial Objects or their parts (which are in turn Editorial Objects) are annotated with the realisation, identification and production information, with classification information and description of the named entities related to the Editorial Object. The editorial views are annotated with the respective audio and visual feature description. All these annotations are discussed in the remaining part of this section.

### 2.3 Realisation, Identification and Production Information

**Material realisation information.** The whole MAD area process depends on the previous phases of the PrestoSpace factory, namely on the digitisation and on the archive inventory processes. The result of the archive inventory process consists in the data structures representing the associations between the materials, identified by UMID numbers, and the realised Editorial Objects. This association can be of varying complexity, depending on whether a single material instance or a set of material instances concur together to realise the Editorial Object. Other information is about the kind of derivation that the material instances may have undergone.

**Identification and language information**. Identification information refers to all the data that contribute to identifying a certain Editorial Object in the editorial, production and publication context. The typical example of such a kind of data are represented by titles and contributors. The analysis pointed out the main aspects related to the identification issue. The P_META standard [PMeta] defines several data sets, together forming an exceptionally complete framework on which to build such identification structures in the MAD area data format context. Language information is also rendered using specific P_META sets.

**Publication and production information.** Publication information, i.e. information about when and on which broadcast service a certain Editorial Object has been published together with ancillary descriptive information about the transmission event, are rendered by the use of P_META sets. These sets give account of individual broadcast publications and of the summary of all the (known) publications of the Editorial Object. Production information includes dates of production, original classification details (genres) and editorial control information, again rendered by using the proper P_META sets.

### 2.4 Description of Visual and Audio Features

Descriptions of audio and visual features are attached to the segments in the editorial view decompositions.

**Visual features of key frames.** Visual color and texture features are extracted from the key frames and standard MPEG-7 descriptors and are attached to the segments representing the key frames. Currently the features ColorLayout, ColorStructure, DominantColor and EdgeHistogram are used.

**Camera motion.** As a continuous camera motion is delimited by a shot boundary, camera motion is described per shot. In order to cope with the fact that camera motion can change significantly within a shot, we decompose the description into camera motion segments, each characterized by homogeneous camera motion within. For search and retrieval purposes an exact description of the global motion is neither necessary nor reasonable. We therefore use a simplified description restricted to nine types of camera motion (fixed, pan left/right, tilt up/down, roll left/right, zoom in/out) and a rough quantization of the amount of motion.

**Visual activity.** Like camera motion, visual activity is described on the shot level. The visual activity over time is described by a sparse sequence of activity samples.

**Automatic speech recognition (ASR) results.** The unit of the segments, for which ASR results are described, is a speech segment as defined by the editorial view. All attributes of the annotation refer to this segment. However, if necessary, a segment may be further decomposed in order to describe a more precise alignment of parts of the text to the time line. The results of automated speech recognition (ASR) are only described as text and not on the phoneme level. Additional optional attributes are the language of the speech, a confidence value and the identification of the speaker. Like in the semantic descriptions discussed below, the identification of the speaker is a reference to the knowledge base.

### 2.5 Semantic Content Description

This section discusses the description of the semantics of the content such as classification of the content and the description of named entities appearing in the content or related to it. The semantic content descriptions are not only very relevant features for retrieval, but the use of ontologies and the common management of the named entities that exist throughout the content collection allow further reasoning based on these annotations.

**Classification.** Classification information, (e.g. genre of an editorial object, content classification of a news story) may be related to Editorial Objects or smaller units defined by the component performing the classification. Classification information is related to all modalities of the content and refers to an external list of classes, such as a thesaurus or ontology. Within the content description, only the reference to this external classification scheme and an optional label (which can be used by the application for display, if it does not access the knowledge base) is described.

**Annotation of named entities.** Surveys have shown that named persons, objects, places and events are the most important criteria for retrieving audiovisual material [ES01]. Thus both the named entities perceivable in the content (e.g. names being mentioned, places shown) or related thematically are annotated in the content description. The multimedia content description describes the occurrence of a related named entity and references the entity in the knowledge base (cf. Section 4.2) using URIs. Thus no redundant information is held in the content description, but existing ontologies can be used. While the multimedia content description is related to one media item, the entities contained in the knowledge base are valid across the content collection.

The MPEG-7 semantics description tools [Mpeg7] are used for the description of the named entity occurrences. The following specialisations of SemanticBaseType are used: Object (with the further specialisation AgentObject for persons and organisations), Event, Concept, SemanticState, SemanticPlace, SemanticTime. The description of the occurrence of a named entity may contain an optional label (which can be used by the application for display, if it does not access the knowledge base), a confidence value, and references to external information.

**Enrichment information.** There may be external information that is related to the content that is referenced, partly because it has been used as input for the semantic analysis (cf. Section 4). This includes for example news articles on the web related to a broadcast news story, scripts, etc. The MPEG-7 RelatedMaterial description scheme is used to describe the related information to a content segment.

## 3 Audiovisual Content Analysis

Automatic analysis tools for audiovisual content are used in the PrestoSpace MAD system to extract metadata from the material and augment the content description. The automatically extracted metadata are used to support manual annotation, as input to the semantic analysis tools and to index the audiovisual content. The state of the art of audiovisual content analysis has been surveyed in [Bai05a] and the analysis tools discussed in the following have been selected. The results of content analysis are described using the metadata model and format discussed above.

**Shot boundary detection.** The shot boundary detection tool segments a video in its primary building blocks, i.e. its shots, and is capable of detecting both abrupt and gradual transitions. Shot boundaries are a prerequisite for other visual content analysis algorithms, content structuring and indexing and serve as a navigation support in the manual documentation tool.

**Key frame and stripe image extraction.** The key frame detector extracts a number of key frames per shot, depending on the amount of visual change. The key frames serve as representations for the shots and are used as input for low-level feature extraction. Stripe images are spatiotemporal representations of the visual essence, created from the content of a fixed or moving column of the visual essence over time. They serve as a help for quick content overview and navigation, especially in the manual documentation tool.

**Camera motion detection.** The camera motion detector analytically describes four basic types of camera motion in the content (pan, tilt, zoom, roll), a rough quantisation of the amount of motion, and the length of the segments in which they appear [Bai05b]. Camera motion information is an important search criteria when reusing archive material, as the visual grammar rules impose constraints on the camera motion in subsequent shots.

**Speech to text transcription.** Extracting text from spoken content of audiovisual material is a fundamental step allowing for several documentation tasks, as well as representing an important core of searchable data in the publication system. In the current set-up of the documentation platform an automatic speech-to-text engine is used, developed by ITC-IRST (Instituto per la Ricerca Superiore di Trento), capable of extracting text from English and Italian speech.

**Audio structuring and segmentation.** This analysis consists in classifying segments of audio in four principal categories (silence, music, speech, noise). This information is mainly considered as a support for manual annotation.

**Low-level visual feature extraction.** The low- level feature extractor describes key frames or shots in terms of their colour, texture and motion features. The tool extracts some of the descriptors specified in the MPEG-7 visual part ([Mpeg7], part 3), namely ColorLayout, ColorStructure, DominantColor, EdgeHistogram and MotionActivity. The descriptors serve as a compact and efficient representation of the visual content of a shot and are used to determine visual similarity between shots.

## 4 Semantic Information

The MAD platform aims at exploiting human language technologies for Information Extraction (IE) from the audiovisual data made available by large archives. The nature and complexity of management, search and reuse of archive materials require complex storage and retrieval functionalities. These activities ask for:

- Recognition and indexing of suitable generalizations of relevant archive concepts as people names, organizations and locations
- Effective retrieval functions that improve indexing at the simple textual level and support conceptual rather than string retrieval
- Interoperability at the levels of abstraction required by audiovisual content. For example, audiovisual essence should be published, queried and exchanged in a distributed fashion. The development of Web publication should support distributed querying and semantic service-based instantiation and invocation. The semantic data descriptions are critical in these activities and interoperable models (ontologies) are needed.

### 4.1 Semantic Analysis

Semantic Analysis is applied in MAD to fit such high-quality requirements from the available multimedia features (e.g. audio features) to suitable generalizations and ontological representations. In the Semantic Web area, the processes going from raw and textual data to ontological annotations are typically called Information Extraction processes. The starting point of the semantic analysis are the results of the automatic speech recognition (ASR) module described above.

The redundancy that audiovisual objects guarantee at the data level needs to be explored in order to govern the retrieval complexity at the proper quality. The problems due to noisy nature of the extracted data (e.g. errors in the ASR that produce mistakes in the grammatical recognition) should be properly limited. The aim is to make as much information as possible available to the overall extraction and retrieval components of MAD. From this perspective larger data sets than just the source AV data should be taken into account. The input textual material should be processed and enriched by the following relevant evidences as semantic metadata:

- Terminological and lexical information local to the AV input data (via ASR)
- Recognition of citations to Named Entities (e.g. people or organization) from local data as well as from reachable external sources
- Automatic computation of useful hyperlinks between the archived AV data (e.g. the individual segments in broadcasted TV journals) and the distributed sources (e.g. Web-based newspaper portals and pages). These sources include assessed textual descriptions of topics related to the AV segment contents and are trusted.
- Ontological information contained in all the above sources, as representation of classes (e.g. geographical locations, organizations or persons), individuals (e.g. John Coltrane or USA/United States) and topical classes (e.g. Education vs. Sport, Foreign Politics vs. Economics).

The modules invoked in the semantic analysis are the following (see Figure 4):

- an *Intaker*, that normalize the input AV segment text obtained by ASR
- the *News Categorizer*, a topical categorization module that assign a specific category to incoming segments
- the *Natural Language Parser*, [BZ02] that recognize lexical units in the ASR transcripts and provides grammatical disambiguation (POS tagging)
- a *Named Entity (NE) Recognizer* that extracts citations of people, organizations, locations and other interesting entities (e.g. dates)
- an *Ontology-based NE recognizer* that links the discovered NE to known individuals and entities of the reference ontology
- a *Web Aligner* that search the Web for pages related (or equivalent) to the source individual AV segments

The Information Extraction chain first applies the *Intaker* module. It collects and normalizes the incoming broadcasted news items as they are transcribed and segmented by the speech recognition GAMP. Then, the *News Categorizer* is invoked to assign the suitable topical categories (and an associated confidence scores) according to the target classification scheme. In the Italian semantic analysis GAMP the RAI internal classification scheme has been adopted. Concurrently, these news items can be parsed (via the *Parser*) to detect Named Entities (via the *Named Entity Recognizer*), these provide a set of significant metadata that can be used by the *Aligner* module to search for candidate news items that are similar to the pages downloaded by a *Web Spider*. The retrieved Web pages are also parsed and indexed according to traditional IR techniques. For each news item, the *Alignment* process selects the suitable Web pages from the set of

the retrieved candidates and sets direct hyperlinks to them. These links are used to include further (external) metadata, auxiliary to the internal ones, to improve the overall accuracy that can be affected by wrong or irrelevant information. Finally, a module using an Ontology to annotate the news item is applied (in PrestoSpace, the KIM platform [Kir05] is used, which is further discussed below). More details on the Italian semantic analysis GAMP can be found in [BCD05], while the English semantic analysis GAMP is discussed in detail in [Dow05].

### 4.2 The Role of Ontological Information

The ontological component in PrestoSpace is managed by the KIM platform [Kir05] which supports information extraction based on an ontology and a massive knowledge base. The KIM platform implements semantic annotation as an innovative model for automatic semantic content enrichment, it enables new information access methods, and extends the existing ones. In this way, KIM supports applications such as highlighting, indexing and retrieval, categorization, generation of advanced metadata, smooth traversal between unstructured text and the available relevant knowledge. The information extraction approach employed in KIM has roots in the conception that certain entities a content refers to are of significant importance for the meaning of the content they appear in. To clarify why named entities constitute an important part of the semantics of the documents, consider the sentence "the first president of the United States". Understanding the meaning of the constituent words is not enough to understand the meaning of the sentence. Unlike words, named entities denote an (often concrete) individual and not a class or any member of the class. When describing the meaning of words, lexical semantics and/or common sense would suffice; to understand the meaning of named entities more specific knowledge about the world is required.

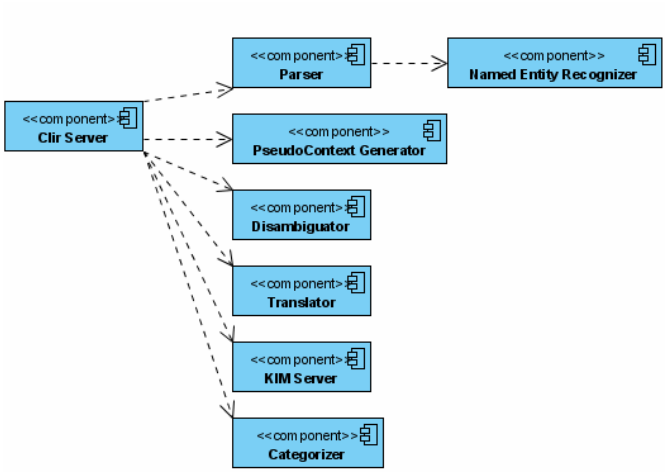


Figure 4. Components of the semantic analysis GAMP.

Semantic annotation is a generation of specific metadata. It is the process of assigning to the named entities in the text links to their semantic descriptions. The semantic annotation process in KIM is based on a simple model of real-world entity classes (i.e. an ontology) and a massive knowledge base. The semantic annotation (metadata) has certain prerequisites for its representation: (1) an ontology (or at least, a taxonomy), which defines the entity classes; (2) entity identifiers, which allow entities to be distinguished and linked to their semantic descriptions; (3) a knowledge base with entity descriptions. KIM relies on two types of ontologies: *upper-level* (PROTON, http://proton.semanticweb.org, roughly domain-independent) and *domain-specific*.

The PROTON upper level ontology encodes the most common aspects of any considerable description, no matter of the specificity of the domain (weather forecast, popular science documentary, etc.), regardless of the specificity of the task in view (for example - classification of movies, access to news emissions, description of the themes of the documentaries). PROTON was designed to address the requirement of being suitable for open-domain general purpose semantic annotations as well as to allow easy extensions according to specific needs. It currently contains about 300 classes and 100 properties.

For the purposes of semantic annotation, indexing, and retrieval of documents, KIM also uses a seed *knowledge base* (KB). The knowledge base (KB), in this context, is a body of formal knowledge about entities, a means for the representation of non-ontological formal knowledge. It consists of instance data – descriptions of entities and their interrelations, i.e. for each entity, the KB contains information about the entity's type, aliases (including a main alias, i.e. official or well-known name), attributes, and relations. The KIM KB provides coverage of popular real-world entities of common interest, which are considered well-known and thus not explicitly introduced in the documents. Most important and used entities in the KIM KB are *geographic names and organizations*. The entities representing geographical features are imported from *GNS* (*GEOnet Names Server*) and other sources. They are organized so as to represent instances of *Location* (and its subclasses) having the property *subRegionOf* as it is applied between *Continents*, *GlobalRegions*, *Countries*, and other subclasses of *Location*. Some subtypes of *Location* which are contained in KIM KB are *Country, Province, County, CountryCapital, City, Ocean, Sea*, etc. The locations are given together with several among their aliases, including English and French aliases, as well as with their geographic coordinates (*Long/Lat*), the designator (*DSG*) and Unique Feature Index (*UFI*), ccording to *GNS*. All this provides a useful basis for cross-linguistic querying and retrieval. The entities in the KB are derived or collected from various sources as geographical and business intelligence gazetteers.

One of the roles of KIM in MAD is to provide a language independent representation for Named Entities as a specific metadata common to the two languages. As an example consider that the *"White House"* is translated in other languages (e.g. in Italian the correct translation is *"Casa Bianca"*). The ontology representation for this entity is via an unique id (i.e. an Uniform Resource Identifier *"URI"*), that is for its nature language independent. This realizes a systematic and consistent approach to multilingual indexing and searching.

# 5 Joint Use of Audiovisual and Semantic Information

## 5.1 Content Segmentation

Editorial parts are the constituent parts of a programme from the editorial point of view, i.e. that of the creators of the programme (e.g. news items in a newscast programme). Several techniques have been investigated to solve the hard problem of identifying editorial parts from the low-level analysis of raw content [Bai05a], though none is solving the problem generally. Due to this, the PrestoSpace MAD unit limits the use of automatic editorial segmentation in the news domain, choosing a multi-layer approach that merges visual and aural information for the detection of news items in the mainstream newscast editions. Lexical segmentation of text is used as a further layer of information to enhance the performances of pure multimedia structure detection based on audio and video features.

The used algorithm is basically divided in two steps. The first phase detects the *relevant* multimedia structure with respect to the editorial segmentation task. The relevant multimedia structure is defined as the particular subset of the total set of shot boundary points, which is the best approximation of the editorial partition. At first, visual boundaries are selected among visual shot clusters boundaries (clusters of visually similar shots), through a filtering mechanism which classifies shot clusters in "boundary cluster" or "non-boundary cluster", based on some of their structural characteristics (e.g. elongation, duration, total number of enclosed shots). The obtained set of shot clusters is further refined by filtering off those clusters which do not show a predefined percentage of temporal coverage w.r.t. the main speaker of the programme, as identified by a speaker labelling process.

The second step of the algorithm makes use of another layer of information, namely the lexical structure of the text. Based on the results of the semantic analysis described above, a topic segmentation is produced. In a fusion step, the segmentations based on audiovisual and lexical information are combined.

## 5.2 Manual Annotation

The use of automatic audiovisual content analysis and semantic analysis can support a human annotator by performing a number of routine tasks, suggesting annotations and providing aids for structuring and navigation, but it will not fully replace a documentalist. In order to ensure a high quality level of the annotation it is necessary to have a human in the loop, who validates the results of automatic content analysis and semantic analysis, does additional structuring, and adds textual information that cannot be added automatically. The annotation tool allows viewing and modifying both the results of audiovisual and semantic analysis tools. The results of the automatic analysis are not only used to automatically extract metadata, but also the create aides for viewing and navigating the content, such as key frames and stripe images.

### 5.3 Retrieval

The rich variety of information extracted by different analysis modules poses several requirements to the Information Retrieval functionalities in the publication phase. First, the user interface should model access methods according to different (and integrated) capabilities:

- Full text search as usually applied by mostly popular search engines
- Natural Languages Questions
- Semantic browsing as navigation through concepts, relations and instances of the ontology

All the above functionalities are to be intended as language neutral: full texts should be searchable in different languages, while ontological information as well as NEs should be properly represented so that language ambiguity and variability are taken into account. Second, all the above search modalities should be offered in a language independent fashion. The discussion of technological solutions to support the above processes is reported below in this document, as they have a relevant impact on the accuracy reachable by the PrestoSpace solutions to CLIR issues.

The viable solutions to the above problem concern:

- The adoption of language neutral representation (via the KIM ontology), and
- Query processing (expansion and translation) for dealing with multilingual information during search.

**Ontology-driven retrieval.** In MAD, the KIM platform is in charge of making available extensive ontological knowledge about the news domain, and supporting indexing and navigation functionalities. It provides novel Knowledge and Information Management infrastructure and services for automatic semantic annotation, indexing, and retrieval of unstructured and semi-structured content/documents. It differs from other systems and approaches in that by providing semantic annotations it supports also IR services based on the results. Different KIM front-end user interfaces are possible given the KIM API, which provides the functionality and infrastructure for the semantic annotation, indexing and retrieval, as well as document management, and KB navigation. The KIM web user interface allows traditional access methods (key word search) and semantic ones (entity search, pattern search), too. Via semantic search the user is allowed to express querying about specific entities restricted by formal constraints over properties. This can be done by navigating the ontology or filling special purpose templates. The interface can return either a set of entities that satisfy the query or the set of documents that refer to these entities. The user can access the document content enriched with the associated metadata on the document level (such as title, author, the target entities, ... ). More details about the KIM and PROTON technology for ontology-driven IR can be found in [Pop04][Kir05].

**Cross-language information retrieval (CLIR)** is supported in the MAD publication platform by a specific server called *CLIR Server*. The CLIR Server includes several components (see Figure 5):

- The *NL Parser*, to extract *Named Entities* and other nouns from the query *q*, in the source language *L*;
- *Pseudo Context Generator*, to generate for each target lexical item *t* in *q*, the most relevant terms that are topically related to *t*;
- *Sense Disambiguator*, to disambiguate all common nouns in the source language *L*;
- *Translator*, to translate the disambiguated common nouns from the source language *L* to the target language *L2*;
- *Kim Server*, to annotate the ontological entries as they are found in a query *q*;
- *Text Categorizer* that classifies the query *q*.

The CLIR Server communicates with these components and manages the internal workflow. The NL parser, Text Categorization and Kim annotation processes are the same used for the Semantic Analysis GAMP.

A distinctive feature of the CLIR server is the adopted technique for Sense Disambiguation and Translation. Translation of all common nouns is required as they are very language specific and must be consistently combined to the language-independent representation of Named Entities.

The sense disambiguation algorithm adopted has been presented in [BCG06]. The aim of the method is to automatically extend the information about a query via text mining techniques, disambiguate nouns through Wordnet senses and use them to select suitable translations in the target language.

In particular, a query expansion process is first applied through a *Latent Semantic Analysis* (LSA, [BDO95]) approach. The initial query *q* is mapped into an LSA space (previously obtained from news corpora in both languages): this allows to associate to all nouns in *q* the closest terms, i.e. a lexicon *dom(q)* associated to the *q*'s topical domain (*Pseudo Context Generation*). Within this lexicon a sense disambiguation process is applied*:* an *n*-ary similarity metric (see [BCZ04]) is used here to rank Wordnet senses of individual nouns in *q* given *dom(q) (Sense Disambiguation)*. As sense ambiguity is much lower within a domain, the sense disambiguation in *dom(q)* is very effective. Preferred senses are finally used to generate translations (*Translation*). The interlingual interfaces of Wordnet, in fact, link *synsets* in different languages. The best senses (synsets) for nouns in q are then used to derive their best translations in the target language. The resulting query includes named entities, query category and all the synonyms of original nouns in the target language. The method runs in a fully automatic way as LSA can be applied without human intervention. The sense disambiguation algorithm is much more effective when combined with LSA as discussed in [BCG06].

**Browsing in publication.** The MAD Publication Platform provides retrieval and browsing functionalities. It deals with instances of documents conforming to the MAD metadata format and makes them available in a web-based representation. It also gives access to the materials exported from the Core Platform. The Publication Platform architecture is based on a Web application as user interface, a DBMS storing the available information related to programmes, and the KIM indexing and search engine.

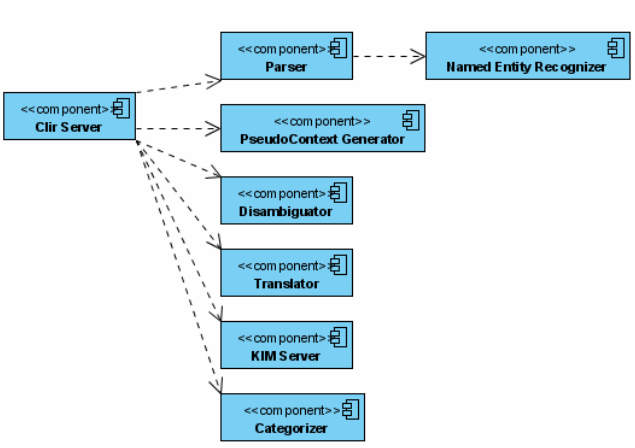


Figure 5. Components of the cross-language information retrieval (CLIR) server.

The search interface supports the various retrieval approaches described in the previous sections, and the user can choose the target of his/her search (e.g. a programme or a news item), which can be filtered by title, broadcast date and service, contributions (e.g. authors, journalists, directors), classification (topics, categories), text of description. As the user selects an item (table row), a *browsing* window is opened which presents all the details of the specific item. The window is made up of four frames: a video preview, the editorial parts tree, the key frames, and an extensible multi-tab frame, each of which is representing a specific elaboration result. The content of all the frames is synchronised during user interaction. The following tabs of the multi-tab frame have been implemented so far:

- *Info* It contains the general metadata about the programme such as title, subtitle, publication dates and channels, contributors.
- *Transcriptions* This tab shows the speech-to-text transcript. The text is divided into segments representing individual news items. The interface also allows the user to select a specific text segment.
- *Semantic analysis* This tab shows a navigable tree that can be explored interactively. It shows the entities found during semantic analysis.
- *Content analysis* Here the user can view the stripe images and the related camera motion information on the timeline.

## 6 Conclusion

In this paper we have presented the data model used in the PrestoSpace MAD project, the audiovisual and semantic content analysis tools used for automatic metadata extraction, and the joint use of audiovisual and semantic metadata for content segmentation and retrieval. The data model is the hub for integrating a number of heterogeneous components for both automatic and manual documentation as well as publication.

The data model has been developed based on an analysis of the processes in a broadcast archive environment and an analysis of existing data models and metadata standards. There is no single standard that fulfils all requirements of the system, but this does not mean that one should not use any of them. In order not to reinvent what already exists, we use metadata standards (MPEG-7, P_Meta) whenever possible and combine them in our data format, adding elements that are not covered by the standards. The structural description of an Editorial Object, as well the audiovisual content description and the annotation of semantics in the content description are based on the MPEG-7 Detailed Audiovisual Profile; production, identification and publication information are described using P_Meta.

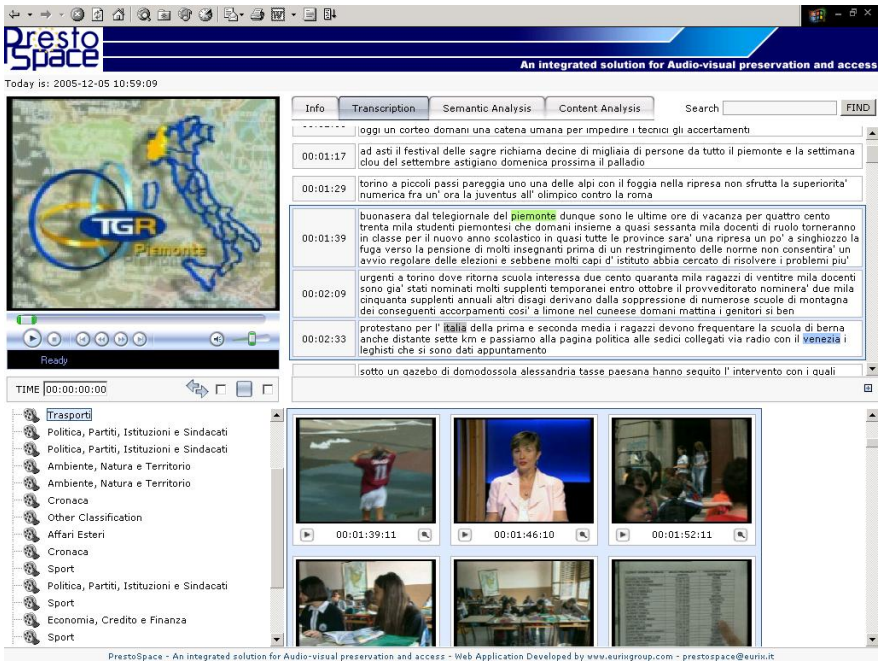


Figure 6. Use of MAD system processes to achieve archive exploitation.

As has been shown in Section 5, the joint use of audiovisual content and semantic metadata is beneficial for both analysis and retrieval applications. Thus the metadata model has to support the integration of both types of metadata. Again, there is no single format that is ideal for describing all that information, but for each of them there are standards designed specifically for this type of metadata, such as MPEG-7 for time-based structural description of audiovisual content and OWL for semantic information. We are convinced that both are complementary and thus the question is not which one to choose, but using a mechanism to integrate them.

In the PrestoSpace system the KIM platform serves as a knowledge base for the named entities across the content set. It is important not to treat this information on a per document basis, as the same entities will appear across document and thus additional information can be gained by the global approach. The audiovisual content description makes use of references to the knowledge base whenever possible, e.g. for classifications and named entities. The MPEG-7 semantics description tools are flexible enough to not only use MPEG-7 classification schemes but any external controlled vocabulary, such as our knowledge base. We use this mechanism to successfully integrate audiovisual content description and semantic information in our data model.

## 7 Acknowledgements

This work has been funded partially under the 6th Framework Programme of the European Union within the IST project “PrestoSpace” (IST-FP6-507366, http://www.prestospace.org).

## 8 Bibliography

[Bai05a] Bailer, W., Höller, F., Messina, A., Airola, D. , Schallauer, P., Hausenblas, M.: State of the Art of Content Analysis Tools for Video, Audio and Speech: PrestoSpace Deliverable 15.3, 2005. URL: http://www.prestospace.org/project/public.en.html

[Bai05b] Bailer, W., Schallauer, P., Thallinger, G.: Joanneum Research at TRECVID 2005 – Camera Motion Detection: Proc. TRECVID Workshop, Gaithersburg, MD, USA, Nov. 2005.

[Bai06] Bailer, W., Schallauer, P.: The Detailed Audiovisual Profile: Enabling Interoperability between MPEG-7 Based Systems: Proc. of 12th International Multi-Media Modeling Conference, Beijing, CN, Jan. 2006.

[BCD05] Basili, R., Cammisa, M., Donati, E.: RitroveRAI: A Web Application for Semantic Indexing and Hyperlinking of Multimedia News, in “International Semantic Web Conference", Y. Gil, E. Motta, V.R. Benjamins, M.A. Musen Eds., Lecture Notes in Computer Science, LN 3279, 97-111, 2005.

[BCG06] Basili, R., Cammisa, M., Gliozzo, A.: Integrating Domain and Paradigmatic Similarity for Unsupervised Sense Tagging, Proceedings of the European Conference on Artificial Intelligence, Riva del Garda, (Italy), 2006.

[BCZ04] Basili, R., Cammisa, R., Zanzotto, F. M.: A semantic similarity measure for unsupervised semantic disambiguation, Proceedings of the Language, Resources and Evaluation LREC 2004 Conference, Lisbon, Portugal, 2004.

[BZ02] Basili, R., Zanzotto, F. M.: Parsing Engineering and Empirical Robustness, 8 (2/3) 97120, Journal of Language Engineering, Cambridge University Press, 2002

[Bau05] Bauer, C., Rosensprung, F., Lajtos, S., Boch, L., Poncin, P., Herben-Leffring, C.: Analysis of current audiovisual documentation models, Mapping of current standards: PrestoSpace Deliverable 15.1, 2005.
URL: http://www.prestospace.org/project/public.en.html

[BDO95] Berry, M. W., Dumais, S. T., O'Brien, G. W.: Using linear algebra for intelligent information retrieval, SIAM Review, Vol. 37, No. 4, pp. 573-595, December 1995.

[Bru00] Brugnara, F., Cettolo, M., Federico, M., Giuliani, D.: A system for the segmentation and transcription of Italian radio news: Proc. RIAO, Content-Based Multimedia Information Access, Paris, France, 2000.

[DDS99] Del Pero, R., Dimino, G., Stroppiana, M.: Multimedia Catalogue – the RAI Experience: EBU Technical Review nr. 280, Geneva, 1999: pp. 1-13.

[Dow05] Dowman, M., Tablan, V., Cunningham, H. and Popov, B.: Web-Assisted Annotation, Semantic Indexing and Search of Television and Radio News. 14th International World Wide Web Conference. Chiba, Japan, 2005.

[ES01] Enser, P., Sandom, C.: Retrieving archival moving imagery – a step too far for CBIR? Proc. Multimedia Content-Based Indexing and Retrieval Workshop, Rocquencourt, France, Sept. 2001: pp. 7-10.

[Kir05] Kiryakov, A., Popov, B., Ognyanoff, D., Manov, D., Kirilov, A., Goranov, M.: Semantic Annotation, Indexing, and Retrieval. Elsevier's Journal of Web Semantics, Vol. 2, Issue (1), 2005.

[Mes06] Messina, A., Boch, L., .Dimino, G., Bailer, W., Schallauer, P., Allasia, W., Basili, R., Groppo, M., Vigilante M.: Creating rich Metadata in the TV Broadcast Archives Environment: the PrestoSpace project: Proc. AXMEDIS06 Conference, Leeds, Dec. 2006.

[Mpeg7] ISO/IEC 15938, Multimedia Content Description Interface.

[PMeta] EBU Tech3295, European Broadcasting Union (EBU) P_META Metadata Exchange Scheme.

[Pop04] Popov, B., Kiryakov, A., Ognyanoff, D. , Manov, D., Kirilov, A.: KIM - a semantic platform for information extraction and retrieval, Journal of Natural Language Engineering, Vol. 10, Issue 3-4, Sep 2004, pp. 375-392, Cambridge University Press.

[Umid] SMPTE 330M, Unique Material Identifier (UMID), Society of Motion Pictures and Television Engineers (SMPTE), 2000.