Creating Rich Metadata in the TV Broadcast Archives Environment: the PrestoSpace Project

A.Messina[1], L. Boch[1], G. Dimino[1]
[1]RAI – Centre for Research and Technology Innovation, Corso Giambone 68, I-10135 Turin, Italy
{a.messina, l.boch, g.dimino}@rai.it

W. Bailer[2], P. Schallauer[2]
[2]JOANNEUM RESEARCH, Institute of Information Systems and Information Management, Steyrergasse 17, 8010 Graz, Austria
{werner.bailer, peter.schallauer}@joanneum.at

W. Allasia[4], M. Groppo[4], M. Vigilante[4]
[4]EURIX, Multimedia Department, c.so Tortona 17, I-10153 Turin, Italy
{allasia,groppo, vigilante}@eurix.it

R. Basili[3]
[3] University of Roma, Tor Vergata, Via del Politecnico 1, I-00133 Roma, Italy
{basili}@info.uniroma2.it

Abstract

This paper describes the part of the European PrestoSpace project dedicated to the study and development of a Metadata Access and Delivery (MAD) system for television broadcast archives. The mission of the MAD system, inside the wider perspective of the PrestoSpace factory, is to generate, validate and deliver to the archive users metadata created through the employment of both automatic and manual information extraction tools. Automatic tools include audiovisual content analysis and semantic analysis of text extracted by automatic speech recognition (ASR). The MAD publication platform provides access and search facilities to the imported and newly produced metadata in a synergic and easy-to-use interface.

1. Introduction

1.1. Scope of the PrestoSpace project

The PrestoSpace Integrated Project was launched in 2004 under the Information Society Technologies priority of the Sixth Framework Programme of the European Community (IST FP6 507336). 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, manage 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 importance of the European audiovisual heritage has been recognised by the Council of Europe, which issued a European Convention for the Protection of the Audiovisual Heritage [11]. That document states that *"... moving image material is a form of cultural expression reflecting contemporary society ... it is an excellent means of recording everyday events, the basis of our history and a reflection of our civilisation;*". This material is also commercially valuable.

The problem is: all this material is at risk, and 70% is already deteriorating, or is inaccessible because it is fragile or on obsolete formats.
There are two areas where new technological solutions are needed.

- Better solutions for the actual conversion from deteriorating, fragile and obsolete formats to sustainable formats;
- An overall solution for combining preservation with access, in a real system capable of providing large-scale public and commercial access.

The main deliverable of the PrestoSpace Project is to develop actual facilities and launch services for audiovisual preservation. The project will start these preservation factories by preparing the business plan, contacting potential investors and working with commercial partners to set up the actual services. These services will exploit the technological and industrial results of the project. Research has a very strong implication on the whole project. In each area solutions have to be found in order to permit the development of new tools or processes. Numerous outstanding research centres are involved in the project in order to develop the building blocks needed for technological development.
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*, concerned with content description and retrieval.

1.2. Project roadmap

The project started in 2004 and has duration of 40 months. Currently, at the end of the second year, most of the envisaged technologies have been developed in the various areas. The first operational tests are being organised under the supervision of a User Group, including many audiovisual archive owners representing both large and small archives, in different domains, from broadcasting to museums. The next and final stage will be mainly devoted to the completion and assessment of the factory components developed and to training and dissemination of the achieved results. A parallel activity has been started to devise and promote a business model that will provide guidelines and strategies to launch sustainable preservation services in Europe based on the project achievements.

2. The Metadata Access and Delivery Area

2.1. Metadata in the archive management process

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 [1].
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.
In the view of the broadcast archives scenario, this entails finding what information schemes are needed to make archive users retrieve audiovisual items with effective levels of accuracy [2], [3].
In this domain, four basic retrieval patterns can be identified:

- *Retrieving audiovisual items by information*. Starting from the specification of metadata constraints the material for which the stated constraints are valid is to be retrieved. This is the traditional use of the information as "metadata".
- *Retrieving information by audiovisual item*. The access to the archive information relies on the audiovisual material as the carrier of the pieces of information the users are interested in.
- *Retrieving information by information*. In this scenario, information is reached through the use of other pieces of information that act as "metadata" with respect to the target information.
- *Retrieving audiovisual item by audiovisual item*. Audiovisual material is sought and retrieved by means of similarity searches based exclusively on the audiovisual content, i.e. regardless of the expressed meaning and content.

The PrestoSpace MAD partners have undertaken a thorough analysis of these aspects coming to the conclusion that the required information for a typical audiovisual archive exploitation process can be divided in the following fundamental classes:

- *Identification information*, e.g. titles, credits, programme publication information.
- *Editorial parts information,* i.e. information about the relevant editorial sub-items of a programme (e.g. news items).

- *Content-related information,* e.g. text of speech transcript, topics, descriptions, aural and visual low level descriptive features.
- *Enrichment information,* i.e. information coming from external sources generically or topically related to the programme content

2.2. Data models and standard formats

Once the required classes of information had been discovered and verified with end-users, the subsequent work was that of producing a data model representing these classes, together with a data format capable of implementing the data model without any loss.

This was achieved through creating a single XML-based document format, taking the best from each of two metadata standards natively orientated to the description of audiovisual objects, MPEG-7 [4] and P_META [5].

Being among the two the closest to the television domain, P_META was adopted due to its complete set of information structures for the identification, classification and publication-related features of a programme.

The MPEG-7 standard was adopted due to its powerful temporal segmentation tools and for its comprehensive set of standard audiovisual descriptors. In order to achieve clear semantics of the elements in the MPEG-7 description, the Detailed Audiovisual Profile [10] is used for the MPEG-7 part of the format.

Figure 1 schematises the resulting document format, also indicating the ad-hoc data structures introduced to realise those information structures not covered by any of the two standards.

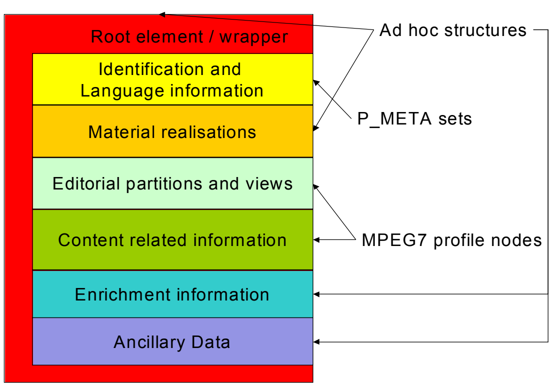

Figure 1 - MAD document format schema.

What is worth being mentioned too, is the use of the SMPTE UMID [6] standard for the unique identification of the instances of audiovisual material throughout the whole platform. This includes original media, digitally re-mastered media, and all the material created during the documentation process (e.g. key frames).

2.3. General architecture

The MAD Platform adopts a modular, extensible architecture. As you can see in Figure 2, the MAD Platform receives as input the digitised media (video and audio files) and returns different materials, such as key frames, camera motions and metadata. These materials are indexed and published on a web server using the Publication Platform (see section 4).

The MAD Platform is made up of a core component (the *Core Platform*) and a set of pluggable software processors named *GAMPs* (acronym of "Generic Activity MAD Processor").

The Core Platform offers the following main services:

- the Workflow Management service, which is responsible for starting processes in the correct order and for resolving dependencies between GAMPs;
- the Essence and Metadata Storage (EMS) system, which stores the audiovisual material sources and the associated metadata;
- the Concurrent Versioning System, which tracks every change to the metadata that takes place during the execution of the various GAMPs and is built on a standard CVS engine;
- the Delivery systems providing access to the enriched metadata and related materials created within the Documentation Platform.

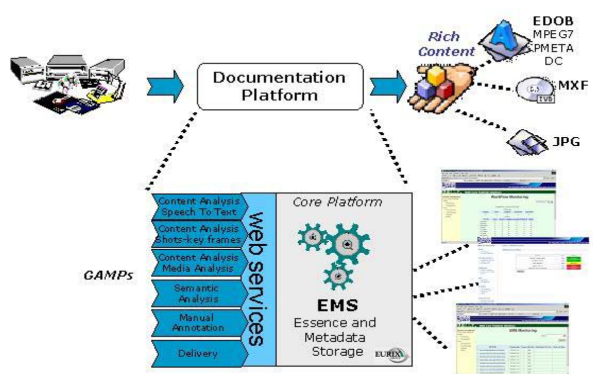

Figure 2 - Architecture of the MAD Platform.

The EMS stores the materials on the file system, and tracks their location by means of a relational database. It's possible to have many copies of the same material, even located on various machines and accessible through a set of protocols (file, HTTP, FTP, SMB, ...).

The GAMPs are the software units that extract the metadata from the digitised materials. The Core Platform maintains a queue in the workflow for every GAMP, which will poll it in order to become aware of any activity to be done. In order to do their jobs the

GAMPs ask the Core Platform for the materials and the associated metadata produced up to the request time.

The current experimental instance of the MAD Platform makes use of three different categories of GAMPs: *Content Analysis*, *Semantic Analysis* and *Manual Annotation*. However any kind of new GAMP can be easily added in the future.

The overall services offered by the Core Platform are available through the following Web Service (SOAP based) interfaces: Workflow, EMS , Administration

Using these services, every GAMP polls the Core Platform asking for a job and related resources and then submits the produced data and notifies the completion of the job to the Workflow Manager. The use of web services implies that the GAMPs could be written in different programming languages and deployed on totally different platforms and operating systems. The overall architecture offers several benefits as follows:

- Modularity: GAMPs can be totally different in functionalities and implementation details, and they still interoperate with the Core Platform;
- Extensibility: in order to insert a new GAMP it is sufficient to add a new process queue to the Workflow engine of the Core Platform;
- Platform independency: the GAMPs can be written in any language, which supports SOAP and web services protocols. The Core Platform itself is written in Java and portable itself to many operating systems;
- Multi-tier distribution: every GAMP can be installed on a different physical system, provided that a network link to the Core Platform exists.
- Furthermore the Core Platform components (as the EMS, the workflow engine and the DBMS) can be installed on different servers as well.

3. Documentation infrastructure

3.1. Content analysis

Automatic Content Analysis tools for audiovisual content are used in the PrestoSpace project to automatically extract metadata from the material and augment the description of the content. The automatically extracted metadata is used to support manual annotation by content structuring (cf. Section 3.3.), to provide input to semantic analysis tools and to index the audiovisual content. We have surveyed the state of the art of the tools for visual, audio and joint audiovisual content analysis in order to assess the feasibility and maturity of the approaches and determine the relevance of the available tools to the archive documentation process. The results of this survey can be found in [7]. The feedback coming from the actual use of the MAD platform by the User Group will enhance the possible advancements of the current state of the art that the project will be able to achieve.

The audiovisual content analysis tools described in the following have been selected for use in the documentation process and integrated using the architecture described above. Due to the modular and extensible architecture of the documentation platform, more time-consuming analysis jobs can be distributed among several clients. The results of content analysis are described using the metadata model and format presented in Section 2.2.

3.1.1. 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 (cuts) and gradual transitions (such as dissolves, fades, wipes, etc.). 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. The approach used for shot boundary detection is an improvement of the one described in [9].

3.1.2. 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 (cf. Section 3.1.8.). 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.

3.1.3. 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. The algorithm is based on feature tracking. The details of the approach and evaluation results can be found in [8]. Camera motion information is an important search criteria when reusing archive material in new productions and to infer higher-level information.

3.1.4. 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 (Istituto per la Ricerca Superiore di Trento) [16], capable of extracting text from English and Italian.

3.1.5. 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.

3.1.6. Editorial parts segmentation. Any modern television archivist considers editorial parts as the basic indispensable entities for the documentation of an archived programme. They can be defined as the constituent parts of the 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 [7], 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.

3.1.7. Reference video clip detection. Reference video clip detection is the task consisting in detecting replica of a reference clip into a visual content. This simple activity is particularly useful when applied in broadcast archived material, where jingles, colour bars and other effects are used as visual separators between the parts of a programme. The produced information constitutes one of the inputs to the more complex editorial segmentation task.

3.1.8. 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 ([4], 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.

3.2. Semantic analysis

The semantic analysis of news items is carried out automatically by two different modules: the linguistic processor and the news categorizer. The former makes grammatical recognition capabilities over both transcriptions and Web pages available and allows recognizing relevant Named Entities. The latter is a text categorizer based on machine learning techniques. It provides the semantic category (e.g. sport, politics) of the incoming news items. Additionally, a module for news alignment selects a suitable set of pages retrieved from the Web, which report on similar facts contained in the target news.

3.2.1 Linguistic Processor. The linguistic processing is carried out by a Natural language parser called CHAOS [12,14]. CHAOS is a server for modular and lexicalized parsing based on a cascade of morph-syntactic recognisers. The main CHAOS modules are: the *tokenizer*, the *morphological analyzer* (who identifies the possible morphological interpretation of each token), a *part of speech tagger*, a *named entity recognizer*, a *chunker* (that collects possibly multiple tokens to form bigger grammatical and unambiguous units called *chunks*), the *temporal expressions recognizer*, a *verb sub-categorization analyzer* (for the recognition of the main verbal dependencies) and a *shallow syntactic analyzer* (for the recognition of remaining and possibly ambiguous dependencies, e.g. prepositional modifiers of verbs and nouns). We apply the CHAOS modules to extract Named Entities and temporal expressions from the target news. Note that temporal expressions are traditionally seen as a kind of Named Entities and that to recognize these latter we need to use the basic modules such as the *tokenizer*, the *morphological analyzer* and the *part of speech tagger.*

3.2.2 News Categorization. Semantic categories are automatically assigned by a text classifier based on a traditional supervised machine learning model. We used an extended version of the profile-based classifier, known as the Rocchio model [13]. Although slightly below[1] the state-of-art (e.g. Support Vector Machines, [17]), these methods have beneficial properties as for the training efficiency and for their relatively simpler impact on the integration of complex software architectures. Given a set of training documents R_i, classified under the topics C_i (positive examples), a set $\overline{R}_i$ of the documents not belonging to C_i (negative examples), a document d_h and a feature f, the Rocchio model [13] defines the weight Ω_f of f in the profile of C_i as:

$$\Omega^i_f = \max\left\{0, \frac{\beta}{|R_i|}\sum_{d_h \in R_i} \omega^h_f - \frac{\gamma}{|\overline{R}_i|}\sum_{d_h \in \overline{R}_i} \omega^h_f\right\} \quad (1)$$

where ω_f^h is the weight of the feature f in the document d_h. In formula (1), the parameters β and γ control the relative impact of positive and negative examples and determine the weight of f in the *i-th* profile. These parameters depend on the training corpus and different settings of their values produce a significant variation in accuracy.

[1] See for example [18] for an extensive comparative study.

In [13] a method to choose optimal parameters is presented. It makes use of a unique parameter γ_i as follows:

(2) $$\Omega_f^i = \max\left\{0, \frac{1}{|R_i|}\sum_{d_h \in R_i}\omega_f^h - \frac{\gamma_i}{|\bar{R}_i|}\sum_{d_h \in \bar{R}_i}\omega_f^h\right\}$$

Since each category C_i has its own set of relevant and irrelevant features, it will have its own γ_i parameter. The optimal γ_i value is estimated by considering the one that maximizes the classification accuracy (of the classifier *i*) on a validation set. The Rocchio classifier figure on the Reuters benchmark is about 83% (f1-measure), close to state-of-the-art classification models, which are in general more computationally expensive [13].

3.2.3 Web Alignement. We developed, a simple spidering process based on the Google's API, to retrieve all the documents published in a target temporal window. This is centred on the broadcasting day by adopting a symmetric span (in the *future* and in the *past*). Accordingly, the temporal distance between the retrieved news and the source transcribed segment is considered as an inversely proportional ranking score. The overall alignment criteria is based on the summation of the following scores:

- Document similarity score, provided by our search engine (a values between 0 and 1) to the candidate web page.
- The time distance *D* between the target news and the candidate web page (normalized with respect to 24 hours).
- The coherency score between the categories of the target news and the candidate web page. We manually defined a contingency table whose entries are scores between category pairs, e.g. to *Politics* and *Foreign politics* pair is assigned a score of 0.7.

We tested our approach with the web site of the Italian newspaper "*La Repubblica*". This publishes news categorized by legacy metadata (e.g. a set of 8 newspaper categories) in standard HTML. Also the news date, e.g. "January, 12, 2003" is made available. In our preliminary test the alignment results were quite satisfactory, BEP of about 65%-72%, [15].

3.3. Manual annotation

The use of automatic content analysis and semantic content 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.

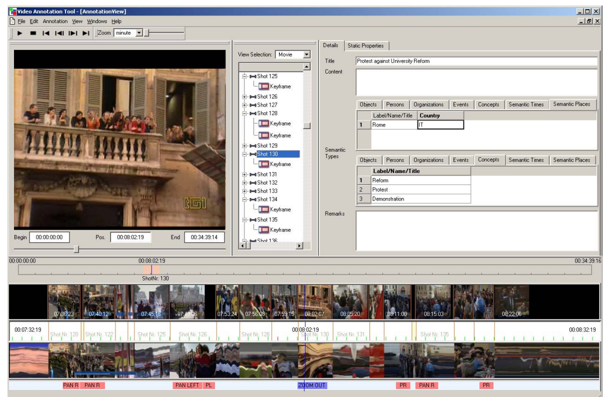

Figure 3 -Interface of manual annotation tool.

These functionalities are provided by a manual annotation tool, which is part of the documentation infrastructure. Depending on the workflow, it can be used at any stage in the documentation process and it will make use of the results that have been generated so far by the automatic tools.

The tool supports the user in efficiently navigating the content by providing a timeline representation (with a adjustable temporal range) showing time-dependent analysis results, which are synchronized among one another and with the player. Figure 3 shows the current prototype of the manual annotation tool. The views in the lower part display content analysis results. Shot boundaries, key frames and camera motion segments may also be edited by the user. The component in the upper centre visualizes a tree structure of the video. It allows relating segments generated by semantic analysis or defined by the user (e.g. news stories) to automatically extracted segments (e.g. shots). The component in the upper right allows the annotation of a number of metadata elements related to the segment selected in the tree. The components for textual annotation make use of controlled vocabularies, and in a next step, they will be fully integrated with the knowledge base used in the semantic analysis subsystem.

4. Publication infrastructure

The Publication Platform provides retrieval and browsing functionalities. It deals with instances of documents in MAD metadata format making them available in a web-based representation and gives access to the materials exported from the Core Platform.

The platform architecture is based on three main components: a web application that serves as the user interface; a DBMS that stores the information related to the available programmes; a text search and indexing engine (Lucene). It is planned to integrate a semantic engine for processing natural language queries.

In accordance with the J2EE guidelines, the Publication Platform is delivered as a Java web archive (WAR), and provides a common web interface for importing, searching, querying and browsing metadata related essence documents.

The Publication Platform offers two main features:

1. data import for submitting materials to be indexed and published
2. data search and browsing

The search interface supports various retrieval approaches and the user can choose to search for a programme or a news item, which can be filtered by programme title, broadcast date and service, contributions (authors, journalists, directors, ..), classification (topics, categories), text of description.

Full text search is applied on the results obtained from the speech-to-text GAMPs, indexed by Lucene. The user can define multiple criteria and can also make use of boolean operators to refine the query. Results are shown as a table that allows the user to sort the items by the column value (title, subtitle, date …). 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 of four frames: a video preview, the editorial parts tree, the key frames, and a multi-tab frame (further tabs could be added in the future), each of which is assigned to represent 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* (see Figure 4). This tab shows the output of the speech-to-text GAMP. The text is divided into segments representing individual news items. The interface also allows the user to select a specific text segment.
- *Semantic analysis* (see Figure 5). This tab shows a navigable tree that can be explored interactively. It shows the entities found by the semantic GAMP.
- *Content analysis*. Here the user can view the stripe images and the related camera motions on the timeline.

The video preview makes use of Windows Media Player, which is the only platform-specific feature currently used. Everything else works fine on whatever operating system and browser, and a custom implementation of a different video player can be done in the future. Using the video player, the seek-able movie let the interface able to synchronise all the tabs on the current timestamp. For example if the user selects a key frame taken just after three minutes from the beginning of the programme, the video will start automatically from that time.

The user can make use of different browsing approaches. It's possible to navigate the news items of the programme using the tree just below the player, and at the bottom you are looking at key frames about shots segmentation.

By selecting the corresponding tabs it is possible to view information about programmes and news items, to view audio transcripts aligned with the timeline, and classifications on semantic analyse and content.

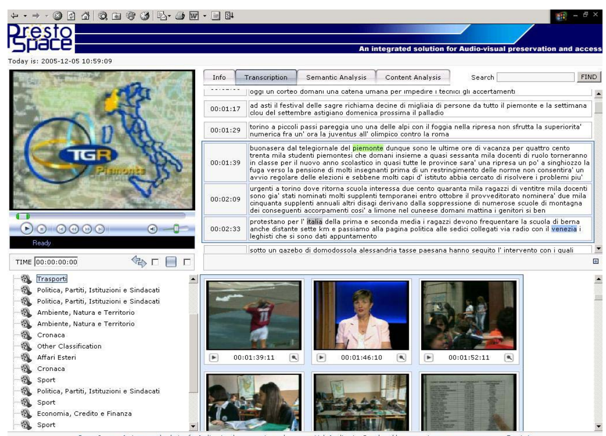

Figure 4 - Speech to text visualization.

5. Conclusions and future work

The results of the work done so far show that multimodal analysis can be used effectively for automatic extraction of indexing and descriptive information from news content. Even if a formal evaluation has not yet been carried out, the level of accuracy obtained seems to be adequate to full text search on the transcript as well as for the more ambitious contextual retrieval based on the exploitation of semantic information.

The next steps will involve the extension of this approach to other content genres and languages, as well as the elaboration of users' feedback about the quality and usability of the results achieved with the current prototype.

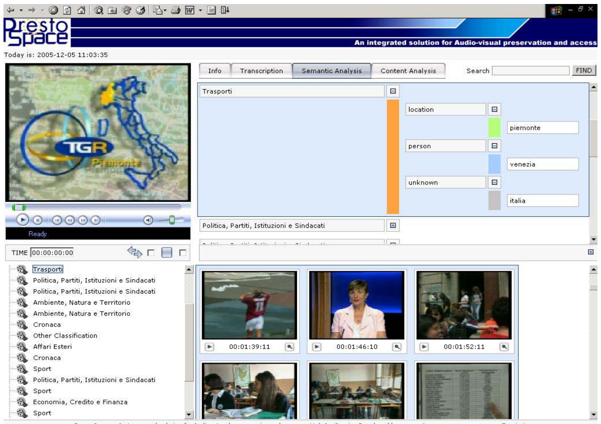

Figure 5 - Semantic analysis visualisation.

6. References

[1] R. Del Pero, G. Dimino, and M. Stroppiana, "Multimedia Catalogue – the RAI experience", *EBU Technical Review nr. 280*, European Broadcasting Union, Geneva, Summer 1999, pp. 1-13.

[2] A. Messina, and D. Airola Gnota, "Automatic Archive Documentation based on Content Analysis", *IBC 2005 Conference Publication*, International Broadcasting Convention, Amsterdam, September 2005, pp. 278-286.

[3] A. Messina, "Documenting the Archive using Content Analysis Techniques", *EBU Technical Review nr. 305*, European Broadcasting Union, Geneva, January 2006.

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

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

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

[7] W. Bailer, F. Höller, A. Messina, D. Airola, P. Schallauer, M. Hausenblas, *State of the Art of Content Analysis Tools for Video, Audio and Speech*, Deliverable 15.3 of the IST PrestoSpace project. URL: http://www.prestospace.org/project/deliverables/D15-3_Content_Analysis_Tools.pdf, March 2005.

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

[9] W. Bailer, H. Mayer, H. Neuschmied, W. Haas, M. Lux, W. Klieber, "Content-based video retrieval and summarization using MPEG-7", *Proc. Internet Imaging V*, San Jose, CA, USA, Jan. 2004, pp. 1-12.

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

[11] European Convention for the Protection of the Audiovisual Heritage, European Treaty Series, No. 183; Strasbourg, 8.XI.2001.

[12] Basili, Roberto, Pazienza, Maria Teresa, Zanzotto, Fabio Massimo, *Efficient Parsing for Information Extraction*, Proceedings of the European Conference on Artificial Intelligence (ECAI98), Brighton, UK, 1998.

[13] R. Basili, A. Moschitti, M.T. Pazienza, "*NLP-driven IR: Evaluating performance over a text classification task*", In Proceeding of the 10th "International Joint Conference of Artificial Intelligence" (IJCAI 2001), August 4th, Seattle, Washington, USA 2001.

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

[15] Roberto Basili, Marco Cammisa, Emanuale Donati, *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.

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

[17] T. Joachims, *Learning to Classify Text using Support Vector Machines*, Kluwer, 2002.

[18] A. Moschitti, R. Basili, "*Complex Linguistic Features for Text Classification: a comprehensive study*", Proceedings of the European Conference on Information Retrieval, ECIR, 2004.