Documenting the Archive — using content analysis techniques

Alberto Messina
RAI CRIT

The purpose of this article is to provide EBU Members with basic information on the possible benefits of employing *content analysis techniques* for documenting their television and radio archives. It is based on the considerable experience gained by RAI over recent years in this field.

The article also looks analytically at the impact these new archive documentation techniques will have on traditional working practices.

In recent years, two major events have revitalised the interest of the broadcast community in automatic information-extraction tools:

a) the phenomenal growth of the Internet; and

b) the dramatic drop in computer prices.

The former has made it possible for generic users to exploit tons of information that was previously unimaginable. This has given researchers in Information and Communication Technology (ICT) the means of finding efficient methods to represent, exchange and search for this immense patrimony of knowledge. The tools which allow us to achieve these goals range from new knowledge representation languages (e.g. OWL) to the application of already-known or recently-invented artificial intelligence techniques (e.g. SVMs, MLPs and HMM to cite the most famous). Among these, ***information extraction algorithms*** play a fundamental role, namely that of making it feasible to extract implicit knowledge from raw data (which is the origin of the "content analysis" concept, depicted in *Fig. 1*).

Typically, the deployment of such techniques entails the use of substantial computing resources – be it in terms of the memory capacity needed or just the pure number-crunching requirements. This is particularly true for algorithms which operate on multimedia information sources such as video and audio, due to their intrinsic weight.

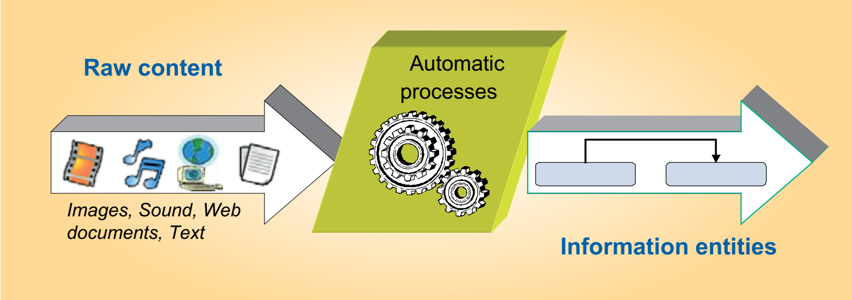


Figure 1
The content analysis concept

Whilst many of the current cutting-edge approaches to multimedia content analysis find their roots some decades ago, they were unable to provide truly viable solutions, due to their immense computing complexity. That is, not until computing platforms unleashed their true awesome power in the last decade or so. Powerful knowledge representation methods, along with unequalled and readily-available computing capacities today, have made it possible to restart the investigation – today, we can apply artificial intelligence machinery to industrial domains with renewed vigour.

As a consequence of this, broadcasters – some of whom are well known for possessing important historical audiovisual archives – obviously represent a primary customer for the application of such techniques. But how can this be realised in a sensible way?

Abbreviations

ASR	Automatic Speech Recognition
HMM	Hidden Markov Model
MLP	Multi-Layer Perceptron
NLP	Natural Language Processing
OWL	Web Ontology Language *(instead of "WOL")*
SVM	Support Vector Machine

Which tools offer what to the archive documentation domain

The archive perspective

From the perspective of large archive owners such as RAI and several other EBU Members, automatic information extraction tools are currently seen as an essential aid to lowering the costs associated with the archive documentation process. The idea of using powerful computers to do the job normally accomplished by skilled human beings [1], clearly represents an extremely valuable opportunity for these broadcasters. This is especially true for archive material genres that do not require high levels of sophisticated annotation to reach an acceptable level of documentation quality, e.g. television newscasts.

However, practical and theoretical evidence shows that the dream of substituting expensive teams of documentalists with dumb clusters of silicon chips is destined to fade away – if not pursued rigorously. This is due to some incontrovertible facts relating to the nature of these machines:

- ***Training issues***

  Automatic information extraction tools have to be "trained" with input data sets that act as "training patterns". They perform better after finding pieces of information that fall within the statistical ranges of the training examples. The number of training phases required tends to equal the potential number of input material categories that are statistically distinct, relative to the information to be extracted. Furthermore, the number and features of these categories may alter over time, depending on the variability of the expressive language used (e.g. news programmes may change their "scenography", "shooting style" and overall "programme format" quite often).

- ***Quality of the results***

  The levels of ***precision*** [2] and ***recall*** [3] become lower as the semantic complexity of the information to be extracted increases. Paradoxically, this aspect concentrates human actions on verifying the most complex information layers where, at the same time, the cost of the intervention is much higher than in the lower layers (because of the higher levels of quality required).

[1] Not allowing for any social considerations associated with the preservation of employment levels.

[2] Here, precision = the ratio between the number of correct information "detections" and the total information "detections".

[3] Here, recall = the ratio between the number of correct information "detections" and the total information present in the material, as would be pointed out by a perfect detector.

- ***Impact on the processes***

  In terms of systems' management, careful studies of the costs implied in setting up, upgrading and developing the systems which host the automatic extraction tools, have to be undertaken. Currently, there is no concrete information on what the impact would be on existing archiving environments if such tools were introduced as large-scale industrial processes. Due to the need to verify the outputs produced by automatic machines, the total efficiency of the documentation process for a certain piece of material may be affected dramatically, even if timely optimizations are made at various points of the chain.

Background: recalling the objective

RAI's experience in this domain has shown that the role of the ***documentation model*** is still of primary importance, even where automatic processes enter the scene as new actors. But in what sense is this true?

To give a synthetic (and imperfect) definition, a documentation model can be defined as a system with which to classify archive items under a closed set of data structures called ***information entities***. The information entities taking part in a documentation model are the means through which we represent real-world entities (physical objects but abstract concepts as well), in a concrete information system *(see Fig. 2)*.

The prime objective of any real-world information system that aims to represent this universe, is to provide concrete means for the ***exploitation*** of the archive items that are represented within it. “Exploitation” can be seen as the activity where the user actually *uses* the archive items. On the other hand, ***fruition*** is the activity where the user simply acquires some information by reading/viewing/hearing the information stored in the system.

Archived items can be of two distinct classes: audiovisual material and information. This distinction, perhaps odd at a first glance, is of key importance. In fact, if the former represents the classical case in which archive users want to obtain the actual pieces of audiovisual material from the archive,

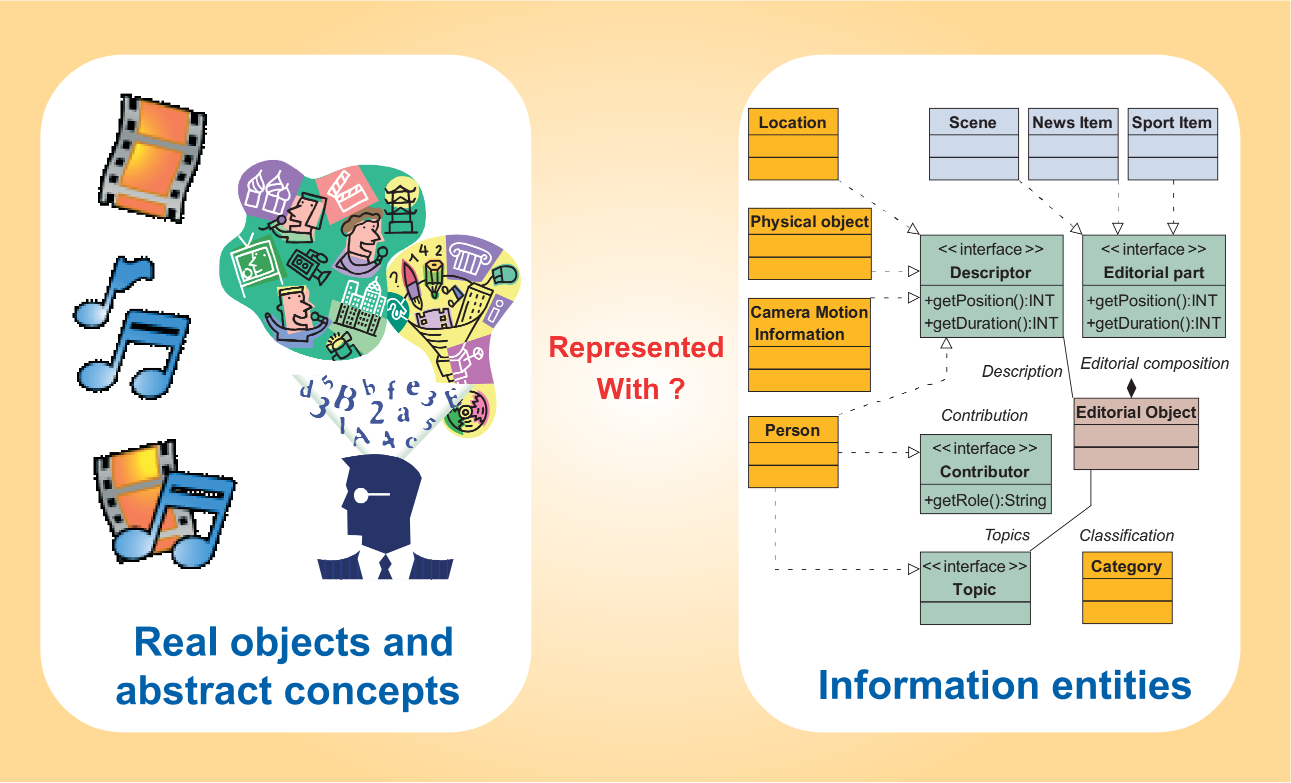


Figure 2
Representation of real-world objects with information entities

the latter refers to the case where information is the principal objective of the users' interaction with the archive. Examples of this can be found in situations where a user is interested in knowing only whether a certain event has happened, or in reading the descriptive details of a location and so on.

In this interpretation, the traditional concept of ***metadata*** also unveils new nuances: in a retrieval process, a piece of "metadata" is an archive item that functions as an informative link towards another item. Four cases are possible:

- ***Retrieving an audiovisual item by means of information***

  This case represents the archive exploitation *per antonomasia*. If the starting point is a specification of the constraints that the associated information should conform to (e.g. assumed values or ranges for attributes, validity of certain relations), the finishing point is actual retrieval of the audiovisual material for which the stated information constraints are valid. This is the traditional way of using the information as "metadata" to retrieve the required A/V material.

- ***Retrieving information by means of an audiovisual item***

  This is a more sophisticated way of accessing archive information. Here, the audiovisual material itself is the carrier of the information the user is interested in. It should be noted that the final objective of the exploitation in this case is information only. The starting point (carrier material) can be reached by any of the four cases indicated here.

- ***Retrieving information by means of information***

  In this scenario, the aim of the user is again to find information. But in this case, it is achieved by using other pieces of information that act as "metadata" for the target information. Typical examples are: finding the topics covered by a particular character in his/her speeches, looking for details about certain historical events.

- ***Retrieving an audiovisual item by means of an audiovisual item***

  This category covers the cases where 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 underlying semantics. The typical instruments used to achieve this kind of exploitation are content-based queries-by-example that make use of audiovisual feature indexes.

In these scenarios, the outstanding concept is that information extraction tools that are based on automatic content analysis have to be assessed on the basis of their ability to provide instances of data with a controllable degree of accuracy.

The recall accuracy, in turn, should be evaluated by comparing the levels of "precision" and "recall" achieved when using content analysis techniques with that obtained using traditional manual annotation. The comparison must of course be made without altering the documentation model or the reference material (learning and test).

In other words, this evaluation model can be thought of as a test system in which, firstly, a full manual annotation task is performed using a certain documentation model D on a set of materials M. In a second phase, still using the same D and M, the documentation task is performed identically by a hybrid documentation system which has automatic extraction tools among its processes. A direct comparison is made by asking the two systems the same set of queries Q. The precision and recall of the retrieved lists of items is then assessed, with reference to the ideal results (known as *a priori*). This is illustrated in *Fig. 3*.

Analysis of the related problems

This section gives a brief summary of various issues relating to the use of automatic information extraction tools.

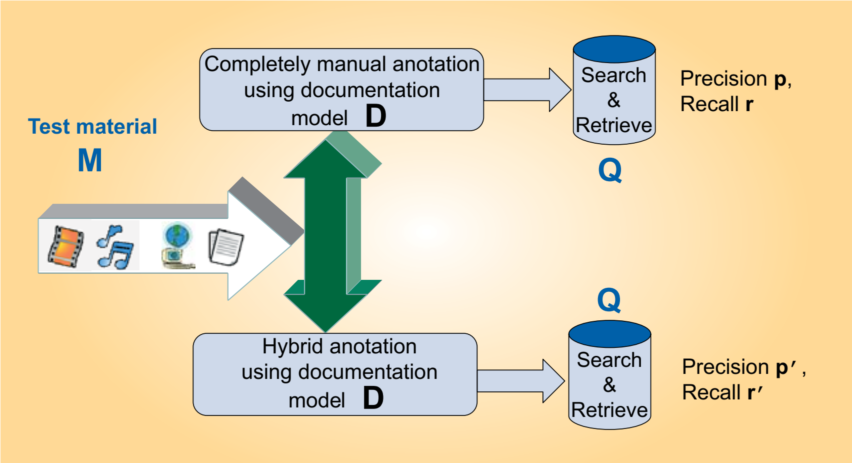

Completely manual anotation using documentation model D
Search & Retrieve
Precision p, Recall r
Test material M
Q
Q
Hybrid anotation using documentation model D
Search & Retrieve
Precision p′ , Recall r′

Figure 3
Comparing retrieval-quality parameters

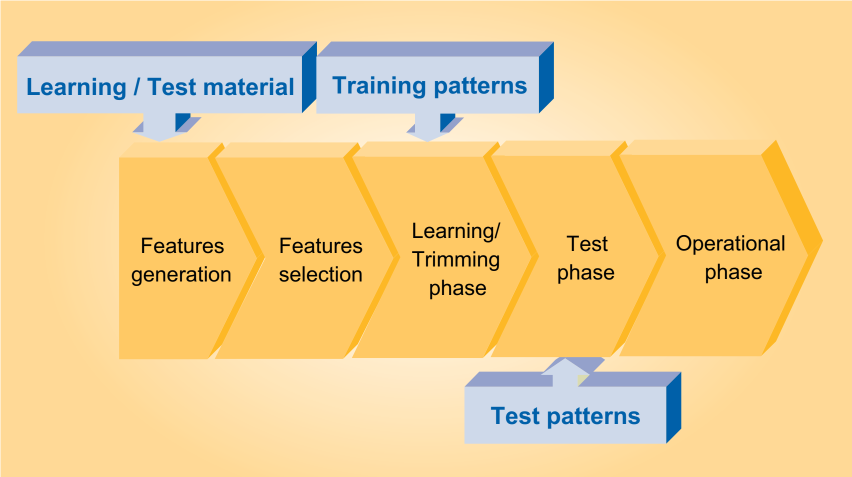


Figure 4
Phases of an automatic information extraction process

Fig. 4 shows a typical integration sequence of an automatic information extraction tool. The first phase, downstream of the test material selection phase, is often called ***feature generation***, i.e. from the selected material a “reasonable set” of features is extracted. Normally this is done on the basis of the type of material (i.e. audio or video) and on evidence and indications found in technical and scientific literature for the class of problems that the tool is designed to address.

In the features selection phase, the most promising extractions from the first phase are selected. In very general terms, the complexity of this task depends on the “separability” characteristics that the classes of information which have to be identified show with respect to selected subsets of the features.

As a trivial example, if the aim of the classifier is to separate black frames from non-black frames in a video sequence, then very likely the tracking of luminance level, averaged over a video frame, is the most promising feature.

In less trivial cases, life can be much more difficult. The learning and trimming phase very much depends on the kind of tool being used. In general it deals with optimization of the cost functions in order to reach a stable set of parameters for the classifiers used. In the test phase, the classifier is tested against patterns that were not in the training set, in order to assess the general performance of the system. Finally, if all these converge, the operational phase can take place. However, there are very challenging problems connected with this process, namely:

- ***High-demanding data elaboration requirements***

  Normally, broadcast archives contain many thousand hours of material to be documented, and the rate at which an active archive grows may reach several dozens of hours per day. This leads to two problems: on the one hand, a great amount of material means a high availability of test patterns but, on the other hand, it rapidly introduces new types and paradigms. This has an impact on the stability of the training parameters and on the number of classifiers that have to be employed. It is very likely that re-training has to be accomplished quite frequently and that new classifiers have to be added during the system's lifetime. In turn, this may entail re-elaborating part of the archived material.

- ***Operational domain variability***

  Classifier training usually gives better performance when the classes that have to be identified and associated to the reference patterns show a higher separability factor, i.e. when there exists a combination of content-related features that allow the determination of a sharp classification boundary. Due to the diverse operational requirements of a broadcast archive, material is rarely classifiable once and for all in an absolute way (e.g. a piece of material can be classified under different specific categories depending on the use that is planned for it, rather than simply on its content characteristics alone).

- ***Application domain mismatch***

  Content analysis and information extraction techniques may be conceived in the context of application domains whose origins and goals may be far from the actual requirements of a broadcast archive documentation process. To give a fashionable example, many current automatic text classification techniques are better suited to work with well-formed textual sources rather than on transcriptions of spoken language (ASR). Domain mismatch problems seriously affect the learning and trimming phase of the algorithms. An important amount of domain knowledge has to be wired into (or around) existing tools, without an *a priori* awareness of the impact that this introduction can bring on the tool performance.

- ***Information granularity misalignment***

  Existing content analysis-based information extraction tools typically work on a piece of input material and provide a set of metadata with a well-defined structure and semantics as their output. The granularity at which extracted data can be associated with the input material cannot be governed by the user (unless it is the user himself that develops the tool as well, but this is not the common situation). For example, a text-based classification tool based on NLP may associate a class label (e.g. "Sport") to a piece of material as a whole, whilst the user might prefer to have a dynamic classification that alters along the material timeline. Furthermore, some classifiers may be constrained by built-in classification schemes that can be difficult (or impossible) to use with legacy or user-defined schemes.

A sensible method

More recently, RAI found it useful to reverse what could be called the “natural” approach to content analysis techniques. Very simply, instead of trying to integrate existing tools in the documentation environment and providing integration components *towards* the documentation model, a more successful approach could be that of starting from the target information model that is to be populated and then, *reversely*, to find out which tools are available and with what level of accuracy they provide their results. This sort of analysis can be called “Process Function Analysis”. Starting from the documentation model specification, the Process Function Analysis identifies the atomic functions that must be provided in order to populate an instance of that model. An example output of a process function analysis is shown in *Table 1*.

The key concept in this case is isolating the elementary documentation process functions (each represented by a row in the table) on the basis of the structure of the documentation model (represented by the “What” columns) and associating a tool function for each atomic functionality (represented by the “How” columns). For example, if the requirement is to provide content descriptions via a description of persons taking part in the scene, then this could be achieved with the use of face and/or voice detection tools based on an analysis of the multimedia content. Again, if the documentation requirement is to identify editorial parts such as scenes in a feature film, then we would need a scene detection and segmentation algorithm.

This strategy brings in some advantages that can be summarized as follows:

- Optimization of tools integration effort. Only tools that can provide functions specific to the information required are evaluated and possibly integrated.
- Optimization of the quality of extracted information. Only by employing the most-promising tools for achieving a certain process function can the quality be optimized.

Table 1
Example output from the process function analysis

What			How	
Area	**Domain**	**Entity**	**Source**	**Result function**
Content	**Descriptions**	Person	Video	Face detection
		Person	Audio	Voice detection
		Location	Video	Location detection
		Physical Object	Video	Object recognition, object tracking
		Camera Motion Information	Video	Camera motion detection
	Topics	Person	Text	Named entity extraction
	Classifications	Category	Text	Text semantic analysis
	Classifications	Category	Video	Location classification (e.g. outdoor/ indoor)
	Descriptions	Silence	Audio	Silence detection
	Text	Text	Audio	Automatic speech transcription
Identification	**Contribution**	Person	Video	Video to text transcription
	-	Awards	Text	Web crawling, web mining
Editorial parts	**-**	Scene	Video & Audio	Scene detection and segmentation
		News Item	Audio	Speaker recognition and clustering
		Sport item	Audio & Video	Highlight event detection

The tools selection phase should take into account on the one hand, the level of “precision” and “recall” that the existing state-of-the-art can offer. On the other hand, being used in an industrial context, it should take into account the amount of effort needed to integrate the existing tools into the documentation infrastructure (for example, in terms of update mechanisms, configuration parameters, execution environments etc.). It is important to stress that the overall decision whether or not to employ an automatic machine that implements one of the found atomic functions should be supported by a weighted combination of both these aspects: tools showing great “precision” and “recall” but with a flawed software architecture should be rejected, just as should tools with excellent software engineering characteristics but which provide poor results.

Brief state-of-the-art classification

A complete state-of-the-art analysis is beyond the scope of the current article, which offers mainly an introduction to the subject. For the interested reader, a good analytical review can be found in [1], while a thorough treatise on the fundamental theory can be found in [5]. Nevertheless, in the context of this article, automatic content-analysis-based information extraction tools can be seen as falling into the following categories:

- ***Low-level feature-based representation methods (LL)***

  To this class belong all the algorithms whose purpose is to extract basic features from audiovisual content and to represent them in data structures suitable for direct retrieval, based on certain similarity features. In the case of video, these features are typically related to **colour** (e.g. colour histograms), **texture** (e.g. texture signature vectors made up of contrast, granularity and directionality measures), **shape** (e.g. curvature and orientation vectors) and **motion** (e.g. motion vector fields). In the case of audio, the typical extracted features relate both to the **frequency** domain (e.g. median frequency, bandwidth, spectral coefficients) and to the **temporal** domain (e.g. zero crossing rate, short time energy). This layer of techniques is based on the assumption that basic features, such as colour, shape, texture, motion and frequency are the dimensions under which human perception classifies images, sounds and video sequences prior to their semantic interpretation. The quality assessment of LL tools should be done according to their precision in representing the salient perceptual aspects of audiovisual material.

- ***Media structuring techniques based on formal features (FS)***

  These techniques frame on LL outputs in order to extract structural information from audiovisual content. A classical problem that FS tries to address is the "shot-detection" problem, i.e. the reconstruction *a posteriori* of the sequence of sensor commutations, either visual or aural, that have taken place during the recording of an audiovisual material. A broader case involving shot detection is the generic "video indexing" problem, i.e. the production of compressed structures of temporal and visual information that take account of the essential aspects of a video sequence, allowing for non-linear content-adaptive consultation of material. In the audio domain, a representative counterpart is the "audio segment classification" problem, i.e. the classification of audio segments into silence, speech, music, noise.

- ***Relevant objects detection and identification methods (OD)***

  This class of techniques relies heavily on LL and partly on FS and their purpose is to address problems relating to the detection and identification of relevant objects by various means associated with the audiovisual content. The assessment of the relevance of objects is highly dependent on the kind of target description which, in turn, is strictly related to the type of application that uses the output of OD methods, as well as to the nature and structure of the selected information model. The most common problems finding use of OD methods are face detection and person recognition problems, physical object detection and tracking.

- ***Content classification techniques (CC)***

  Content classification is considered as one of the basic functionalities expected from the automatic elaboration of audiovisual content. Typical approaches are based on the selection of a proper probabilistic model that is able to maximize the likelihood of the set of low-level features data taken as the training set, and subsequently to use such models to classify the new items. Classification techniques can take into account higher-level information as well, such as OD or MS information, to improve the quality of the probabilistic models. Text-based methods perform the classification task by exploiting statistical or linguistic models, typically downstream of an automatic speech-to-text transcription. Another aspect strictly relating to classification is "class labelling", i.e. the association of keywords with the identified classes, as a first attempt at collating feature-based classifications with the actual meaning expressed by audiovisual content.

- ***Semantic-driven structuring techniques (SS)***

  Very broadly speaking, semantic extraction methods are basically built on top of the other techniques (LL, FS, OD and CC) and tackle the still outstanding problem of bridging the gap between the low-level information (typically LL, FS) and the semantic content of audiovisual material as it would be expressed by a human when perceiving the content. More specifically, semantic-driven structuring techniques (SS) aim at identifying the editorial components of an audiovisual work, i.e. the constituent parts that the creator of the work had conceived as such. As the literature clearly shows [5], this task generally depends greatly on the content genre, due to the consistent dynamism and variance that characterizes the television language, e.g. with respect to dimensions such as target audience, programme formats and purpose. For these reasons, facing the problem of identifying scenes in a movie could be completely different, in terms of automatic techniques and successful employment of tools, from the problem of identifying news items, sport items or entertainment programmes' subject stories.

As with any classification system, the aforegoing represents the viewpoint of the author only. Other views on this matter are presented in [2] and [3].

Evaluating the impact on the processes

On the basis of existing practices and of the advancement of the scientific and technical state of the art, it can be fairly concluded that automatic documentation tools are, on average, good candidates for deploying in the archive documentation process. This is true provided that it is not forgotten that human supervision will always be needed in some form (e.g. learning supervision and update, results trimming).

How can the opportunity to adopt, or not to adopt, such tools be evaluated? A possible answer consists in comparing the quality of the extracted information with the human-extracted information, given a fixed error rate.

By making this comparison, we could arrive at a trade-off condition expressed here in mathematical form:

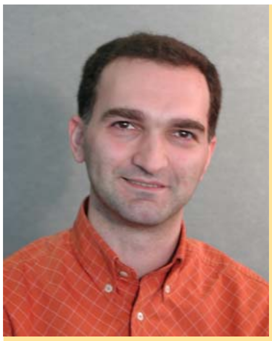
$$\sum_{i=1}^{N} C_i^{E}\left(\varepsilon_i\right) + C_i^{CHK}\left(\varepsilon_i\right) < \sum_{i=i}^{N} C_i^{H}\left(\varepsilon_i\right) \tag{1}$$

... where *N* is the number of identified process functions (i.e. of the rows of *Table 1*) for which the employment of an automatic machine is evaluated, ε is the accepted retrieval error (e.g. the accepted F-measure), C^E is the elaboration cost, C^{CHK} is the checkpoint cost and C^H is the cost of full-human documentation.

Note that costs relating to automatic processes, and costs relating to human processes, are of a different nature. Therefore, to make a sensible comparison, they have to be expressed in some equivalent measurement unit. However, while for human costs it is easier to rely on standard cost unit parameters (e.g. manpower), for automatic components this evaluation could be a bit tricky. Equipment amortization quotes, software licence fees, hardware and software lifecycle costs and incremental infrastructure costs could be important elements for yielding some concrete indications in this direction.

Equation (1) is sufficient to allow for decisions, on the assumption that the introduction of automatic machines has an impact that is limited just to the corresponding functionalities that a specific machine is devoted to accomplish (e.g. automatic speech to text vs. manual transcription). However, a complete impact evaluation should also take into account possible (positive) synergies with the processes and functions that are destined to remain human-based (e.g. the introduction of a face-detection algorithm can lower the cost of analytic programme-participant identification made by a human documentalist).

These considerations, in synthetic terms, mean that the opportunity to employ automatic tools is convenient if the cost of automatic elaboration plus human supervision is lower than the cost of human elaboration at a fixed level of precision/recall on the archive retrieval side. This condition has to be evaluated and averaged over the total number of process functions in which content analysis tools are planned.



Alberto Messina works for the *RAI – Radiotelevisione Italiana* Centre for Research & Technological Innovation, Turin.

He began his collaboration with RAI when he developed his M.Sc. thesis about objective quality evaluation of MPEG-2 video coding. After finishing the university course in 1996, he joined RAI as a research engineer, starting his career as one of the designers and developers of RAI's Multimedia Catalogue. His current interests range from studying the tools and technologies that enable knowledge exchange and representation in the broadcast environment (in the form of file formats, metadata standards and content management systems) ... to the domain of content analysis and information extraction algorithms. Recently, he has started research activities into semantic information extraction from the analysis of audiovisual material.

Involved in several other RAI interdisciplinary projects, Mr Messina formerly collaborated in the EBU project group P/META and is currently an active member of projects P/TVFILE, P/MAG and P/CP. He is also currently working in the EU PrestoSpace project in the Metadata Access and Delivery area.

Conclusions

EBU Members owning important archives may currently be in the position of evaluating the use of automatic content-analysis-based information extraction tools in their archive documentation environment. This could be due to the present revival of interest that such techniques are getting from industrial actors, in conjunction with two modern-day factors: the availability of massive computing power at low cost, and the parallel availability of advanced knowledge representation tools.

Broadcast archives represent an obvious application scenario but, as this article set out to show, a careful analysis should be conducted following a rigorous approach, in order not to waste precious resources in exploring what can turn out to be just useless (though fascinating) toys.

In particular, RAI's experience in this field suggests that documentation models should be retained as the pivotal decision point for the selection and integration of the available tools in the documentation process. Furthermore, extensive studies should be carried out by broadcasters to evaluate the impact that this introduction may have in terms of differential costs.

References

[1] W. Bailer, F. Höller, A. Messina, D. Airola, P. Schallauer and M. Hausenblas: **State of the art of Content Analysis Tools for Video, Audio and Speech**
Deliverable 15.3 of the IST PrestoSpace projoct.

[2] C.G.M. Snoek and M. Worring: **Multimodal Video Indexing: A Review of the State-of-the-art**
Multimedia Tools and Applications, 25 (1): pp 5 - 35, January 2005.

[3] M. Roach, J. Mason, L.-Q. Xu and F.W.M. Stentiford: **Recent trends in video analysis: a taxonomy of video classification problems**
6th IASTED Int. Conf. on Internet and Multimedia Systems and Applications, Hawaii, Aug 12 - 14, 2002.

[4] A. Messina and D. Airola Gnota: **Automatic Archive Documentation Based on Content Analysis**
IBC 2005 Conference Publication.

[5] S. Theodoridis and K. Koutroumbas: **Pattern recognition – second edition**
Academic Press, 2003.