# Integrating research into video and audio digitized archives into textual research: Examples from research on renewable energy

Krishna Chandramouli, Roberta Turra, Giorgio Pedrazzi, Foteini Tsaglioti, Vaso Aggelopoulou and Aristotle Tympas

### *Introduction*

The ongoing digitization of media and other textual archives is already changing research in the history of technology of science [Tzokas et al., this session]. At the same time, research in the history of technology and science (and research throughout the humanities and the social sciences, as well as research by journalists and other professionals and amateurs) is also changing by the availability of non-textual digitized media archives, namely video and audio archives. Addressing the challenges of developing an integrated framework for explicit exploitation of ever increasing audiovisual data, we present an overview of the Papyrus interdisciplinary search engine.

In order to demonstrate the advantages for exploiting audiovisual information for investigating histographical issues, two domains of interest were selected, namely history of wind power and biotechnology (for the choice of the two Papyrus domains, refer to Katifori et al., this session). However, in this paper we will only focus on the wind power domain. A large amount of content received from Deutsche Welle (DW) and Agence France-Presse (AFP) contained audiovisual material. Therefore, in this paper we present an integrated Papyrus framework for extracting implicit knowledge embedded in these multimedia items. A detailed discussion on the construction of the Papyrus Ontology is presented in [Katifori et al., this session].

To set the stage for introducing these contributions, we briefly introduce the historiographical framework that was taken into account during the development of the multimedia framework. This framework is formed by combining general historiographical suggestions on how to properly study technological change,[1] specific historiographical suggestions from the available historiography of wind power,[2] and, suggestions concerning the increasing importance of audiovisual archives for the study of the history of recent technology and science.[3]

Central to this historiographical framework has been the hypothesis that audiovisual archives are not linear extensions of textual archives. There is historiographically important information that we can get only through audio and/or video sources. For example, videos on wind power can offer unique insight on issues concerning the rhetorical/narrative strategies that have been used to discuss the advantages and disadvantages of wind power. In the multimedia framework that we used to develop the Papyrus prototype, we found several such strategies. The list includes the simultaneous display (within the same picture) of old windmills and new wind farms so as to convey a sense of continuity between wind farms and windmills. It also

includes the simultaneous display of wind farms and conventional energy generation installations, especially generation plants that generate visible amounts of smoke. In this case, the visual strategy aims at contrasting the two (wind power and conventional energy installations). In most cases, these strategies don't make it to the text. To research them, a historian of wind power has to study video and/or audio material. Given the availability of a great amount of video and audio, we have focused on how Papyrus could help this researcher to access it. The contributions outlined in the following sections of this paper are focused on the same example, namely Papyrus-assisted research on the relationship between windmills and wind farms as displayed in audio and video resources.

In **Figure 1**, an integrated framework for audiovisual media processing is presented. The video item uploaded to the Papyrus repository is analysed with visual, audio and textual components. The visual analysis tools include shot boundary detection module, highlight extraction module, which is followed by low-level, and high-level feature extraction. On the other hand, the audio stream extracted from the video is further processed with analysis components namely speaker diarisation followed by speech recognition module. The textual transcript output extracted from the ASR module is further analysed to extract NER. In addition, additional processing components available for textual analysis include concept extractor, which could extract concepts such as "hill", etc. The metadata generated from the analysis components are stored in the Papyrus metadata model. Based on the user query, the metadata model is searched and corresponding media items are extracted according to the user requirements.

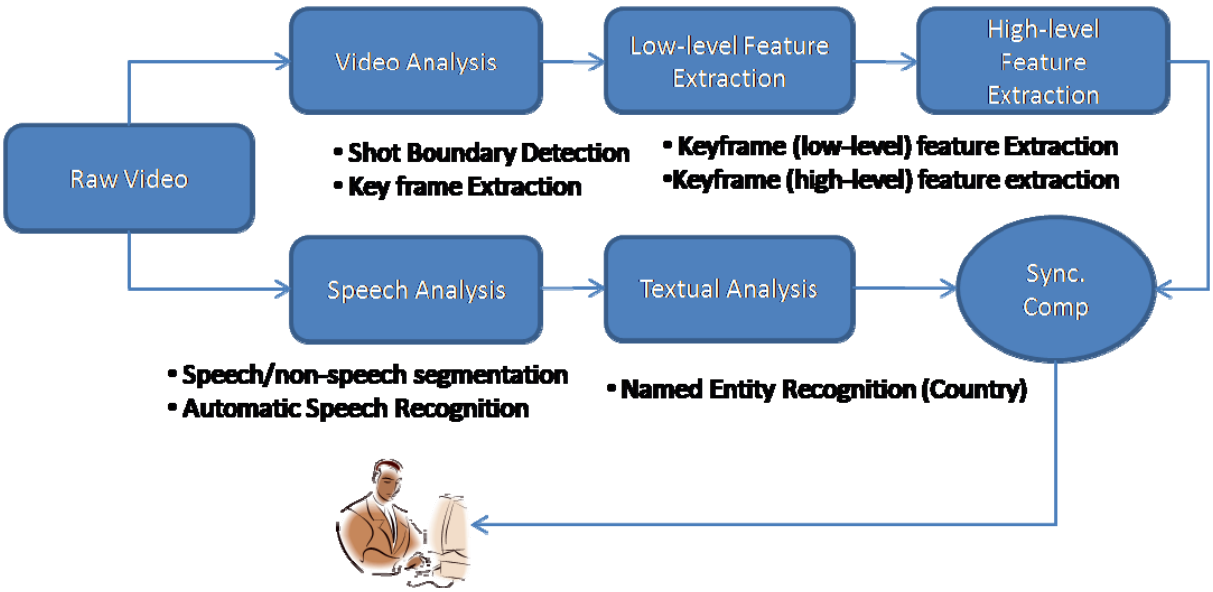


**Figure 1 - A generic framework for multimodal component**

***The video component analysis***

In addition to the above module, continuous research was carried out on developing a temporal segmentation algorithm using MPEG – 7 Colour Layout Descriptor (CLD) [1]. The MPEG – 7 CLD is a compact and resolution invariant representation of colour specifically developed for high-speed image retrieval. However, the computational effectiveness of the descriptor has often been exploited for temporal segmentation of the video. In general, the descriptor is designed to capture the spatial distribution of colour in an image or an arbitrary shaped region. The spatial distribution of colour constitutes an effective descriptor for sketch based region image retrieval; content filtering using image indexing, and visualization. The functionality of this descriptor can also be achieved using a combination of grid structure descriptor and grid-wise dominant colours. However, such a combination would require a relatively large number of bits, and matching will be more complex and expensive. The CLD uses representative colours on a $8X8$ grid followed by a Discrete Cosine Transform (DCT) and encoding of the resulting coefficients. The feature extraction process consists of two parts; grid based representative colour selection and the DCT transform with quantization. The DC values are quantized to 6 bits and the remaining to 5 bits each. These results demonstrate that the CLD is quite effective in image retrieval. The results also compare favourably with a grid based dominant colour approach wherein the image is partitioned and dominant colours for these partitions are used to represent the layout. For matching between two CLD's $(DY,\ DCr,\ DCb)$ and $DY',\ DCr',\ DCb')$, L2 measure is used. For detecting visually coherent scenes, a thresholding scheme is applied.

From the analysis of the Papyrus videos, the challenges of developing a temporal segmentation module include the following

- to account for the transition of the objects in a scene
- to account for fades and dissolves in a shot
- to account for shot and scene changes

Addressing the above challenges, and to detect fades and dissolves types of temporal segmentation, a time-delay module of the same has been developed. The time-delay module accounts for the slow change in the visual characteristics of the shot. In addition, the module also considers the transition (or camera span across a view) between shots to extract the shot boundaries.

***The Keyframe Extraction***

For the extraction of keyframes from the video, a measure of visual dissimilarity is derived by implementing a supervised classifier. The visual dissimilarity derived between frame $f_A$ and $f_B$ is generated, by training the classifier with MPEG – 7 feature set of frame $f_A$ and frames successive to $f_A$ along temporal line such as $f_{Ai}$ where $i\hat{I}\{0,1,..,N\}$ and $N$ is the total number of frames in the video are presented to the classifier as a test set. The classifier output provides a measure of dissimilarity

between frames $f_{Ai}$ with respect to $f_A$. Hence, the algorithm is considered to be a supervised classification, with frame $f_{Ai}$ labelled as positive (or '1') and the successive frames which belong to this class are clustered together as long as classifier assigns label '1' to frames $f_{Ai}$. If a frame in the sequence of $f_{Ai}$ is labelled as '2' denoting a high change of visual dissimilarity, that frame $f_{Ai}^2$ (where 2 denotes the label assigned by the classifier) is considered as the training sample for the successive frames.

In Self Organising Maps (SOM),[4] input patterns are fully connected to all neurons via adaptable weights and during the training process, neighbouring input patterns are projected into the lattice corresponding to adjacent neurons. SOM enjoys the merit of input space density approximation and independence of the order of input patterns. Like K-Means algorithm SOM also needs to predefine the size of the lattice. In basic SOM training algorithm the prototype vector are trained with equations (2-12).

$$m_i(t+1) = m_i(t) + h_{ci}(t)[x - m_i(t)] \quad \textbf{(1)}$$

Where $m$ is the weight of the neurons in the SOM network $h_{ci}(t)$ is the neighbourhood function that is defined in (2-13).

$$hci(t) = \alpha(t)\exp\left(\frac{||rc - ri||^2}{2\alpha^2(t)}\right) \quad \textbf{(2)}$$

Where, $\alpha(t)$ is the monotonically decreasing learning is rate and $r$ represents the position of the corresponding neuron. From the experimental results, it was noted that using a single layer SOM elimination of true negative images by the classifier was limited to those feature vectors, which are represented by the term $x - m_i(t)$ in the training function. Hence, we propose a Dual Layer SOM (DL-SOM) to improve the performance of the SOM. The algorithm workflow and DL-SOM network structure is presented in Figure 2. The evaluation function for the second layer is presented in equation (2-14).

$$m_i(t+1) = m_i(t) + h_{ci}(t)[x + m_i(t)] \quad \textbf{(3)}$$

The output of the classifier is a measure of visual dissimilarity from the classifier. The output is further analysed by filtering the values through a high pass filter by detecting the positive slope encountered in the results. Then the corresponding frames are selected as the key frames or visual highlight of the video. In the workflow, there is a misclassification tolerance of 3 frames, which was experimentally determined. Also, these frames are used in the further analysis of feature and event detection. Since, scene filtering is achieved based on the algorithm of feature detection, the next section will present the feature extraction algorithm [2]. For the extraction of high-level visual features and events, a rectangular mesh structure is trained with both positive and negative samples from the pre-defined training models. The feature detector is a binary classifier, assigning labels to the input feature vectors. The network structure is presented in **Figure 2**, where $X$ is the input feature vector. The

training of the network neurons is performed using particle swarm optimization. The input feature vector from the training model is presented to the network. The winner node based on the competitive learning is selected. The features from the selected winner node and the input training feature are presented to PSO. The $d-$ dimension optimization problem to be solved by PSO is the $L1$ metric between the winner node feature vector to the input feature vector. The particle swarm for each dimension of the input feature is initialized randomly. The evaluation function for each particle in each dimension is calculated and accordingly the $pbest$ and $gbest$ values for the particle swarm is updated. The velocity and position of each particle in each dimension is updated. The iteration is continued until the result of the evaluation function is less than threshold $e_{th}$. The training of the algorithm is continued until all the input patterns from the training models are exhausted.

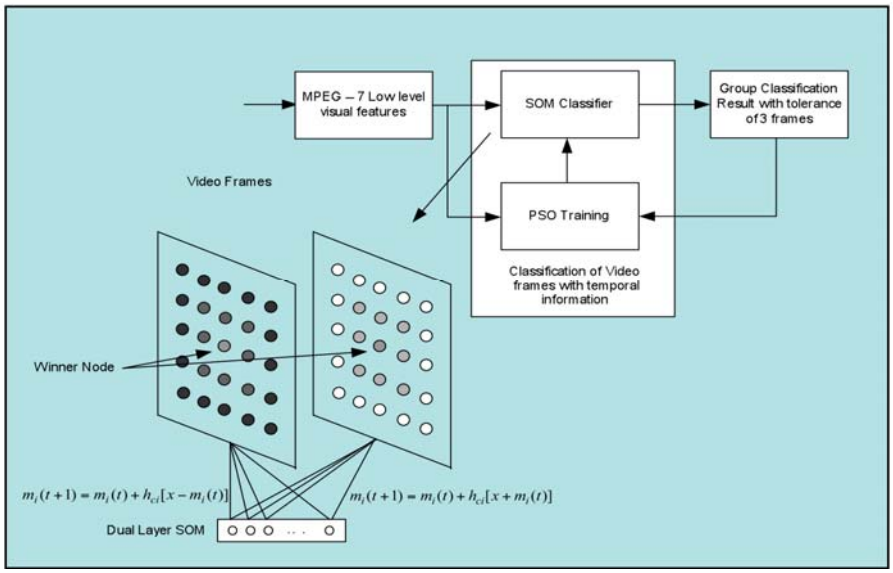


**Figure 2 - Dual Layer SOM and PSO based Highlight detection**

### *Classification model*

One of the key challenges in developing an automatic classification model is the presence of "Semantic Gap", which is succinctly defined as the gap between low-level features and high-level semantic features. Addressing this problem, a large number of indexing and retrieval algorithms have been presented in the literature. Although the performance of the machine learning techniques has been largely improved, the machine learning outcomes are still a far away from the results generated by human cognition. In tackling the problems of enhancing the performance of machine learning algorithms, recent developments in optimisation techniques have been inspired by problem solving abilities of biological organisms such as bird flocking and fish schooling. One such technique developed by Eberhart and Kennedy is called "Particle Swarm Optimisation (PSO)". In comparison to other

evolutionary computation algorithms, the PSO algorithm considers the following two main assertions as listed below [3]:

- Mind is Social: Learning from experience and emulating the successful behaviours of others, people are able to adapt to complex environments through discovery of relatively optimal patterns of attitudes, beliefs and behaviours.
- Particle swarm are a useful computational intelligence methodology: Central to the concept of computational intelligence is system adaptation that enables or facilitates intelligent behaviour in complex and changing environments. Swarm intelligence comprises of three steps namely evaluate, compare and imitate. Each particle goes through these stages by performing simple mathematical operations in solving a more complex optimisation problem.

Following the advantages listed above for the use of PSO algorithm, a Self Organising Map (SOM) based visual classifier has been developed and integrated in the Papyrus system for semantic indexing of the visual medium. The neural network architecture is based on the nervous systems component and can be categorised as feedforward, feedback and competitive [4]. Feedforward networks transform a set of input signals into a set of output signals. The desired input-output transformation is usually determined by external, supervised adjustment of the system parameters. In feedback networks [7], the input information defines the initial activity state of the feedback system, and after state transitions the asymptotic final state is identified as the outcome of the consumption. In competitive learning networks, neighbouring cells in a neural network compete in their activities by means of mutual lateral interactions and develop adaptively into specific detectors of different signal patters.

In competitive neural networks, active neurons reinforce their neighbourhood within certain regions, while suppressing the activities of the other neurons [5]. This is called on-center/off-surround competition. The objective of SOM is to represent high-dimensional input patterns with prototype vectors that can be visualised in a usually two-dimensional lattice structure [6]. Each unit in the lattice is called a neuron, and adjacent neurons are connected to each other, which results in a clear topology of how the network fits itself to the input space. Input patterns are fully connected to all neurons via adaptable weights and during the training process, neighbouring input patterns are projected into the lattice, corresponding to the adjacent neurons. SOM enjoys the merit of input space density approximation and independence of the order of input patterns. A detailed discussion on the implementation of the classifier has been presented in [8]. In **Figure 3**, an overview of semantic concept co-existance is presented. A thorough evaluation of the video analysis components has been presented in [9].

***The audio component analysis***

The analysis of the audio component of video items achieves two main objectives: on one side, it complements the visual component analysis enabling, through a multimodal analysis, the generation of higher level, semantically relevant, metadata and, on the other side, it provides a semantic indexing of video items similar to the one provided for textual news items, in order to make them available in a uniform, coherent manner.

Multimodal analysis exploits the combination of visual and audio features extracted from the digital media and the interaction between different layers and data streams present in the same multimedia document to provide semantic categories extracted by the combination of multiple modalities. Audio features, in particular, provide the basic content structure by identifying video segments characterized by narration, interviews and noise or music. Segments are labelled by type (e.g. speech / non speech), gender (male / female) and speaker. Speech segments are further analysed through a speech recognition process to provide the topic being discussed, the main concepts expressed and the related Named Entities.

**Figure 3 – An example of classification models (with concepts, windfarm, sky, windmill and vegetation)**

The following paragraphs describe how each task has been achieved, starting from the speaker segmentation process that is the basis for content structuring and speech segments identification, followed by a description of the speech recognition process

that generates transcriptions which can finally be analysed by Natural Language Processing techniques. The last paragraph illustrates a novel method for analysing textual content based on Wikipedia as a linguistic resource that has been tailored for speech transcriptions to reduce the impact of speech recognition errors on the metadata generation process.

**Speaker segmentation**

Speaker segmentation, also known as speaker diarisation, refers to the process of automatically transcribing a given audio data source in terms of "who spoke when" [10] giving an insight on audio items structure by identifying segments with homogeneous audio features and by providing a descriptive label of their content (e.g. "speech", "male", "speaker A", "noise" ...).

A typical Speaker Diarisation system conceptually performs these tasks:

- **Audio Feature extraction**: features extracted from the audio stream are intended to suggest information about the speakers in order to enable the system to separate them optimally.
- **Speech activity detection**: an audio stream may consist of some acoustic activities like speech, noise, music, background conversation and silence. Non-speech regions should be detected and removed from the audio stream.
- **Speaker change detection**: inside every speech region, a speaker change (or speaker turn) detector is used to find points in the audio stream which are candidates for speaker change points.
- **Gender detection**: it allows, for the segments classified as speech, to detect if the speaker is a male, a female or a child.
- **Speaker clustering**: segmented regions, belonging to the same speaker, are grouped together. This does not entail whether such segments come from the same acoustic file or different ones.

These tasks can be performed by different algorithms applied in different order, mixed together and repeated iteratively.

By structuring the audio/video stream into speaker turns, speaker diarisation has already proven its usefulness for the indexation of broadcast news, and multimedia objects in general, making possible, for example, to track people across recordings. Speaker diarisation is also useful as a preliminary step in the task of automatic transcription.

The usual output of a speaker diarisation system is a list of time slices (usually represented by their start and end time, or by their start time and the duration) with a description of each slice (usually represented as a set of tag). This information, either alone or integrated with information extracted from other modalities, may contribute extensively to the overall semantic interpretation of multimedia data [11].

The Papyrus Speaker Diarisation Framework (PSDF) includes tools for audio format conversion, features extraction, speaker segmentation, speaker clustering, speech activity detection and gender detection. The PSDF provides three algorithms that implement a complete speaker diarisation system, in order to provide the most reliable results to the content structuring and multimodal analysis.

Speech activity detection is a central task in speaker diarisation and evaluation measures, like the Diarisation Error Rate (DER), are directly affected by the performance on this task. Speaker diarisation is often used for speaker tracking and speech activity detection allows a finer tracking by excluding audio regions where the speaker is not talking. Moreover speech activity detection also helps to avoid confusing homogeneous noise segments with a speaker. The PSDF includes an implementation of the most common techniques for speech activity detection. Among these, the most robust to noise and context have been selected and used for the speaker diarisation algorithms.

In the PSDF two gender classification tools are also available that aim to divide the segments into common groupings of gender in order to supply more side information about the speakers in the final output.

Timing information of audio segments, speaker labels, speech/non-speech tags, male/female tags are all metadata provided by the PSDF and can be further used combined with the analysis results of other modalities.

**Automatic speech recognition**

Automatic Speech Recognition (ASR) is the process of converting spoken words to text. ASR supports the conceptual querying of video content and the synchronization to any kind of complementary resource. The potential of ASR-based indexing has been demonstrated most successfully in the broadcast news domain [12]. In fact, despite several years of research in this field, ASR systems work reliably only under rather constrained conditions, where restrictive assumptions, described in table 1, can be made. States of art performance levels for Large Vocabulary task (i. e. Broadcast News speech) are between 10-20% Word Error Rate (WER) depending on the language, type of speech and audio quality. For other domains, like in the Papyrus case, values under 50% are difficult to obtain [13].

The ASR task of the Papyrus project is a continuous, spontaneous, large-vocabulary speech recognition task of different speakers, over different channels in a noisy environment. Therefore the task requires a system stable to different environment conditions, that doesn't need training on individual speaker's voice and stable to different speaker accents. The mean WER for Papyrus videos is 48,1% with a high variability from video to video due to different audio and speech conditions, ranging from read speech in a studio environment to spontaneous speech under noisy acoustic conditions. This guarantees, anyway, sufficient accuracy for a robust textual analysis [14].

| Factors | Best case | Worst case | Papyrus case |
|---|---|---|---|
| Vocabulary size | Small vocabulary | Large vocabulary | Large vocabulary |
| ASR type | Dictation | Continuous speech | Continuous speech |
| Speech type | Reading | Spontaneous | Both |
| Speaker accent | Perfect match with the acoustic model training set | Non-native speakers outside the acoustic model training set | Both |
| Channel characteristics | Microphone | Telephone | Microphone |
| Environment | No kind of noise | Noise, Music, Overlapping speech | Noise, Music, Overlapping speech |

**Table 1- Main factors affecting ASR performance**

In **Figure 4** the whole process of audio analysis is described, with some details on components affecting ASR.

An ASR engine requires, for a given language:

1. Acoustic model: describes the basic sounds units of the language (phonemes)
2. Vocabulary: describes possible pronunciations of all the words of the language
3. Language model: describes how the words are related to each other in the language

Results from different software (Sphinx3 from Carnegie Mellon University and Sonic from the Colorado University), using different parameters, were compared for the five videos (English version) for which the reference texts had been provided.

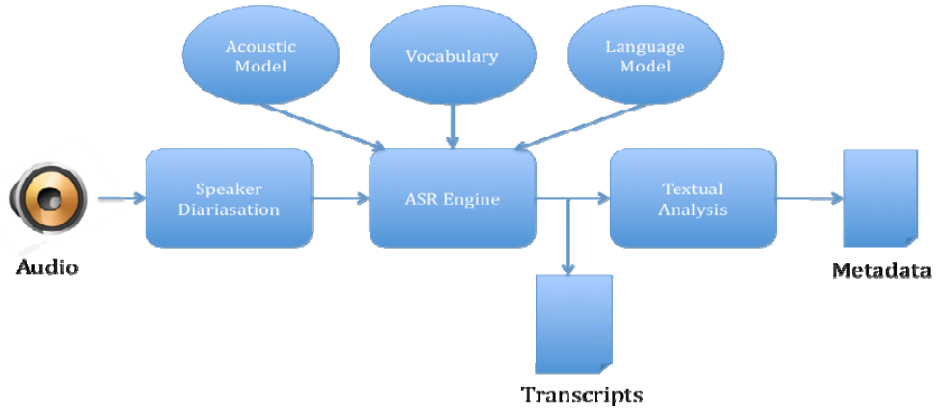


**Figure 4 – ASR Workflow**

The best results have been obtained using the acoustic model HUB4 distributed with the latest version of Sphinx4 for Java[5] and a language model that combines the lm_giga_64k language model and a language model specific for energy (Energy).

Pronunciations for words in the Energy language model but not in standard CMU dictionary (7a) have been added to the vocabulary.

The specific language model for energy has been generated starting form Deutsche Welle and Agence France Presse news on renewable energy and then combined with the lm_giga_64k language model[6]. The language model resulting from the combination accounts for 67000 words, 3000 of which specific to the energy domain.

This combination of resources, specifically tailored for the energy domain, improved recognition accuracy of 10% with respect to the standard resources.

**Concept Mapping**

While keyword extraction has been widely investigated in the text domain, there is less effort on speech transcripts [15]. Linguistic analysis of speech transcriptions is affected by a) the word recognition errors, b) the lack of punctuation and c) the lack of linguistic structure that characterizes the spontaneous speech. A method is therefore necessary to reduce errors and increase precision in the metadata generation process. The proposed method mainly relies on Wikipedia as a validation tool of the extracted linguistic constructs in terms of meaningfulness and relevance to the context.

The method is implemented in a specific tool, the Concept Mapper, which has been developed for analysing both textual news items and audio transcripts to identify the most relevant concepts and to connect them to the proper ontology identifiers. Textual news items and audio transcripts are treated differently in the concept selection stage. Links to the historiographical issues are achieved indirectly through the mapping between the History Ontology and the News Ontology. The Concept Mapper role is to map news content to the News Ontology, by leveraging textual and spoken language technologies and complementary resources such as Wikipedia.

Wikipedia provides both a linguistic resource and a source of additional metadata for semantic indexing. For each detected concept, in fact, the tool provides the following information:

- The keywords that were identified as representative of the concept and their frequency in the news item
- Additional information provided by the Wikipedia page that describes the concept:
  - Title
  - Translations to other languages
  - Anchors text and Redirects
  - Categories
- A score of relevance for the news item (internal relatedness)

- A score of relevance for the domain (external relatedness)
- The ontology identifier of the concept

The Concept Mapper is implemented in four steps:

1) candidate keywords extraction (noun phrases are selected using a shallow parsing procedure [16])

2) "anchor search" in Wikipedia content (exploiting all the available alternative ways of referring to the same concept, i.e. the anchors) and candidate keywords disambiguation [17] (whenever a noun phrase refers to more concepts, or Wikipedia pages, the one that best relates to the news item is chosen, using the internal relatedness measure [18]as a semantic proximity index)

3) keyword ranking and selection (internal and external relatedness are used to choose which of these concepts are relevant enough to the story and to the Papyrus domain to be retained as semantic metadata)

4) ontology connection (an ontology identifier is associated to each concept, when available)

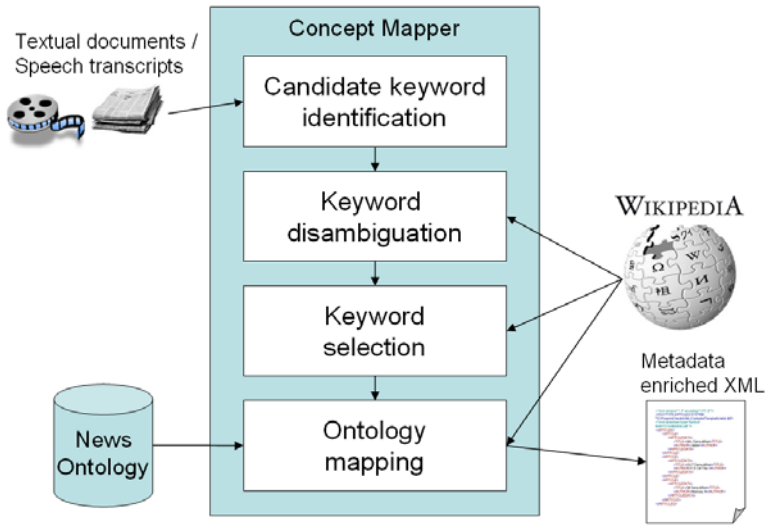


**Figure 5 - Architecture of the concept mapper**

The main issues in this process are the detection of erroneous nominal phrases across sentences (due to the lack of punctuation) on one side, and the loss of correct nominal phrases due to the speech recognition errors and to repetitions and stammering of the spontaneous speech, on the other side.

While the loss of information is not easily recoverable and will affect the system recall, the detection of erroneous chunks can be reduced by filtering the results with a Wikipedia validation process. This will improve the system precision, avoiding most of the speech recognition mistakes to affect the metadata generation.

The validation process is essentially a two step process. In the first step, only nominal phrases that are linkable to a Wikipedia page are kept, since this makes possible to

assign a meaning to the nominal phrase. Even if spurious nominal phrases are eliminated at this stage, it is still possible that irrelevant chunks are kept. The second step aims at identifying them by checking whether their meaning is pertinent to the context (both the internal context of the news item and the external context of the news domain). For this purpose the internal and external relatedness measures can be used, as well as the frequency of the nominal phrase. Spoken language is indeed more redundant than the written one and repetitions of terms within a shot prove their relevance even when the ASR texts include errors and lack of structure.

To identify the most appropriate criteria for nominal phrases selection, a manual annotation of the reference texts is necessary in order to define the "reference nominal phrases" that the system should be able to extract. The "reference nominal phrases" are those concepts that mostly reflect the news content and are agreed on by domain experts.

Once the list of "reference nominal phrases" is available, different selection criteria (based on the internal relatedness, external relatedness, frequency, commonness, confidence and any appropriate combination of them) can be compared in order to maximise precision and recall of the concepts retrieved from the transcripts.

To illustrate the process of criteria selection, results obtained analysing the longest available video item can be presented. The analysis of the reference text of "*332134 6 2007 made in germany schottland ausbau windenergie english*" led to the (manual) identification of 39 relevant concepts, among the 103 nominal phrases that were actually identified and had a corresponding page in Wikipedia. These can be considered representative of the news content. Among them are: renewable energy, Scotland, Scottish Power, rural, hill, sight, tourism, turbine, wilderness, wildlife, wind farm and wind power.

The textual analysis of the ASR transcription led to the automatic identification of 123 nominal phrases (after Wikipedia 1st step validation). Since only 20 of them correspond to the "reference nominal phrases", this implies that the concept recall is 51,3% (20 correctly identified concepts over 39 reference concepts) and cannot be improved. On the other hand, the precision value of 16,3% (20 correctly identified concepts over 123 identified concepts) can be improved by identifying the criteria to select most of the correct concepts out of the automatically retrieved ones. In order not to affect the system recall too seriously, the F score, instead of the precision, will be maximized.

Figure 6 shows the F score trend, for each selection criteria, as the selection threshold decreases and, consequently, the number of selected nominal phrases increases. The graph shows how the F score is maximised by all the criteria by selecting a number of nominal phrases around 20. Furthermore, the graph shows that, the external relatedness of the chunk, multiplied by the number of its occurrences, achieves the highest F score at almost any level of the threshold (and number of selected chunks). In particular, the maximum is reached at the threshold 0,45 of the "frequency *

external relatedness" measure, selecting 21 chunks: as 15 of them match with the "reference nominal phrases", the recall is 38,5% and the precision 71,4%.

The final criteria that has been implemented is therefore to select from ASR transcriptions only nominal phrases with external relatedness (multiplied by the frequency) above 0.45. The better performance of the external relatedness with respect to the internal one is anyway justified by the presence of speech recognition errors that affect the news item internal context, favouring the external (domain) context as more reliable.

From a qualitative point of view, the retrieved nominal phrases cover all the main topic of the news item and, with respect to precision, it should be noticed that most of the "erroneously" identified concepts are actually concepts correctly identified and correctly disambiguated that don't satisfy the chosen relevance requirement (euro, Europe, meter, engineering, people, pipeline transport). Thus discrepancies between the manually assigned relevance and the system generated relatedness account for most precision errors (relevance overestimation) and for a large fraction of recall errors (relevance underestimation). With respect to this, it should be noticed that concept relevance is highly subjective and that the observed disagreement on the degree of relevance falls within the inter-rater assessed disagreement (Cohen's kappa coefficient of 0.45).

Concerning the geographical locations, in particular, since their identification is one of the multimodal analysis objectives, the method described enables the retrieval of generic locations (e.g. hills, shore, sea, mountains, countryside and etc.) as well as nations and main regions and towns, but mostly fails on small towns as ASR vocabularies don't provide the pronunciation for them (nor Wikipedia provides a page for them). The selected concepts, together with the additional Wikipedia related information, provide semantic metadata to the news item that enable a semantic search. Furthermore, the Concept Mapper identifies, for each selected concept, the proper News Ontology link by exploiting the ontology structure and maximising the group relatedness, i.e. the semantic proximity between the concept and the group of concepts (synonym set) as defined in the ontology [19].

The reason for using a Wikipedia annotation as intermediate step to obtain an ontology annotation is due to the fact that the Papyrus News Ontology is a domain restricted ontology (does not comprehend "tourism" for example, although it can be a quite meaningful metadata) and that it doesn't provide confidence measures for filtering the textual analysis results. In the "*CLS 332134 6 2007 made in germany schottland ausbau windenergie english*" video, for example, the spoken word "paradise" is erroneously recognised as "Paris" which would be validated by the ontology, whereas filtering it with the internal relatedness eliminates it (the video is actually set in Scotland).

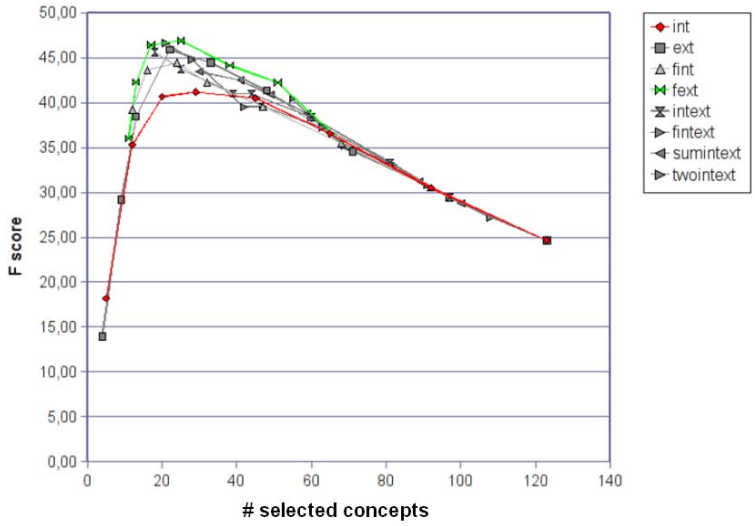


**Figure 6 – Comparison of selection criteria**

Therefore, the proposed method automatically identifies relevant concepts in textual documents and automatically maps them to their formalization in a given domain ontology. This enables automatic annotation of texts and semantic metadata generation exploiting both Wikipedia knowledge and the Ontology knowledge. This method has already been implemented in the Papyrus (*Cultural and Historical Digital Libraries Dynamically Mined from News Archives*) prototype to provide metadata generation for the semantic search functionality and to provide content mapping to the News Ontology for the cross discipline search functionality. It analyses both textual content and speech transcripts in English and French, in two domains (renewable energy and biotechnology) and can easily be extended to other languages and domains. The temporal information that is provided along with the metadata enables the semantic indexing and also the synchronization of video and audio segmentations. This allows improving metadata quality as well as scene segmentation, by simultaneously taking into account the information provided through different modalities and is part of the ongoing research activity.

### *Papyrus Interdisciplinary Search Engine*

The audiovisual framework presented in this paper has been integrated into an online Papyrus interdisciplinary search engine.

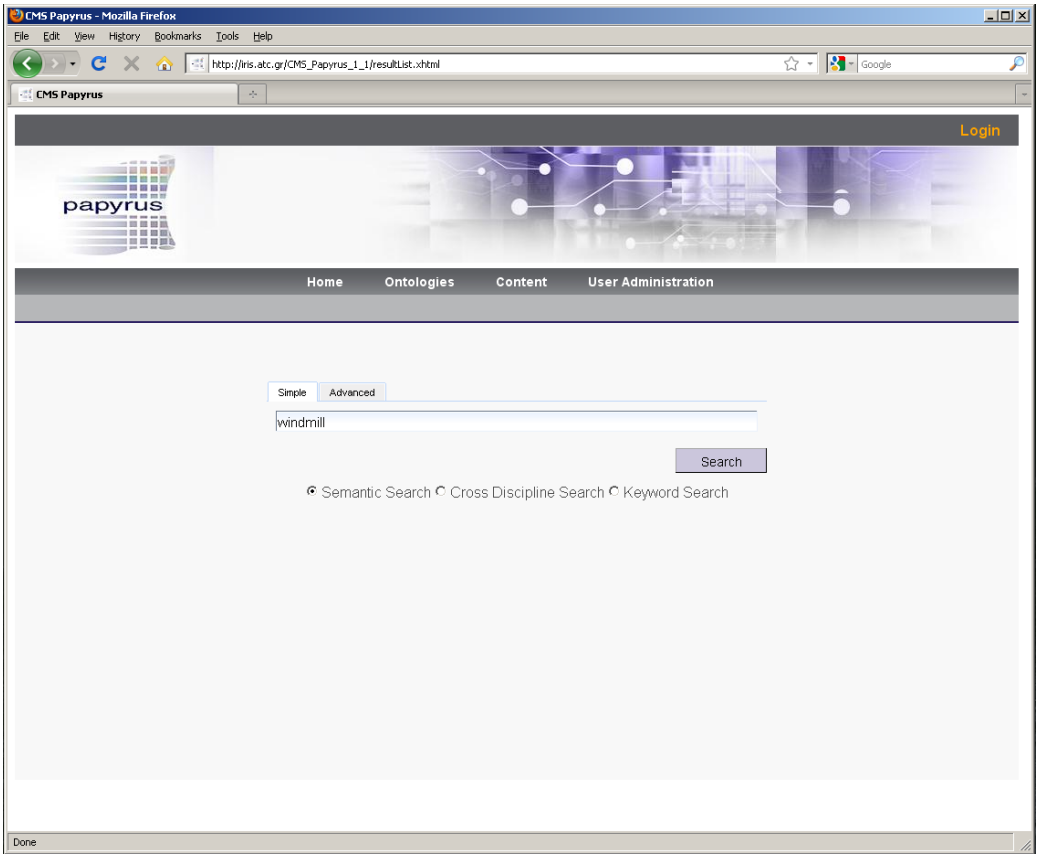


**Figure 7 – An overview of the Papyrus search engine interface**

### *Conclusion and future work*

In conclusion, the integrated framework presented in this paper provides an easy and flexible access to the previously unexplored audiovisual items for research in issues of historiographical importance. The integrated framework is a part of the Papyrus interdisciplinary search engine, which can be accessed through the online portal.[7] In future we will focus on evaluating the performance of the integrated system along with other usability issues.

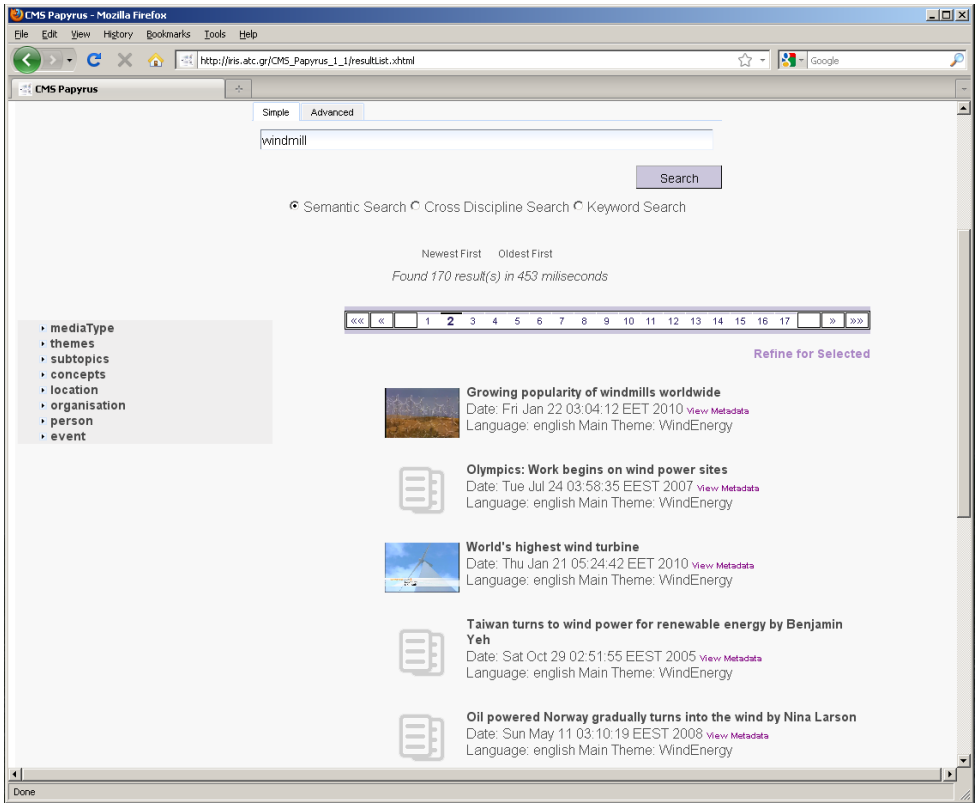


**Figure 8 – The screenshot of the Papyrus results page for the query "windmill"**

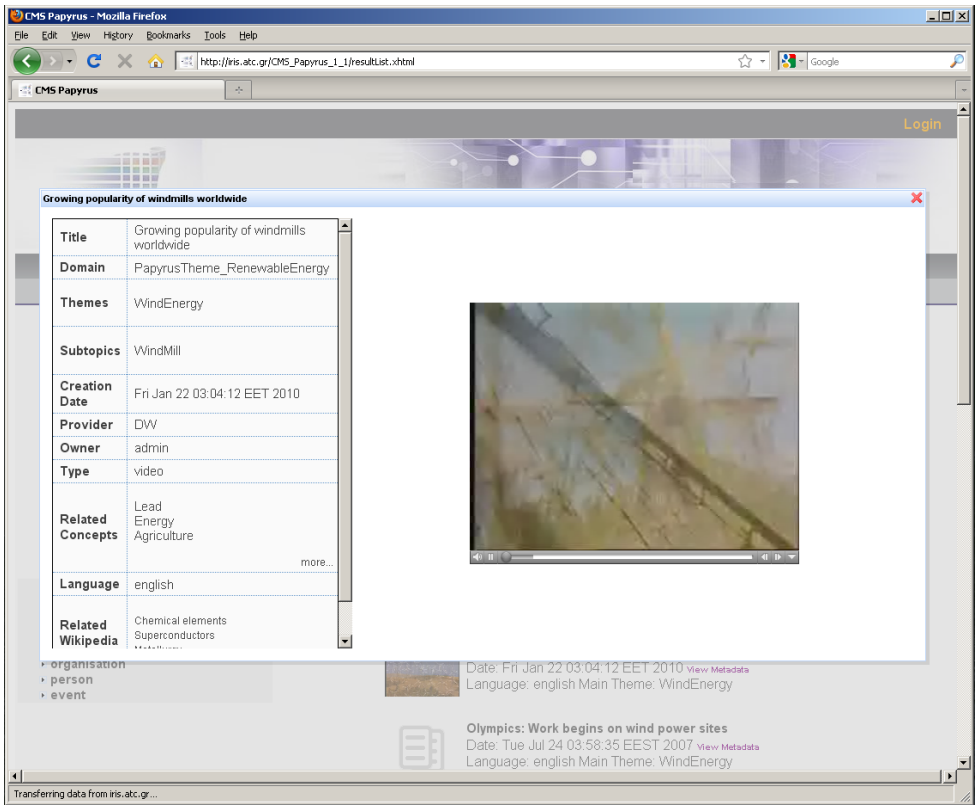


**Figure 9 – A screenshot of the audiovisual metadata**

## Notes

[1] For a review of such suggestions, see Aristotle Tympas, "Methods in the History of Technology", in Colin Hempstead (ed.), *Encyclopedia of 20th Century Technology*, New York: Routledge, (2005), pp. 485-489. For suggestions concerning the history of technology in Europe, see the articles in a special issue of *History and Technology* 21, no. 1 (2005).
[2] For a sample of books on the history of wind power, see T. Lindsay Baker, *A Field Guide to American Windmills*, Norman: University of Oklahoma Press, 1985, Richard Hills, *Power from* Wind*: A History of Windmill Technology*, New York: Cambridge University Press, 1994, Matthias Heymann, *Die Geschichte der Windenergienutzung 1890-1990,* Frankfurt: Campus-Verlag, 1995*,* Robert Righter, *Wind Energy in America: A History*, Norman: University of Oklahoma Press, 1996. For insightful historiographical suggestions, see also Matthias Heymann, *"*Signs of Hubris: The Shaping of Wind Technology Styles in Germany, Denmark, and the United States, 1940-1990"*, Technology and Culture* 39, no. 4 (1998): 641-670 and Geert Verbong, "Wind Power in the Netherlands 1970-1995", *Centaurus* 41, no. 1-2, (1999): 137-160.
[3] See the references in [Tzokas et al., this session], footnotes 7-10.
[4] A brief discussion on the theoretical motivation for the use of SOM is presented the next section.
[5] Other acoustic models (old version of HUB4, Voxforge and WSJ) showed degrading performance.
[6] http://www.inference.phy.cam.ac.uk/kv227/lm_giga/
[7] http://iris.atc.gr/CMS_Papyrus_1_1/

## References

[1] Manjunath, B. S., P. Salenbier and T. Sikora, *Introduction to MPEG – 7, Multimedia content description interface*, New York: Wiley, 2003.

[2] Chandramouli, K. And E. Izquierdo, "Visual Highlight Detection using Particle Swarm Optimisation", *Latin-American Conference on Networked and Electronic Media*, 2009.

[3] Kennedy, J. and R. C. Eberhart, *Swarm Intelligence*, San Francisco, CA: Morgan Kaufmann, 2001.

[4] Rumelhart, D. E., G. E. Hinton, R. J. Williams, "Learning internal representations by error propagation", In D. E. Rumelhart and J. L. McClelland (eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, (vol. 1, pp. 318-362.), Cambridge, MA: MIT Press, 1986.

[5] Hopfield, J. J., "Neural networks and physical systems with emergent collective computational abilities", *Proceedings of the National* Academy *of Sciences*, 79:2554-2558, 1982.

[6] Inoue, M., "Image Retrieval: Research and use in the information explosion", *Progress in Informations* 6 (2009): 3-14.

[7] Kohonen, T. "The Self Organising Map", *Proceedings of IEEE* 78, no. 4 (1990): 1464-1480

[8] Chandramouli, K. and E. Izquierdo, "Image Retrieval using Particle Swarm Optimisation", in M. C. Angelides, P. Mylonas and M. Wallace (eds.), *Advances in Semantic Media Adaptation and Personlisation,* (pp. 297-319)*, CRC Press*, 2009.

[9] Chandramouli, K., et al., "Techniques for Multimodal content analysis", *Technical Report*, 2009.

[10] Tranter, Sue E. and Douglas A. Reynolds, "An overview of automatic speaker diarization systems", *IEEE Transactions* on *Audio*, *Speech*, and *Language* Processing 14, no. 5 (2006): 1557-1565.

[11] Friedland, G., H. Hung and C. Yeo, "Multi-modal Speaker Diarization of Real-World Meetings Using Compressed-Domain Video Features", *Tech.Rep. 08-007*, ICSI, October, 2008.

[12] Huijbregts, M.A.H. and Ordelman, R.J.F. and de Jong, F.M.G., "Annotation of Heterogeneous Multimedia Content Using Automatic Speech Recognition). In *Proceedings of the Second*

*International Conference on Semantic and Digital Media Technologies*, SAMT 2007, 5-7 Dec 2007, Genoa, Italy, 2007.

[13] Rehatschek, H. and Sorschag, R. and Rettenbacher, B. and Zeiner, H. and Nioche, J. and de Jong, F.M.G. and Ordelman, R.J.F. and van Leeuwen, D., "Mediacampaign: A Multimodal Semantic Analysis System for Advertisement Campaign Detection". In: *Proceedings of international workshop on Content-Based Multimedia Indexing*, CBMI 2008., 18-20 June 2008, pp. 85-92, London, UK.

[14] Garofolo, J.S., C.G.P. Auzanne, and E.M. Voorhees, "The TRECSDR Track: A Success Story", In *Eighth Text Retrieval Conference*, pp. 107-129, Washington, 2000.

[15] Liu, F., D. Pennell, F. Liu, and Y. Liu, "Unsupervised approaches for automatic keyword extraction using meeting transcripts", in *NAACL '09: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Morristown, NJ, USA, 2009, pp. 620–628, Association for Computational Linguistics.

[16] Schmid, Helmut, "Probabilistic part-of-speech tagging using decision trees", In *Proceedings of the International Conference on New Methods in Language Processing*, pp. 44–49, 1994. (available at http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger1.pdf)

[17] Milne, D. and I. Witten, "Learning to link with Wikipedia" in *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, New York, NY, USA, 2008, pp. 509–518, ACM.

[18] D. Milne and I.Witten, An effective, low-cost measure of semantic relatedness obtained fromWikipedia links, 2009.

[19] Reiter, Nils, Matthias Hartung, and Anette Frank, "A Resource-Poor Approach for Linking Ontology Classes to Wikipedia Articles", in Johan Bos and Rodolfo Delmonte (eds.), *Semantics in Text Processing. STEP 2008 Conference Proceedings*, vol. 1 of Research in Computational Semantics, pp. 381–387, College Publications, 2008. (available at http://www.aclweb.org/anthology/W08-2231)