REVEAL THIS: Public Final Activity Report

Project ref.	FP6-IST-511689
Project Acronym	REVEAL THIS
Project full title	Retrieval of Video and Language for The Home user in an Information Society
Distribution	Restricted

Public Final Activity Report November 2004 – April 2007

Project Ref. no.:	FP6-IST-511689
Acronym:	REVEAL THIS
Full Title:	Retrieval of Video and Language for The Home user in an Information Society
Reporting period:	November 2004 – April 2007
Project Partner:	Name: Stelios Piperidis Institution/Company: Institute for Language and Speech Processing Address: Artemidos 6 & Epidavrou, 151 25 Marousi, Athens , Greece Phone: +30-210-6875 421 Fax : +30-210-6852 620 E-mail: spip@ilsp.gr Public project web site: http://www.reveal-this.org

REVEAL THIS Consortium	
Partner Short Name	**Partner Full Name**
ILSP	Institute for Language and Speech Processing
SAIL	SAIL LABS Technology AG
XRCE	XEROX The Document Company
KUL	Katholieke Universiteit Leuven
USG	University of Strathclyde
BeTv	BeTv
TVEyes	TVEyes UK Ltd

Table of Contents

Executive Summary

REVEAL THIS
Retrieval of Video and Language for The Home user in an Information Society

URL : http://www.reveal-this.org

REVEAL THIS addresses a basic need underlying content organisation, filtering, delivery and consumption by developing systems that will help European citizens and content providers keep up with the explosion of digital content that is scattered over different platforms, like TV, radio, World Wide Web, different media, such as speech, text, image and video, as well as different languages.

REVEAL THIS develops content processing technology able to capture, semantically index, categorise and cross-link multimedia and multilingual digital content. Users of the system will satisfy their information needs through personalized semantic search and retrieval, summaries of content and translation of them into their desired language.

To achieve its objective, the REVEAL THIS project set the following measurable innovative technological & scientific objectives:

- enrichment of multilingual multimedia content with semantic information like topic of a multimedia document, speakers, actors, facts and events mentioned, as well as keyframes relevant to user profiles
- semantically linking multimedia information presented in different media and languages
- development of cross-media categorization and summarization engines
- deployment of cross-language information retrieval and machine translation technologies in order to allow the user to search for and retrieve information according to his language preferences

These objectives aim at providing a technology suite that will enable the development of a fully operational personalized entertainment system. **Such a system can be used** (a) **by content providers**, to add value to their content, restructure and re-purpose it and offer their subscribers, individual or corporate users, personalized content, or (b) **directly by end users**, for gathering, filtering and categorizing information collected from a wide variety of sources in accordance with user preferences.

The project was implemented by: Institute for Speech and Language Processing / IRIS (Greece)- Coordinator, SAIL LABS Technology AG (Austria), Xerox - The Document Company S.A.S (France), Katholieke Universiteit Leuven R&D (Belgium), University of Strathclyde (United Kingdom), BeTV SA (Belgium), TVEyes UK Ltd (United Kingdom)

1 REVEAL THIS in the period November 2004 –April 2007

1.1 Summary of Activities

During its thirty-month duration, REVEAL THIS completed the development and evaluation of all its subsystems and of the final integrated prototype; in particular, the Cross-media Content Analysis & Indexing Subsystem, the Cross-media Categorisation, the Cross-media Summarisation, and the Cross-lingual Translation Subsystem, were designed, developed, evaluated and integrated in a prototype that implements two different information access scenarios : a ***pull scenario*** by which a user ***retrieves*** multimedia information through a web based user interface, and a ***push scenario*** by which a user receives information in his mobile, ***filtered*** according to his profile (as declared in the system). User requirements analysis and market and technology watch guided the whole software design lifecycle.

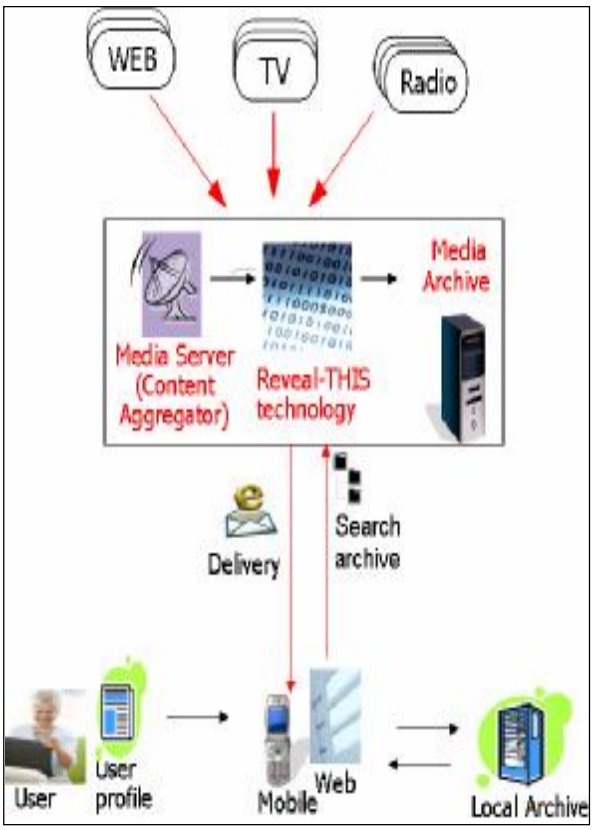

Figure 1: REVEAL THIS use scenarios

The information the user can access consists of European politics issues, sourced from European Parliament related data, associated national news TV & radio programmes, web text data, and travel information, sourced from travel documentaries, national TV programmes and web texts. Information is sourced in English and Greek. Usability testing of the integrated prototype was undertaken for both project scenarios with both English and Greek-speaking users.

Cross-media Content Analysis and Indexing Subsystem

In the final version of the main REVEAL THIS subsystem, the Cross-media Content Analysis and Indexing Subsystem (CCAIS) has been released catering for processing single media and automatically generating metadata such as:

- speaker turn and identity, transcriptions and stories for speech data,
- named entities (persons, places, organizations), topics and facts for web text and transcribed speech
- keyframes, shots, faces (detected and identified) and image categories for video and images

The metadata produced by the single-media processing components are aligned, synchronized and encoded in XML. The result is an XML/MPEG-7 document

containing all information gathered and linked to the corresponding points of the source material (text, audio, video). Segmentation suggested by audio processing (speaker turns) and topic detection is taken into account to produce unified segments. Several consecutive segments are aggregated to form "stories", i.e. sections of the document that deal with the same topics. The cross-media indexing component decides on the most appropriate indexing terms per story.

Crossing media (*or cross-mediality*) in REVEAL THIS is conceived of as the process of intelinking evidence, in the form of indexical data, provided by the different media participating in the message formation process within the same multimedia document (e.g. a video file).

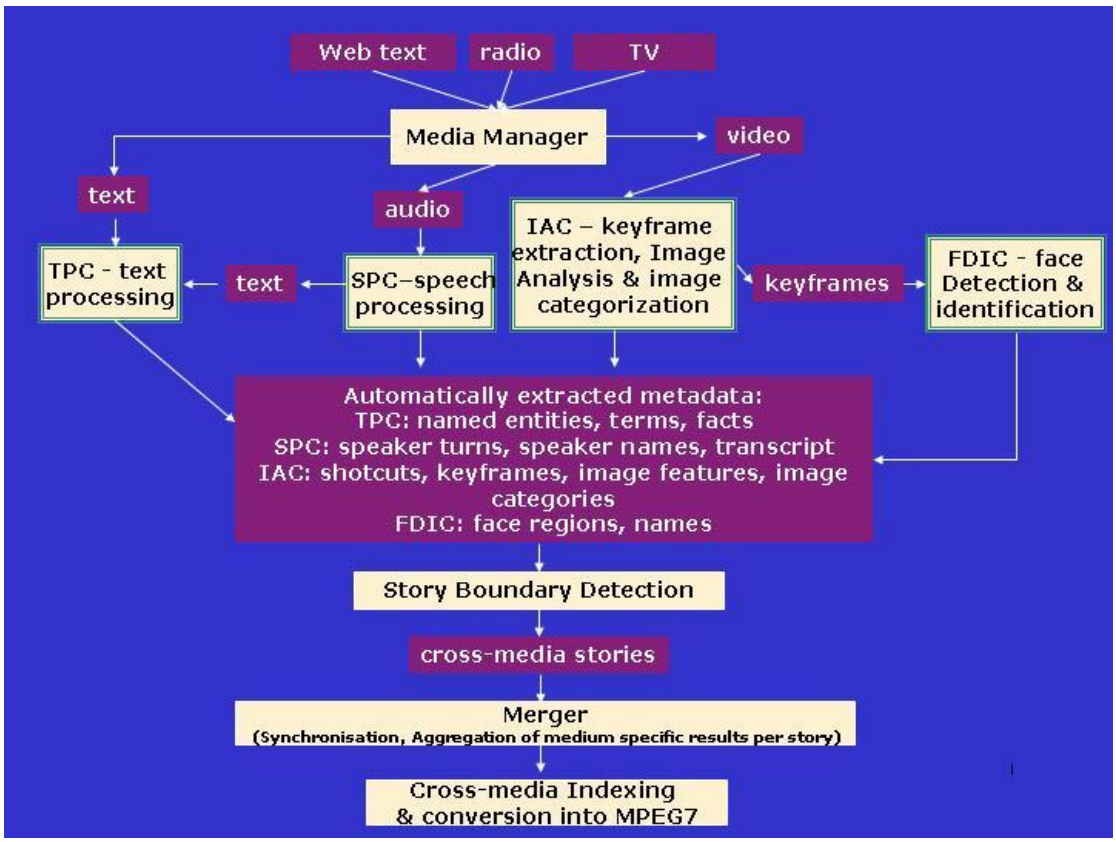

Figure 2: REVEAL THIS Cross-media Content Analysis and Indexing Subsystem

Deploying this indexical information and representations, REVEAL THIS uses state-of-the-art technologies for categorizing, summarizing and translating multimedia documents.

Cross-media categorisation

The categorization subsystem operates along the cross-lingual and cross-media directions. In the cross-lingual dimension, a categorizer based on a pivot language category model (in English) is deployed. Documents in other languages go through a translation phase, thus enabling their categorisation. Such a strategy overcomes constraints at the level of the training set, which often are not sufficiently big for building models in all languages involved.

In the cross-media dimension, documents containing not only text or images but a combination of different types of media (text, image, speech, video) are considered. The multiple-view fusion method adopted builds 'on top' of two single-media categorizers, a textual and an image categorizer, without the need to re-train them. Data annotated manually for both textual and image categories is used for training the cross-media categorizer. In that set dependencies between single-media category systems are exploited in order to refine the categorization decisions made.

Cross-media summarisation Subsystem

The task of the Cross-media Summarization Subsystem (CSS) is to determine and present the most salient parts according to users' profiles and interests by fusing video, audio and textual metadata. The CSS subsystem consists of three major components: the textual-based summarization component, the visual-based summarization component, and the cross-media summarization component aiming at the fusion of the two analyses and creating a self-contained object. Additionally, the Cross-media summarization subsystem provides the necessary visualization interfaces enabling the user to preview a specific multimedia object before downloading a file of interest. Cross-media summaries for politics, news and travel-related audiovisual files have been designed and implemented in the project; the module developed interfaces also with a web-based translation service for creating "translated" versions of the summaries for English and Greek (both directions).

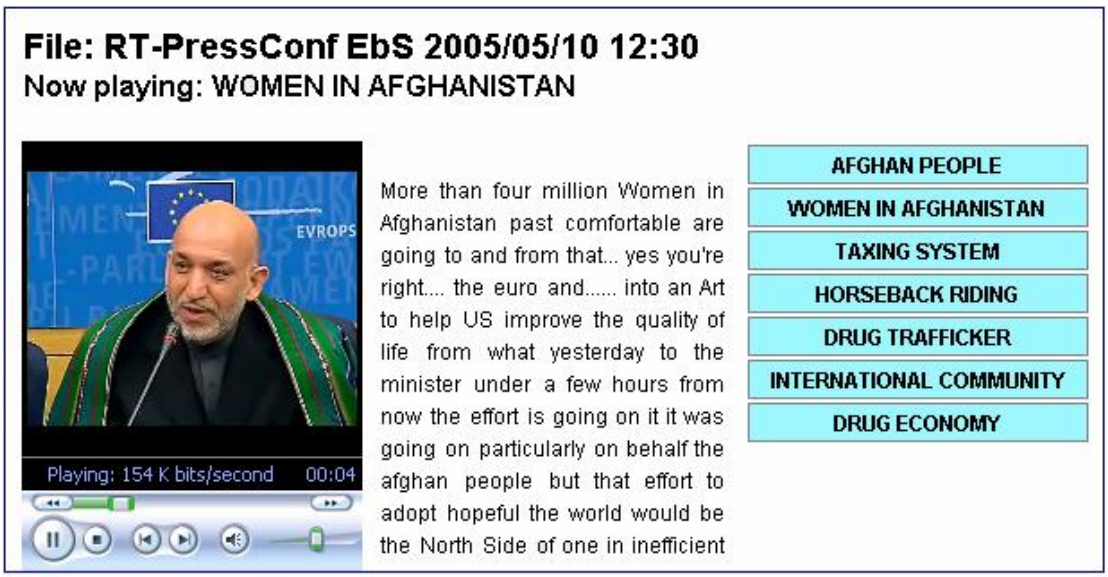

Figure 3: Multimedia summary (file summary)

Cross-lingual Translation

The Cross-lingual Translation subsystem (CLTS) allows users to query documents written in different languages, to categorise content expressed in different languages and to preview language specific summaries. A bilingual lexicon extraction module is used to generate lexical equivalences used for query translation purposes, but also to

replace keywords in a target language, in case a document is linguistically not well formed (e.g. output from a speech recognizer) and thus not effectively translated. Last, a statistical machine translation module is responsible for providing translations of the textual part of the summaries produced by the Cross-Media Summarization Subsystem. The translation pair is English and Greek and translation takes place in both directions.

REVEAL THIS Integrated Application Prototype

The final system consists of two major components: The *Multimedia Indexer* (based on the cross-media indexing component – CMIC) and the *Media Server*. All results of the Indexer together with the original media files are uploaded to the Media Server. Media files are uploaded in two formats: in Window Media format (.wmv/.wma) and in 3gp format (targeted at mobile devices). The server supports *searching for content* (pull scenario) or *filtering content* (push scenario) based on the metadata. It provides multi-lingual search, multi-media presentation of retrieved content, personalization, summarisation, and delivery to different devices and in particular to PCs/Laptops and mobile phones. The web-interface of the prototype provides full search and retrieval functionalities targeting professional users with demanding information needs, while the mobile-interface provides a light-weight access to the different functionalities of the prototype, suitable for mobiles and for laymen.

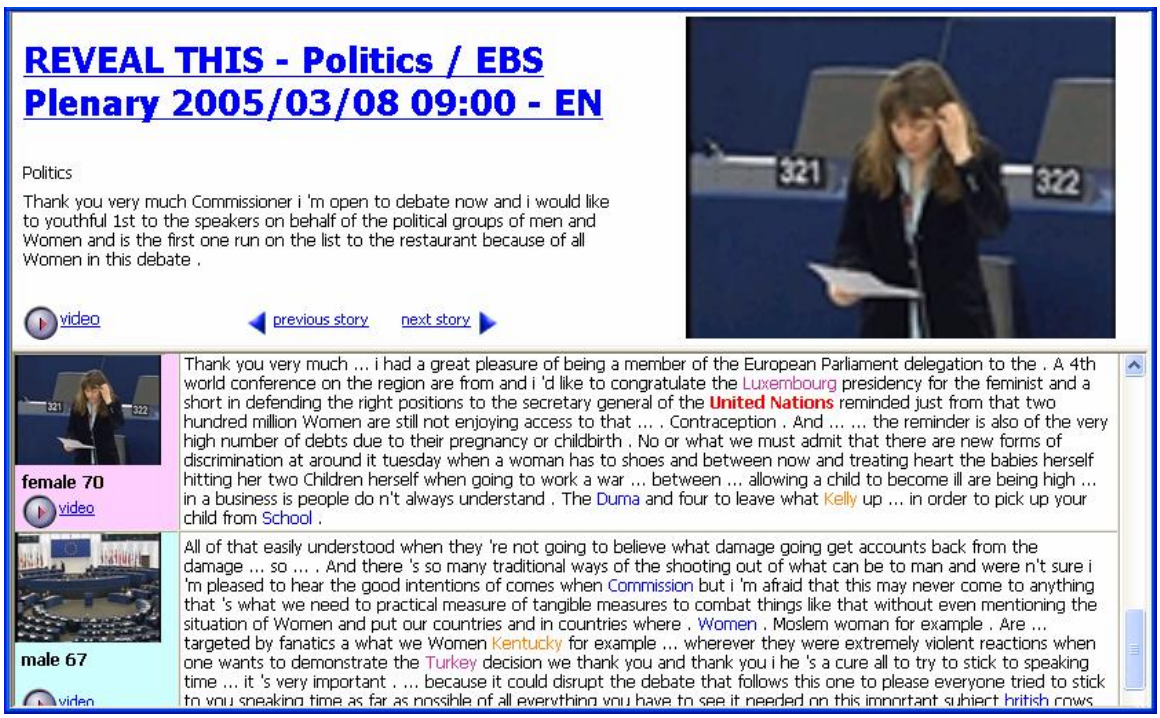

Figure 4: Story View in the REVEAL THIS web-interface

Figure 5: Screenshots of the REVEAL THIS mobile-application interface

In the last project stage, usability evaluation of the REVEAL THIS prototype took place. The system was tested in the two domains of EU politics and travel information, and in two languages, English and Greek. Advanced users, experienced in multimedia search, used the prototype in its pull (retrieval) mode (the web interface), while novice users, with limited or no search experience used it in its push (filtering) mode through mobile phones. The results showed that the innovative REVEAL THIS functionalities were more than welcome by both user groups, the performance was generally satisfactory for them, though expectations for much better performance in terms of e.g. speech transcription and translation were evident.

REVEAL THIS provides a technology suite enabling the development of a fully operational personalized entertainment system. Such a system can be used by content providers, to add value to their content, and directly by end users, for gathering, filtering and categorizing multimedia information.

Contact person
Stelios Piperidis
Department of Language Technology Applications
Institute for Language and Speech Processing / IRIS
Artemidos 6 & Epidavrou, 151 25 Marousi, Athens, Greece
E-mail: spip@ilsp.gr

2 Project objectives and major achievements

2.1 Project Objectives

The main objective of the REVEAL THIS project is the design, development and testing of an integrated infrastructure that allows the user to store, categorize and retrieve multimedia and multi-lingual digital content across different sources (TV, radio, music, Web), with a view to personalize the user experience with these sources.

To achieve the main objective, the REVEAL THIS project has set the following measurable innovative technological & scientific objectives:

- Augment the content of multimedia documents with semantic information like: characteristic keyframes, identified speakers, faces, entities, topics and fact information.
- Develop cross-media and cross-language representations.
- Build high-level functionalities, namely categorization, and summarization, based on cross-media and cross-language representations.
- Provide cross-lingual capabilities (cross-lingual information retrieval, categorization and machine translation of indicative summaries) based on the latest statistical machine translation technology.
- Integrate the above technologies in a unified platform equipped with semantic search, retrieval engine, personalization and delivery mechanisms to at least two different devices (mobiles and PCs).

These objectives aim at providing a technology suite that will enable the development of a fully operational personalized entertainment system. **Such a system can be used** (a) **by content providers**, to add value to their content, restructure and re-purpose it and offer personalized content to their subscribers (individual or corporate users), or (b) **directly by end users**, for gathering, filtering and categorizing information collected from a wide variety of sources in accordance with user preferences.

In order to achieve the first scientific objective and effectively render the above functionalities possible, it is necessary to provide additional indices that pertain to:

- **text:** named entities (e.g. names of persons, places, organizations), terms, topics and facts
- **speech:** speakers (e.g. speaker identity), transcriptions and stories
- **video and images:** persons, faces, keyframes and image categories.

REVEAL THIS deploys state-of-the-art technologies and components, namely indexing techniques for each media:

- Speech processing component – automatic speech recognition, automatic speaker identification.
- Image analysis component – shots, keyframes, image categories.
- Face analysis component – face recognition, face identification.

- Text processing component – entity, term and fact extraction, topic detection, story segmentation.

The metadata/indices produced by the above components are aligned, synchronized and encoded in XML. Once documents* have been enriched and indexed, representations suitable for crossing media and languages in the processes of retrieval, categorization and summarization are developed.

Crossing media (*or cross-mediality*) in REVEAL THIS is conceived of as the process of intelinking evidence, in the form of indexical data, provided by the different media participating in the message formation process within the same multimedia document (e.g. a video file).

To achieve this second scientific objective, REVEAL THIS exploits the links between different media explicated in each document. A promising approach lies in exploiting similarities between media in each document and deriving semantic representations that are "naturally" linked across media supported by evidence provided by the different processing components.

Based on the above indexical information and representations, REVEAL THIS deploys state-of-the-art technologies for categorizing and summarizing documents.

However, the diversity of the documents reaching the citizens of an integrated, interconnected Europe does not only lie on the different media used, but also on the different languages through which content is mediated. REVEAL THIS develops a cross-lingual translation subsystem equipped with state-of-the-art components capable of automatically extracting multilingual glossaries from specific collections via word/term alignment techniques, so as to complement existing multilingual resources in the provision of domain-tuned retrieval and categorization technologies. Furthermore, REVEAL THIS has the capability of presenting summaries of documents in different languages, by using state-of-the-art statistical machine translation models.

Finally, REVEAL THIS integrates the above technologies in a unified infrastructure equipped with search and retrieval, personalization and delivery mechanisms to two different devices (mobiles and PCs). In this context, state-of-the-art techniques of distributed information retrieval, related to multimedia resource selection, data fusion and presentation of results, are coupled with document summarization and categorization to enable the user to effectively search and browse the large amount of multimedia content gathered.

2.2 REVEAL THIS work performed in the period November 2004-April 2007 and main achievements

All the objectives of the project Description of Work (DoW) have been met and a number of contributions beyond the state of the art have been achieved.

* When we use the term document, we implicitly refer to a multimedia document, written in, spoken in and/or associated with one or more languages.

A summary of the work undertaken is presented below. Work results have been documented in the respective deliverables, a tabular presentation of which is appended at the end of this section. More detailed accounts of the work performed are presented in the next section "Progress per work-package", while a table listing contributions beyond the state of the art, per work-package follows.

Cross-media Content Analysis and Indexing Subsystem

In the final version of the main REVEAL THIS subsystem, the Cross-media Content Analysis and Indexing Subsystem (CCAIS) has been released catering for processing single media and automatically generating metadata such as:

- speaker turn and identity, transcriptions and stories for speech data,
- named entities (persons, places, organizations), topics and facts for web text and transcribed speech
- keyframes, shots, faces (detected and identified) and image categories for video and images

The metadata produced by the single-media processing components are aligned, synchronized and encoded in XML. The result is an XML/MPEG-7 document containing all information gathered and linked to the corresponding points of the source material (text, audio, video). Segmentation suggested by audio processing (speaker turns) and topic detection is taken into account to produce unified segments. Several consecutive segments are aggregated to form "stories", i.e. sections of the document that deal with the same topics. The cross-media indexing component decides on the most appropriate indexing terms per story.

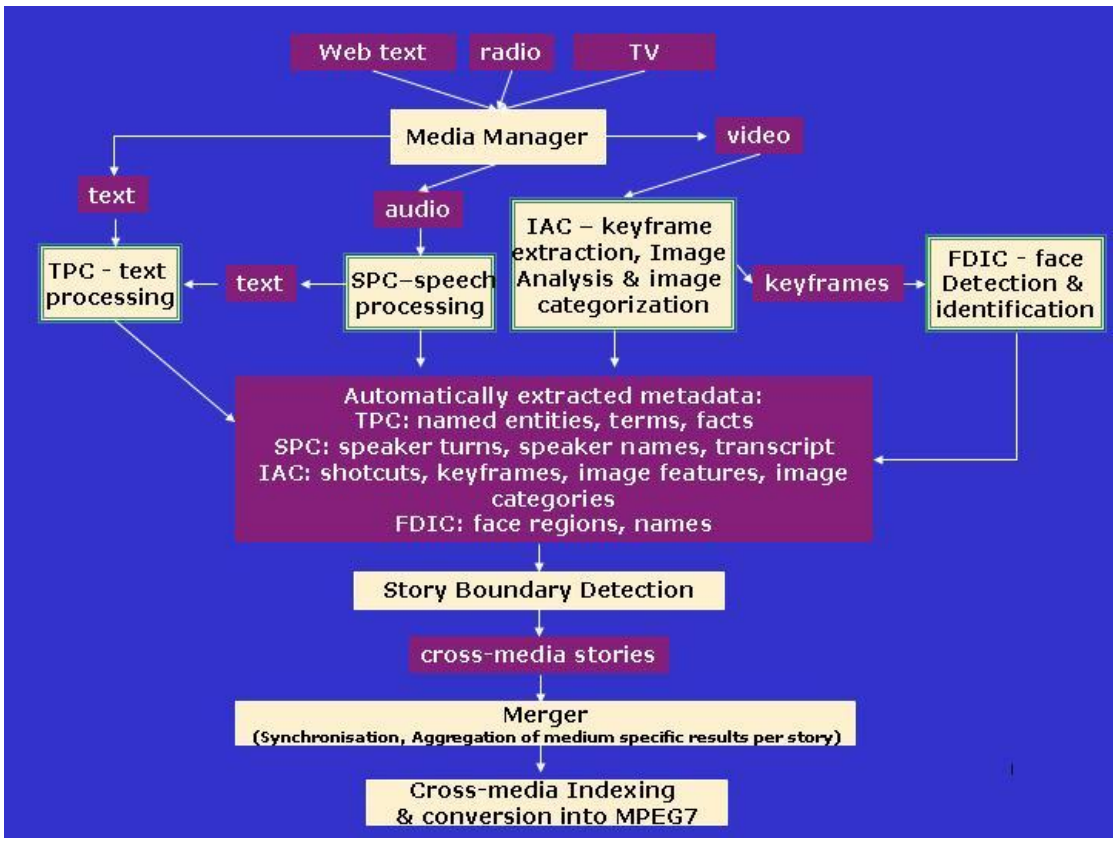

Figure 1: REVEAL THIS Cross-media Content Analysis and Indexing Subsystem

After deploying this indexical information and representations, REVEAL THIS makes use of state-of-the-art technologies for categorizing, summarizing and translating multimedia documents.

For a detailed presentation of the work on components of the Cross-media Content Analysis and Indexing Subsystem, please refer to Workpackage 4, in the next section "Progress per Work-Package".

Cross-media categorisation

The categorization subsystem operates along the cross-lingual and cross-media directions. In the cross-lingual dimension, a categorizer based on a pivot language category model (in English) is deployed. Documents in other languages go through a translation phase, thus enabling their categorisation. Such strategy overcomes constraints at the level of the training set, which often are not sufficiently big for building models in all languages involved.

In the cross-media dimension, documents containing not only text or images but a combination of different types of media (text, image, speech, video) are considered. The multiple-view fusion method adopted builds 'on top' of two single-media categorizers, a textual and an image categorizer, without the need to re-train them. Data annotated manually for both textual and image categories was used for training the cross-media categorizer. In that set, dependencies between single-media category systems have been exploited in order to refine the categorization decisions made.

For a detailed presentation of the work on the Cross-media Categorization Component, please refer to Workpackage 5, in the next section "Progress per Work-Package".

Cross-media summarisation Subsystem

The task of the Cross-media Summarization Subsystem (CSS) is to determine and present the most salient parts according to users' profiles and interests by fusing video, audio and textual metadata. The CSS subsystem consists of three major components: the textual-based summarization component, the visual-based summarization component, and the cross-media summarization component aiming at the fusion of the two analyses and creating a self-contained object. Additionally, the Cross-media summarization subsystem provides the necessary visualization interfaces enabling the user to preview a specific multimedia object before downloading a file of interest. Cross-media summaries for politics, news and travel-related audiovisual files have been designed and implemented in the project; the module interfaces also with a web-based translation service for creating "translated" versions of the summaries for English and Greek (both directions).

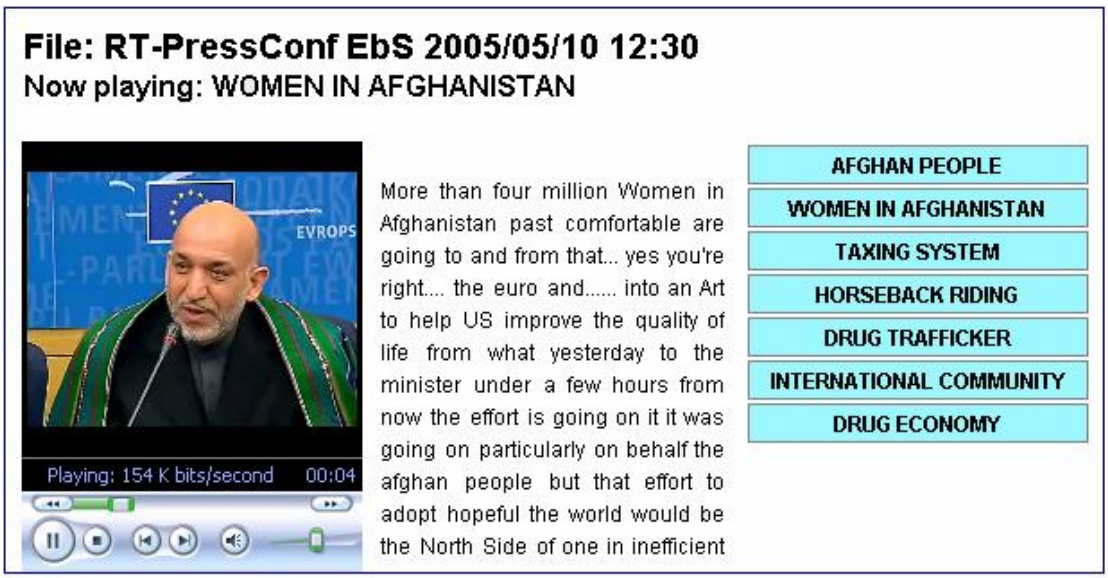

Figure 2: Multimedia summary (file summary)

For a detailed presentation of the Cross-media Summarisation Subsystem, please refer to Workpackage 6, in the next section "Progress per Work-Package".

Cross-lingual Translation

The Cross-lingual Translation subsystem (CLTS) allows users to query documents written in different languages, to categorise content expressed in different languages and to preview language specific summaries. A bilingual lexicon extraction module is used to generate lexical equivalences used for query translation purposes, but also to replace keywords in a target language, in case a document is linguistically not well formed (e.g. output from a speech recognizer) and thus not effectively translated. Last, a statistical machine translation module is responsible for providing translations of the textual part of the summaries produced by the Cross-Media Summarization Subsystem. The translation pair is English and Greek and translation takes place in both directions.

For a detailed presentation of the Cross-lingual Translation Subsystem, please refer to Workpackage 7, in the next section "Progress per Work-Package".

REVEAL THIS Integrated Application Prototype

The final system consists of two major components: The *Multimedia Indexer* and the *Media Server*. All results of the Indexer together with the original media files are uploaded to the Media Server. Media files are uploaded in two formats: in Windows Media format (.wmv/.wma) and in 3gp format (targeted at mobile devices). The server supports *searching for content* (pull scenario) or *filtering content* (push scenario) based on the metadata. It provides multi-lingual search, multi-media presentation of retrieved content, personalization, summarisation, and delivery to different devices and in particular to PCs/Laptops and mobile phones. The web-

interface of the prototype provides full search and retrieval functionalities targeting professional users with demanding information needs, while the mobile-interface provides a light-weight access to the different functionalities of the prototype, suitable for mobile-access and for laymen.

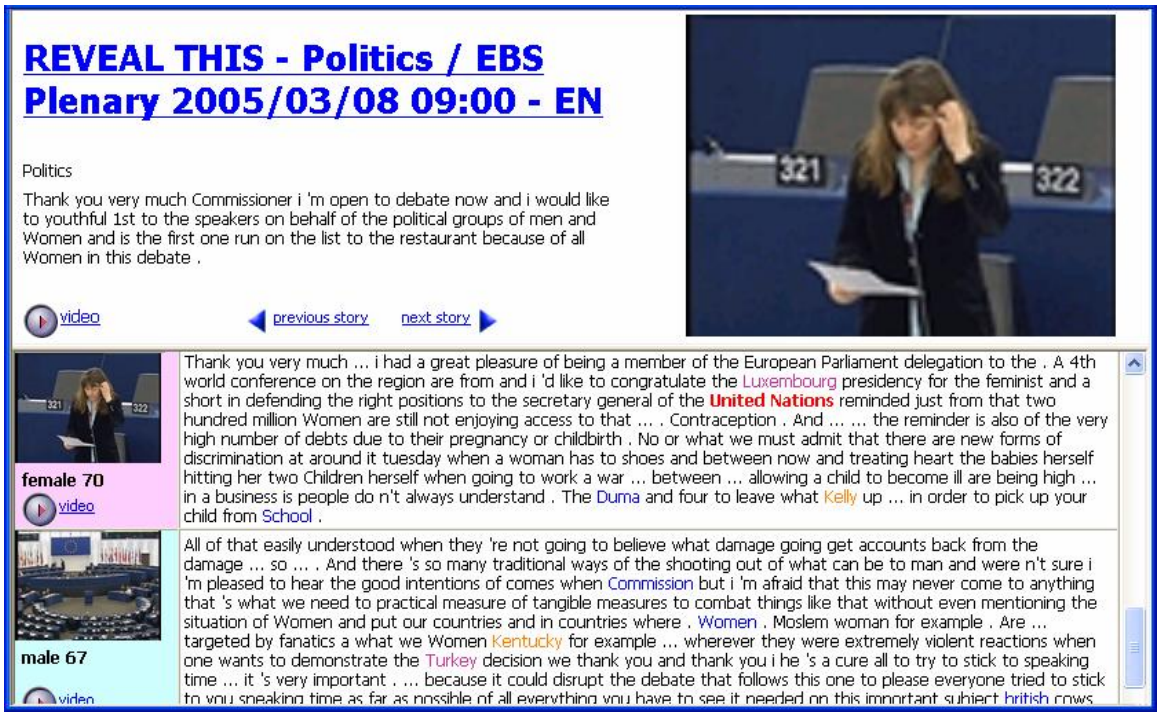

Figure 3: Story View in the REVEAL THIS web-interface

Figure 4: Screenshots of the REVEAL THIS mobile-application interface

For a detailed presentation of the Integrated REVEAL THIS prototype, please refer to Workpackage 8, in the next section “Workpackage progress in the reporting period”.

Last, usability evaluation of the REVEAL THIS prototype has taken place. The system was tested in the two domains of EU politics/news and travel and in two languages, English and Greek. Advanced users, experienced in multimedia search, used the prototype in its pull (retrieval) mode (the web interface), while novice users, with limited or no search experience used it in its push (filtering) mode through mobile phones. The results showed that the innovative REVEAL THIS functionalities were more than welcome by both user groups, the performance was generally satisfactory for them, though expectations for much better performance in terms of e.g. speech transcription and translation were evident.

A detailed presentation of the evaluation task and its results is available in Workpackage 9, in the next section "Progress per Work-Package".

3 Progress per Work-Package

3.1 WP2 User Requirements and Data Collection

Workpackage 2 consisted in three main tasks :

- conducting a survey of the market and analysing user requirements
- drawing up a development and evaluation plan for the whole project duration
- collecting the data necessary for training, testing and evaluating the REVEAL THIS prototype and its subsystems and components.

In particular, the first task was to **carry out and analyse the user requirements** and **the market survey**. The analysis followed two distinct directions: a User Requirements study complemented by a Market Survey. The first studied the need of potential content aggregators' users. The second analysed the characteristics, the players and needs of the content aggregator market with a view on the opportunities that such market might offer to a potential content aggregator service provider. Together, they provided a complete picture of the offer and the demand of the content aggregator market that was used to drive the high level design of the R-T system.

Since there was/is no system available that incorporates all the technology that R-T makes available and that gathers and integrates information in the way envisaged in the R-T proposal, the user requirements study based itself on existing systems that are as close as possible to R-T. This enables to show users existing examples of services that are working and that could be considerably improved with R-T technology. The choice of example systems was directed to *content aggregators*. In this broad category of systems we put Web content aggregators, RSS feed readers, TV monitoring systems, personal video recorders, etc. Each one of these systems is capable of gathering information from a multitude of multimedia sources and, to some extent integrate it somehow to present it in a coherent service designed for a specific purpose. Some examples of content aggregator systems were presented to the potential R-T system users (henceforth referred simply as users) to focus their attention to specific existing functionalities, so that they could comment on their importance, ease of use, and other aspects. It would have been difficult otherwise to elicit user requirements without some example of what R-T will or could be.

The first survey was conducted in Spring 2005 and keeping in view the objectives of R-T, five content aggregators had been selected: Google News, TVEyes, Newsburst, Awasu, and Headlinespot. The choice of systems reflected the popularity and functionalities of systems available at the time and was carried out in strict collaboration with project partners, drawing from their experience with similar systems. Finding specific appropriate examples to use in the user study was not simple. Each system has its strong point and its weakness and may or not be considered a good example of content aggregators. However, since we could not use them all, we tried to select a few that were as representative as possible of the different systems that can be considered content aggregators.

User requirements analysis in R-T was an ongoing process. Many new products were/are becoming available in the market and that may change user expectations and requirements dramatically. Therefore**,** a second survey was carried, nearly a year later, including this time a number of new products available in the market. This was done in order to keep in line the multimedia aspects of market systems with the R-T objectives; as a result, new aggregators featuring multimedia services were included and the questionnaire was re-designed accordingly. This second survey used the following systems: Google News, TVEyes, Yahoo video, topix.net, and Blinkx. In addition, it focused on different user groups and was generally more directed toward laymen than the first survey. Both surveys elicited users' opinion on some of the characteristics of these systems in two different ways: An online questionnaire and a personal interview.

In summary, looking at the findings of the surveys, the results of the user expectations and experience highlight the needs for R-T to design and implement a system that combines ease of use with advanced search functionalities. This enables the user to access a not-too-large number of high quality and reputable sources in both a push and pull way. In particular, the users indicated the need to access archival information with a combination of advanced search functionalities and the ability to browse and navigate the result set. (Details were documented in D2.1 and its updated version).

The ***market survey*** presented a picture of the internet multimedia market today, and the prime use made by users seeking information. Data was gathered by means of interviews, meetings and discussions (including US industry leaders). The Survey focused initially on the existing search market and the major players like Google and Yahoo!. The distinction was made between the primary function of a search engine, i.e. to provide a link for the user to a destination site where content may be found; and the emergence of search company strategies in providing aggregated and personalised services on the host site. Focus was on to the development of the key role of a Content Aggregator, the role of which will dramatically increase with the exponential explosion of content available on the Web. Details were documented in D2.1 and its updated versions, while further market surveys throughout the duration of the project were carried out and documented as part of WP10.

The combined results of the user requirement study and the market analysis enabled to identify the user groups to be targeted, and the characteristics of the information to be dealt with. In addition, it enabled the drawing of a development and evaluation plan that assured coherence of development and proper involvement of users at the

different stages of the project and in particular in the evaluation that was carried out in WP9.

The plan addressed questions related to the scope/range of involvement of user groups at the different stages of the project, and the provision of utilising user feedback effectively and systematically. Although a thorough investigation on where, when and how often information is gathered and verified was carried out as part of WP9, the plan indicated where to gather the necessary multimedia and multilingual data to be used in the rest of the project. The volume, quality and content of the data collected needed to be suitable for real tasks. Part of these data was also to be used during the design and development phases of the project for the training of the speech, text, and image processing components in WP4. Another part of the data was kept aside, to be used in the final evaluation of the system, during WP9. Furthermore, the REVEAL THIS ***development plan*** (Deliverable D2.2) adopted a stratified incremental approach for the development of the integrated REVEAL THIS prototype two main families of demonstration prototypes, with four intermediate releases in total were planned. In fact, the first version consisted in a targeted end-of-year 1 demonstrator for peer reviewing purposes, an improved version of which resulted in Prototype 1, available in M15 of the project. An improved version, Prototype 2, including the final versions of the content categoriser, summariser and translator was planned to be available in M20 of the project, while the Final Prototype was planned to be available by the end of the second year of the project. The development plan of the prototypes was organised according to domains, languages, improving component performance and increasing system functionality.

Last, for the development and testing of the REVEAL THIS subsystems and the usability evaluation of the final integrated prototype a large data collection was compliled with a number of idiosyncracies. The REVEAL THIS **corpus** is uniquely multifaceted (cf. D2.3 and updates for details on the corpus):

- It is *multiple domain*, covering politics (EU politics in particular), traveling (travel information for tourists), news (national news in Greece, Britain, Belgium/France and web news on politics, travel and health) and health.
- It is *multimodal*, i.e., it consists of video, audio, text and images (the latter accompany text, i.e., illustrated text documents)
- It is *multi-source*, i.e., it comes from a variety of sources such as TV channels (satellite and terrestrial), radio broadcasts, and the web.
- It is *multilingual*, i.e., it contains data in more than one language, and in particular in English (EN), Greek (EL) and French (FR).
- Part of it is *parallel* in all three languages (e.g. all European Parliament speeches and most press-conferences, some travel documentaries etc.), another part is parallel in two languages (e.g. travel documentaries in EN accompanied by subtitles in EL), and the rest is *comparable* in the wider sense of the term, i.e., thematically interrelated data.
- It contains multimodal documents of different genre, which guarantee a significant variety in modality-specific characteristics

The corpus has been divided into 4 so called core development data sets (in short Core): Cores 1, 2 and 3 consist of data covering the domain of politics which are collected at different points in time, while Core 4 covers the domain of travel. Data in Cores 1 and 2 are used for training and development purposes in the politics domain, while Core 3 has been reserved for testing. A subset of data in Core 4 is used for training and development purposes in the travel domain, while another subset of it has been reserved for testing.

3.2 WP3 System Specifications

The purpose of this work-package was to define in full detail the technical specifications for all software modules that would be implemented in workpackages WP4-7. The consortium established a common understanding of how technical possibilities and user requirements could be unified and defined a system architecture. System design, in terms of subsystems and components, data flow and interfaces between different components were defined. Specifications of the speech processing component, shot-cut detection, keyframe extraction and face detection and identification were made available. The consortium focussed on issues such as online vs. offline processes for both summarization and cross-lingual translation, and the distinction between categorization and indexing in terms of the optimal component configuration.

Progress towards objectives

The original system design was slightly revised during the course of the project for accommodating specific implementation needs. All details are documented in D3 and its updates.

3.3 WP 4 Cross-media Content Analysis and Indexing Subsystem

The goal of this work package was the development of the Cross-media content analysis subsystem that analyses multi-media data streams. REVEAL THIS combines technologies falling into three categories, corresponding to the research fields from which they have sprung: Automatic Speech Recognition (ASR), Image Analysis (IA) and Natural Language Processing (NLP). The whole subsystem consists of (i) the Speech Processing Component (SPC), performing speech recognition and speaker identification, (ii) the Image Analysis Component (IAC) performing spatiotemporal video decomposition (shots, subshots) and categorizing the contents of key frames, (iii) the Face Analysis Component (FDIC) performing face detection and face identification, (iv) the Text Processing Component (TPC) performing fact extraction and (v) the Cross-media Indexing component (CMIC) establishing links between different multimedia objects across media. Finally, technical evaluation has been completed for each and every component of the analysis and indexing subsystem.

Task 4.1.1 Speech Processing Component

Objective

The Speech Processing Component receives the audio track of TV and radio programs and performs a sequence of analysis steps including segmentation based on audio

characteristics, speech recognition and speaker clustering. WP4 aims at the adaptation of the Speech Processing Component (SPC) in the target domain. The SPC comprises speaker identification (SI) and automatic speech recognition (ASR) modules. To support this goal, new models (acoustic model, language model and speaker-id) should be built on the basis of training data collected in WP2. Moreover, new ways of building acoustic models should be investigated in order to handle colloquial and spontaneous speech better, and therefore improve the recognition accuracy.

Status at beginning of the project

SAIL contributed to this project the Speech Recognition Engine, running in real-time, the acoustic and language models in high quality for English and an initial system for Greek. A very slow and long-running process for building the acoustic models and the language models was originally in place. The training procedure for speaker-id training was depending a lot on the know-how of a person to operate a sequence of independent programs.

Progress throughout the project

SAIL focussed its activities on the following tasks with the objective of improving its speech recognition engines for both languages of the project and for tuning them to the specific domains covered in the project:

a) New model language building process:
The most valuable achievements of the first year were improvements in the way the Language Models were built, which affected considerably the performance of the speech processing component.

b) Improvement of Greek Language Model:
Using this functionality and a lot of effort for creating a new training corpus, cleaning and tokenizing that corpus and building a variety of acoustic and language models, SAIL was able to improve the Greek system to the level of other languages both in terms of recognition accuracy as well as in terms of speed and memory footprint required.

c) New acoustic model building process:
To be able to use more training data for acoustic training and utilize this data better, SAIL revisited its acoustic model training process. Its legacy process used to be limited to processing a total of 150 hrs of data which was no longer adequate given the state of the art in model building with data collections approaching 1500 hrs. It therefore implemented a new training process and interfaced this to its automatic speech recognition (ASR) decoder using traditional Maximum Likelihood estimation. Using this new framework, it implemented a new training method based on the "Maximum Mutual Information Estimation" criterion.

d) Application of a new training process:
During the second reporting period, SAIL worked on both basic research and project application specific tasks. It tested its new Language Model building process on domain-specific data and observed up to **15% improvements** in Word Error Rate (WER). Using its new acoustic model training process from start-to-end in both Greek

and English it saw a statistically significant improvement in recognition accuracy on the acoustic model side while still retaining the same amount of training data. As shown in table 5, the improvements in several test sets largely outweigh the small reductions in WER in a few others. Particularly the languages of the Reveal project profit from the changes. The line el.score refers to the Greek test set, while the line biznews is an English test set.

	old		new	
	correct	wer	correct	wer
Aljazeera_20041016_0300_News.score	64	39,8	64,5	39
Al_Arabia_20041015_0400_News.score	78,2	30,3	79,1	29,4
biznes_2006_01_04_pcm.score	87	17,8	87	16,6
BR199912051845.score	77,4	33,1	77,2	33,4
el.score	79,7	21,3	79,7	21
ES_20000614_2100.score	85,5	30,8	85,6	31,2
RTBF_20011206_1250_Journal_Televise.score	72,5	38,3	72,4	38,6

Table 5: Recognition accuracy with new acoustic model (same training data) for EN and EL.

e) Speed improvements and data compression
When using a larger amount of language model data, the resulting models are also growing in size, yielding to a slow-down in decoding due to the sparseness of memory access during decoding time which leads to congestion on the memory bus. We implemented a compression scheme to shrink the size of the Language Models by up to 40% with only insignificant loss in speech recognition accuracy.

f) Cross-media segmentation
Using input from KUL, SAIL worked on cross-media segmentation of audio/video streams. This cross-media research did not seem to work satisfactorily (see D4.4.v2.0 for details).

g) Speaker Identification
Using data provided by REVEAL THIS, we trained a Speaker Identification Model and re-processed all demo and benchmark data for both Greek and English data sets.

Generally, SAIL's systems were benchmarked using publicly available data from TC-Star, another EU project concerned with Speech to Speech Translation. SAIL built new models better matching the application domains of REVEAL THIS and reprocessed the demonstration and benchmark data numerous times.

Self-Evaluation against initial Objectives

In what follows, we would like to compare SAIL's achievements in the project with the initial objectives:

WP4 aims at the adaptation of the Speech Processing Component (SPC) in the target domain.

The results in the target domain have significantly improved over initial results. Therefore the overall goal is met.

To support this goal, new models (acoustic model, language model and speaker-id) are built on the basis of training data collected in WP2.

New models have been built based on the data derived from WP2. All models work and are installed in the demonstrator.

Moreover, new ways of building acoustic models are investigated in order to handle colloquial and spontaneous speech better, and therefore improve the recognition accuracy.

New training procedures, for both the acoustic and the language model training have been developed. The new models for English and Greek have been built on the basis of subject training procedures. Several intermediate technical problems have been overcome and the final training system is robust and ready. It is already being deployed in other projects.

Without doubt the position of the SPC in the international markets has been improved. The fundations have been laid for keeping up with the technological advancement in other geographies, particularly in North-America.

Task 4.1.2 Image Analysis

Objectives

The Image Analysis Component performs analysis of the image content of video keyframes to determine tags relating to the object and activity content of those keyframes. This component will also exploit video processing functionality to segment video into shots and select keyframes. The main modules of the image analysis component include methods:
1) To detect and describe characteristic patches in scenes,
2) To describe and classify these patches into a “visual alphabet” which is a discrete set of alternatives that is suitable for subsequent classification by machine learning methods,
3) To model characteristic 2D geometric configurations of these patches.

Status at the beginning of the project

The Generic Visual Categorizer (GVC) XRCE had at the beginning of the project was based on bag of keypatches but used only one visual vocabulary. Furthermore, no category system related to the domains covered in the project were available, as well of course as no associated image categorizers. Lastly, patches extraction and description modules were based on external tools, a dependency XRCE wanted to avoid for the project.

Progress Towards Objectives

XRCE developed a novel approach to GVC. Many state-of-the-art GVC systems are built around a vocabulary of visual terms and characterize images with one histogram

of visual word counts. The proposed approach is based on a universal vocabulary, which describes the content of all the considered classes of images, and class vocabularies obtained through the adaptation of the universal vocabulary using class-specific data. An image is characterized by a set of histograms - one per class - where each histogram describes whether the image content is best modeled by the universal vocabulary or the corresponding class vocabulary. XRCE found experimentally that this novel representation greatly improves the performance compared to an approach based on a single vocabulary at the cost of a modest increase in the computation (improvement in the range of 8% - absolute difference).

XRCE tested different approaches to score normalization. The output of the categorizer is one score per class, with the scores taking values in the range minus infinity and plus infinity. Proper score normalization is important to (i) set a robust threshold for the acceptance / rejection decision and to (ii) combine the output of the GVC system with, e.g. the text categorization system. Different approaches were tested and found that a simple sigmoid fit provided the best results.

XRCE defined an image category system related to the project domains. To this end, it pre-selected over 10,000 representative frames from Discovery Travel data eliminating non-relevant images (advertisement, duplicate removal, etc.). A subset of 5,000 images was manually annotated by several users, each image being annotated by at least 2 users. It analyzed these annotations with respect to (i) image class distribution with respect to our set of 42 predefined categories (ii) correlation between different annotators for different classes. As a result, we decided to retain for each image as main labels only the labels chosen by several users. However we kept the other labels in order to prevent the usage of the respective image as negative training example for the corresponding categories.

The previous version of the categorizer used as low level features texture information extracted from grey-level images. On a set of 27 categories which are of interest to the project (i.e. relating to travel and leisure activities) XRCE obtained a 66% correct rate with such features. To improve the performance of its system, it experimented with different local color features while keeping the same framework for classification. It considered features such as local mean and standard deviation computed on a regular grid and different color histograms in various color spaces: RGB, LUV, CrCb, R/(R+G+B) G/(R+G+B). The best classification accuracy of 64% was obtained through the concatenation of local mean and standard deviation in RGB space. XRCE also tested different approaches to merge the orientation histograms and color features: at the low-level feature, at the high-level feature (i.e. at the histogram level) and at the score level (late fusion). The experiments showed that the latter approach performed best and a classification accuracy of approximately 74% could be achieved.

XRCE further developed two new versions of the categorizer:
In the first one, weak local geometry through the context considered for each local patch was integrated. This was done by the computation of high-level local features given by local histograms of low-level visual word co-occurrences in a spatial neighborhood. Similarly to the previous system based on low level local features (SIFT and color), universal and class-adapted vocabularies of these high level features were built and one-versus-all linear classifiers were trained and tested on the bi-partite

histograms. The approach based on these high-level features improves on a large number of categories on both highly structured objects (e.g. Bicycling and Urban) and weakly structured scenes (e.g. Beach and Wintersports). See further details in [†].

The second categorization system was build using Fisher kernels, which allows to characterize a signal with a gradient vector derived from a generative probability model and to subsequently feed this representation to a discriminative classifier. XRCE applied this framework to image categorization where the input signals are images and where the underlying generative model is the visual vocabulary of low-level features in images. XRCE also has shown that Fisher kernels can actually be understood as an extension of the popular bag-of-visual words. The approach demonstrates similar performance as the adapted vocabulary approach; however it has much lower computational cost both at training and test time. See further details in[‡].

Finally, the image categorizer has been integrated with the text categorizer in the multi-media categorizer (see WP 5).

Task 4.1.3 Face Detection and Identification

Objectives

The purpose of the face analysis component is to detect and identify faces in video keyframes. This task naturally divides into two parts, a face detection component (FDC) and a face identification component (FIC). These components operate sequentially. First, the FDC scans the image to find the locations and sizes of human faces visible in the image, if any. The information extracted in this step is then handed over to the FIC, which analyses the faces to discover their identity. The main challenge of the face analysis component is to make this process as robust as possible, while keeping computation times reasonable.

Status at the beginning of the project

A working face detection algorithm, based on earlier work done by KUL, was available at the start of the project. However, KUL had no previous experience in the face recognition domain. So the face identification component had to be designed and implemented from scratch. Given the time constraints of the project and the complexity of the task, this implies that most of the work on the face analysis component would go towards developing a state-of-the-art face recognition algorithm.

Progress towards objectives

[†] Florent Perronnin, Universal and Adapted Vocabularies for Generic Visual Categorization, IEEE Transactions on Pattern Analysis and machine Intelligence, paper accepted with major revisions.

[‡] Florent Perronnin and Christopher Dance, Fisher Kernels on Visual Vocabularies for Image Categorization, CVPR 2007.

Concerning face detection, the method developed earlier by KUL was shown to be outperformed by the Viola and Jones method. Therefore, a new component was based on the OpenCV implementation of the latter algorithm. The resulting component is very efficient and capable of detecting faces in the frontal pose as well as faces which are seen at an angle. Full profile views are usually not detected.

For the face identification, the decision was made to follow the approach pioneered by Blanz and Vetter. This method was chosen since it was deemed most suited to the task of recognizing faces under a wide variety of circumstances such as variations in viewing angles and illumination. The approach makes use of a so-called 3D Morphable Model (3DMM), which is a statistical way to describe the 3D shape and texture (i.e. colour) of the human face. In the method of Blanz and Vetter, a 3DMM is used to approximate the 3D shape and texture of a face in an image, and the identity of the face is determined by comparing the resulting model parameters to a database of known faces. Neither the algorithm, nor any suitable 3DMMs were publicly available at the start of the project, so everything had to be developed from scratch.
The 3DMM developed by KUL for this project was learned from 107 laser-scanned faces available in the USF DARPA HumanID 3D Face Database. This effort was completed during the first year of the project.

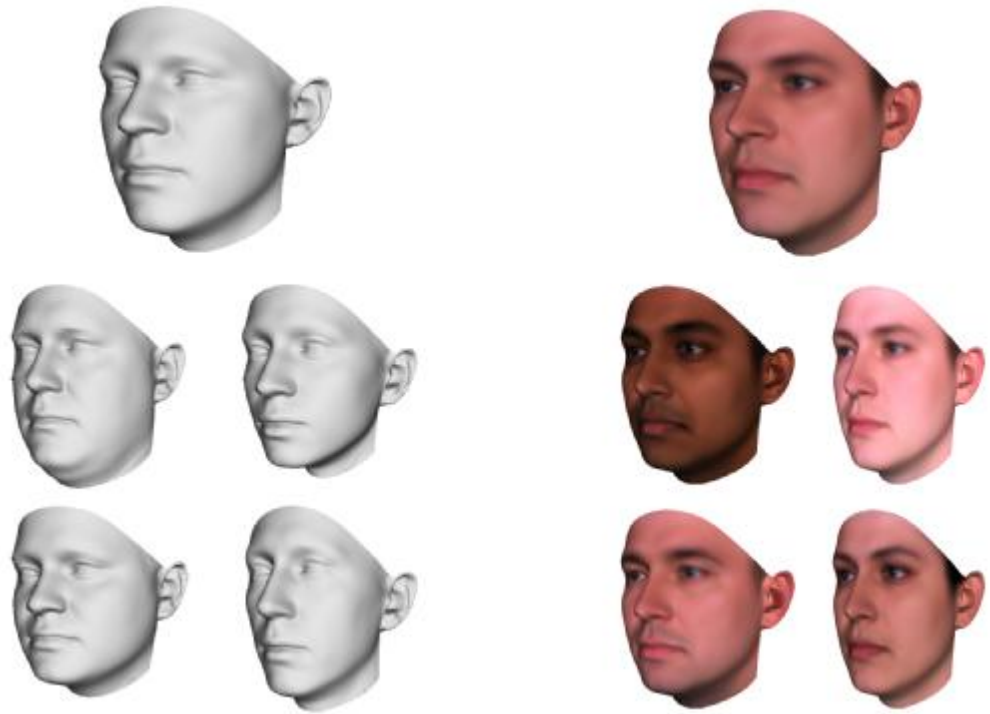

Figure 5: The 3DMM developed by KUL. Left: average shape and the first two modes of the shape variations. Right: average texture and the first two modes of the texture variations.

Most of the time spent on the FIC went into the development of the fitting algorithm, which matches the shape and texture of the 3DMM to the appearance of a face in an image. A major part of this was the development of an intelligent, modular optimization framework, enabling researchers to easily experiment with alternative configurations of the model rendering pipeline. Using this framework, a 3DMM fitting algorithm similar to the one of Blanz and Vetter was built and later extended to account for partially occluded faces. The extended version of the algorithm was

particularly successful, improving recognition rates in all scenarios and most significantly for faces showing non-neutral facial expressions or partially occluded by foreign objects (glasses, hair, microphones, or earsets). Another important achievement was the development of a method for automatically determining the 3D orientation of a face in an image based on the information supplied by the FDC and the colour content of the image (i.e., the locations of skin colour pixels). This made the face recognition algorithm fully automatic, whereas the method of Blanz and Vetter still relied on manual intervention to initialize the orientation of the face.
The final system was trained and optimized to recognize the faces of 21 politicians appearing in EbS video footage of the European Parliament. A recognition rate of 86 percent was achieved on EbS video frames, which is very high considering the difficulty of the material.

Figure 6: Result of applying the fitting algorithm to typical images from the EbS footage. Top row: original images. Bottom row: original images with the 3DMM fitting result superimposed.

Finally, KUL decided to address the lack of good 3DMMs available to the research community, which it sees as a major obstacle towards scientific progress in this domain. Therefore, the last few months of the project were spent on improving the quality of the 3DMM itself, with the intention of releasing the resulting model to the research community.

Task 4.1.4 Text Processing Component (TPC)

Objectives

The Text Processing Component (TPC) deals with the processing of the generated transcriptions produced by the ASR engine as also with written textual streams (web data). The TPC comprises Named Entity Recognition (NERC), Term Extraction (TE), and Fact Extraction (FET) modules.

Status at the beginning of the project

Generic language processing tools (including part of speech taggers, morphosyntactic analysers, chunkers) as well as term extractors and named entity recognisers for the

finance domain were available for Greek and English. These tools were not directly usable on project domain data. Fact extraction technology had to be developed from scratch.

Progress towards objectives

*1. **Named entity recognition***

The task of the NERC module is to identify all named locations, persons and organizations in the text transcriptions produced by the ASR engine or /and in the textual web data. To this end, the NE recognition system (MENER) developed at ILSP, was trained on material properly annotated. For each annotation level concerning the target domains (politics and health) and languages (EN & EL), initial guidelines were compiled and provided by a team of linguists. Manual annotation of EL and EN data was performed pertaining to the domains of news, politics and travel. To sustain compatibility with the ACE schema, REVEAL THIS catered for an extended annotation schema with subtypes for further classifying the spotted NE's. Due to the type and content of the project data, however, only subtypes of LOC entities were used. Moreover, to sustain consistency in annotation, a condensed classification schema was finally used. To this end, mappings of the extended classification schema were performed: GPE was mapped on LOC (location), and LOC sub-classification was dropped.

Moreover, training of the models for the aforementioned languages and domains was not solely based on the REVEAL THIS data. The training material used for the English news model originated from the English data provided at the CoNLL-2003 shared task. The material is part of the Reuters corpus. It consists of three parts: the core data (204k tokens), a development set (51k) and a test set (46k), all provided in a line-delimited textual format that ILSP converted into an equivalent XML representation. Additionally, the training corpus of the Greek news model was an agglomeration of mainly news documents from the following sources: various Internet news sites (9k tokens), the Greek Business Channel (GBC), a media provider in Greece (35k) and data that originated from the European Parliament (EP) web site (66k). GBC data consists of short daily news bulletin files, covering economical, domestic and world news, which briefly mention 6-8 events per file. Annotations relevant to the TPC were carried out by means of a purpose-designed tool developed at ILSP, called Marker.

The NERC module combines sentence-based local evidence about words, with global evidence, which is a collection of features drawn from other occurrences of those words within the same document (global features). In this respect, ILSP conducted small experiments on the basis of which we decided to further enhance these features either locally or globally. To this end, statistical data were further coupled with frequency information on *character sequences of 4-characters length occurring within spotted NE's*. Additionally, the frequency of words occurring with a *capital initial* in the training data was estimated. Finally, word accent in Greek texts was eliminated in order to normalise texts and cater for the enclitic accent due to intonation. The inclusion of the above features returned a slight improvement in the initial results.

Unseen data from the REVEAL-This corpus, amounting to approximately 63k tokens for EN and 16k for EL in the News – Politics domain, as well as, 7k for EL and 4k for EN for the Travel domain have been hand annotated following a formally configured annotation schema. To estimate system performance, the evaluation software provided for the CoNLL2003 shared task was used.

Quantitative results of the NERC module have been presented in detail in D4.4. The module performed well in Greek news and politics texts (94.87% F-measure) outperforming the best CONLL2003 system in the same domain (different languages, different training and test set). It achieved a 74.27% F-measure in Greek travel data , and a 71.4% F-measure in English travel data. However, its performance in English news and politics was lower (68.98% F-measure); low performance has been attributed mainly to the fact that training of the model for the aforementioned language and domain has not been solely based on the REVEAL-This data. The training material used for the English news model originates from the English data provided at the CoNLL-2003 shared task and is part of the Reuters corpus, while the testing was done exclusively on European Parliament data, which have its own idiosyncrasies mainly with regard to organisation names.

2. Term extraction

The task of the TE module is to identify all terminological units. To this end, the TE module developed at ILSP was further developed and tuned for the domain-specific needs of the project. The TE module is a hybrid system comprising a term pattern grammar based on finite-state technology, and a statistical filter, used for the removal of grammar-extracted terms lacking statistical evidence. In this context, the following resources were compiled:

Grammars: Candidate multiword terms are recognized by context free grammars that ILSP developed for EN and EL respectively.

Corpora: In order to calculate the inverse document frequency corpus (FrCo) of grammar-produced candidate terms we compiled databases that contain corpus and document frequencies for each lexical token. More specifically, in the case of Greek, a reference corpus collection of ~10K texts in the politics domain was used. Pre-processing for this included tokenization, lemmatization and POS tagging. In the case of English, the Reuters-21578 Text Categorization Test Collection (~18K texts) was used. Pre-processing stages included tokenization, lemmatization and stemming.

The Term Extraction module has been tested on two different corpora, one per language (EN, EL). For the Greek language, a manually annotated corpus of approximately 65k words was used. The English corpus comprises of several manually annotated texts taken from the REVEAL THIS core1, core2 and travel datasets, approximately 63k words in size. Gold annotations on these data were prepared using the ILSP Marker software.

Evaluation of the Term Extraction module was measured with the standard recall and precision figures. The system's recall was considered satisfactory. On the other hand, the precision figure was considerably, but not unexpectedly, low since term grammars inherently produce numerous terms because of the generality of the rules. Analysis of

the results was carried out revealing the prevalent impact of three major factors responsible for low precision:

(a) Incorrect part-of-speech identification (e.g. adverb instead of adjective or verb instead of noun),

(b) many human-produced terms are included in a bigger machine-generated term since the grammar pattern can sometimes capture very lengthy candidate terms (e.g. human spotted "beach" and the term extraction module spotted the greater "most famous beach", human spotted "island" and the module spotted "rocky island soil"),

(c) inadequate frequency statistics. Many human-crafted terms have either low frequency in the news texts because of small or zero frequency in the reference corpus, resulting in a very small tf-idf performance.

3. Fact extraction

The task of the FET module is to identify and extract the facts that give evidence of the domains targeted by the project. Fact extraction is expected to bring substantial help towards the ultimate goal of the TPC, namely the rich indexing of the content across the domains and the languages of interest. With respect to this objective, ILSP decided to provide more than one versions of the module, employing different methodologies in order to cover different processing scenarios. FET v1 deals either with an original text document or with text produced by the speech recognition component, targeting all the facts expressed in the REVEAL THIS data. Processing concerns English and Greek data in the domains of Travel and EU politics. According to the implemented algorithm, for each sentence facts are anchored on Named Entities denoting people, locations and organizations and/or Terms.

A new version of the Fact Extraction Module (FET v2) was finally released by ILSP, which (a) takes advantage of the fact that preprocessed TPC files (originating from written textual streams) contain dependency-based syntactic representations for each sentence, (b) integrates information from a bilingual, domain-specific shallow ontology, and, (c) exploits information concerning the semantic arguments of each fact.

In order to fulfil the above requirements, the TPC pipeline processing was expanded to accommodate syntactic analysis. In the case of written textual streams, pre-processing also involved syntactic analysis and recognition of dependency relations between tokens of each sentence. SPC-derived data were not processed at this level, since initial experiments by ILSP showed that errors in automatically generated transcripts make this input unsuitable for full syntactic parsing. A dependency-based representation was chosen on the basis that it allows for more intuitive descriptions of a number of phenomena, including long-distance dependencies, as well as structures specific to languages like Greek that exhibit a flexible or free word order. At the same time, dependency representations capture syntactic relations that in many cases are close to a semantic representation of each sentence (e.g. Subject-Agent).

For syntactic analysis of EN data, ILSP used the Stanford parser on 48 hierarchically-organised relations. For syntactic analysis of EL data, it exploited the MaltParser platform, via which it trained a memory-based dependency parser for Greek. Training data at the level of syntax consisted of ~70 KWords annotated using a dependency-based syntactic scheme. This set of manually annotated training data constitute the first-ever Greek Dependency Treebank, which was developed within REVEAL THIS for the needs of the fact extractor and which was released to the research community for use in the CONLL 2007 dependency parsing challenge (cf. section 4 on contributions beyond the state of art and appendix A on dissemination activities). The dependency relation set in this data comprises 25 main relations, and is similar but *not* identical to the dependency set used for EN. Parser evaluation on gold data showed an overall LAS (labeled attachment score) of 78.06% and an overall UAS (unlabeled attachment score) of 71.11%. Precision and recall for the subject relation reached 70.94% and 76.46% respectively.

Facts extracted were further enriched with semantic information pertaining to the *semantic type* and the *participants* for each fact. Semantic types correspond to nodes of shallow domain-specific ontologies that are expected to provide the user with valuable information about the facts he/she wants to retrieve. To this end, ILSP built a lexical database that was based on the identification of the most significant events attested in the data. This lexical database includes the following:

- lemma: it corresponds to the verb of interest
- example: it denotes the specific sense/ use of the verb
- semantic arguments: it corresponds to the participants involved in the event denoted by the verb
- type: it corresponds to the type of the of the event denoted by the verb
- subtype: it corresponds to the subtype of the event denoted by the verb

Task 4.1.5 Cross-Media Indexing Component (CMIC)

Objectives

The Cross-Media indexing component is the last layer of the chain of processes in the cross-media indexing and analysis subsystem (CMIC). CMIC leverages the individual potential of indexing information provided by different analysis modules such as speech, text and image to improve the effectiveness of information retrieval/filtering at a later stage. It hypothesizes that a system, which combines different media descriptions of the same multi-medium segment in a semantic space, will perform better at retrieval/filtering time. In order to validate this hypothesis, CMIC takes the medium-specific analyses of an audio-visual segment as input and creates links among them in order to show topical similarity in the text space.

Status at the beginning of the project

There were information retrieval and filtering tools available to deal with individual modalities, however, none of them was designed to cope with cross-media indexing.

Progress towards objectives

The indexing task requires merging and processing of a large number of digital content streams online and proactively searching for semantic links between the results of the pre-processing components that give evidence to support a topical similarity in a per story basis. The component's input is the Merger module's output; an XML file that contains every pre-processing unit's output. After analysis, the Cross-Media indexer creates an MPEG-7 representation in two major parts. The first part contains the merged file that is mapped to its MPEG-7 corresponding description tags. The second part contains the enrichment, i.e. scores for expressing how representative each medium-specific description within a story is for its indexing. The retrieval engine subsequently uses this part. When necessary, the MPEG-7 standard is extended to meet the specific needs of the Reveal-This project.

The Cross-Media Indexing component developed by USG is a standalone, independent unit that is able to process and transform the results of pre-processing units. Fundamentally, it achieves this through three processes. The first and the second stages are together called an Analysis phase and the later is called Indexing. The analysis phase corresponds to parsing and transformation of an input stream, noise filtering and lexical analysis tasks. However, the most important phase is the indexing phase, which has to deal with a number of challenges in the process of establishing links between media. These are: a) the amount of indexing information and noisy data; b) uncertainties in the single media; c) missing data, and d) inconsistencies between values received. The indexing approach should help to avoid or otherwise minimise these problems as much as possible. CMIC works on the textual space using Dempster-Shafer's Theory of Evidence. This approach is used for establishing links among the modal descriptions in order to depict topical similarity in the textual space. The theory has been already extensively studied in image retrieval and structured document retrieval, but has never been applied in a multimedia retrieval context.

In general, indexing and searching for multimedia units constitute a great challenge due to their inherent diversity and composite nature. Unlike text, other modalities such as audio (i.e speech) and image are difficult to work with. The diversity of modalities require dissimilar examination techniques specific to their characteristics. Cross-Media (the term Multimodal is often used interchangeably) analysis and retrieval research is a highly interdisciplinary field and has recently received attention of various other research communities. It is the product of recent advances in speech, image, text analysis and retrieval research and many others. In its widely studied form, the subject matter of the field is composite content such as audio-visual segments or even multimedia rich web sites. Together with the TRECVID initiative and thus the availability of a test collection, numerous works have been done reporting varying results on the assumptions of the research area.

Evaluating the accuracy and robustness of individual processing models should be done in their own domain and with specially crafted test beds. However, when it comes to evaluating the performance of the merged and enriched representation of these processors, one needs to find a solution.

In an attempt to validate its arguments and thus its prototype system, USG considered two different testing strategies; known item search and task oriented user test, both involving the construction of a test collection using real users. The first one is a task

oriented user experiment known as known item search. The second approach is to annotate the stories in a small collection by expert judgements. The challenge is preparing and building the collection with relevance judgments easily and quickly. Accomplishing this task properly is very time consuming.

For the evaluation of CMIC (by USG), ILSP reserved a subset of the REVEAL THIS corpus, comprising of five hours of English data including EU Parliament plenary sessions and press-conferences, news, and travel documentaries; ILSP designed and implemented the manual annotation of the test-data with relevance judgements that were used as ground truth for evaluating the performance of CMIC. Despite being relatively small, the intention was to have a data collection completely annotated with graded expert relevance judgments. ILSP guided the creation of a set of 87 queries by laymen users according to sketched information need scenarios and the subsequent creation of relevance judgments by three laymen users (relevance of each story in the test-data collection against each of the 87 queries). Experts were asked to go through the collection exhaustively for each query and provide relevance judgements. The experts were also asked to justify their judgements when considering stories as "partially relevant" to indicate how strict/lenient they had been in their judgements. The segments that were not mentioned at all were judged as irrelevant by the annotators.

The evaluation results suggest that the Dempster-Shafer (DS) approach is not performing well and worse than a baseline system that uses a standard information retrieval (IR) model. USG could not find a clear explanation for this, as the model it was using was not dissimilar to what has been used successfully by other researchers in similar contexts. In addition, some toy testing of the model seemed to suggest that it should perform at least as effectively as the baseline. Further more, despite being better than this model, the precision and recall values for the baseline system did not seem very high either. This may also be the result of factors such as the size of the collection or the imprecision of individual modalities. Further analysis of this empirical result also showed that there is no positive performance contribution of modalities that are other than speech and, although not significantly, in some cases it was observed that modalities other than speech altogether had decreased the performance in both the baseline and the DS system. In other words, running both systems using speech only content yielded better performance. Given this setup and experiments, USG concluded that the speech modality seems to be dominant over other modalities and that the contribution of other modalities to CMIC is negative, rather than positive.

While this could also be explained by some fault in the cross-media indexing model, further analysis somehow pointed to a different direction. It suggested that the bad performance could be due instead to some problem with the test collection. In fact, the above conclusions could be explained also by the following three possible situations: 1) The test queries might be biased towards speech 2) The size of the vocabulary is larger for speech than for any other modality 3) Reducing every single modality to textual space may not be appropriate for capturing relations. Further analysis, while not confirming any of the above hypotheses, confirmed that in the collection speech is by far the dominant medium and only if treated as such the CMIC performs decently. The use of evidence provided by other media is often detrimental to the performance of CMIC. This indicates that relevant documents in the test collection have been

almost always identified by relevance clues found in speech and not in other modalities. However, in some cases this produced some relevance errors, as the relevance was to be indeed found in other modalities and not just in speech. The DS model might catch these situations, but in doing so it produces a false positive according to the relevance assessments. This might be a plausible explanation of the poor performance that is supported by the failure analysis of some document retrieved. Further experimentation would be needed, but the REVEAL THIS test collection is the only one available for the specific domains and with a per story structure and building a new or larger one is beyond the scope and resources of the project. ILSP will make the test-collection available to the research community along with the guidelines for setting up the query formation and eliciting the relevance judgements, in an attempt to address the lack of such resources for evaluation of cross-media information retrieval systems.

3.4 WP5 Cross-media categorization subsystem

Objectives

The goal of this WP is to develop state-of-the-art categorization components that work across media and across languages to provide categorization capabilities (for filtering, routing and/or archiving).

Status at the beginning of the project

At the beginning of the project, XRCE had a generic text categorization system, with a version working on large numbers of categories. Neither category systems nor categorizers were available in the domains retained in the project (Travel, Health and Politics). In addition, the existing categorization system did not include multilingual and multimedia solutions.

Progress Towards Objectives

During this project, XRCE worked on the following:

It developed category systems for the two domains (Politics, Travel), and associated text categorizers working at different granularity levels. Expert users in politics for example can have access to a detailed set of categories, while more naïve users will be able to rely on broader categories, better aligned to their knowledge of field.

In the case where the number of categories is large, XRCE derived an upper bound for the error rate for each potential category system to be used at a given level within a cascade of categorizers. This upper bound leads to a better choice of a cascade of categorizers in large scale categorization. Indeed, XRCE was able to show an improvement of ca. 20% in categorization accuracy over the status at the beginning of the project, on a category system derived from DMOZ[§].

[§] http://dmoz.org/

It explored calibration of categorizers in order to improve the performance of the different categorizations (text-based, image-based, multi-media) to be used and assessed within REVEAL THIS. This work consisted in implementing first versions of calibration engines based on one-dimensional support vector machines, and non-parametric methods. This technique has lead to a ca. 5% gain in categorization accuracy.

To deal with documents in Greek, XRCE built a cross-lingual categorizer. This tool integrated bilingual lexicons built in WP7 into the text-based categorization chain. This has led to a first version of a cross-lingual categorizer for the two domains. The finalized categorizers for politics and travel can take as input either English or Greek documents. the categorization models use category profiles in English. Internal representation of input documents is translated from Greek to English prior to categorization using bilingual lexicons derived from existing dictionaries and parallel corpora.

XRCE conducted an evaluation of its multilingual categorization tool, by assessing whether English and Greek documents were treated "in the same way". The results on a parallel collection (namely EuroParl) indicated strong similar, even though not identical, results for the two languages. Evaluation results are reported in D5.2 in detail.

Furthermore, XRCE designed two methods for cross-media categorization, exploiting the generic visual categorizer developed in WP4. After experimentation with various approaches, the final cross-media categorization prototype was designed to implement the multiple view approach (cf. deliverable D5.2). In this approach, the results of an image categorizer are levered with the outputs of a text categorizer, and vice versa. This is done by exploiting the correlations between image and text labels. Preliminary experiments in this direction showed a significant increase in the results of the image categorizer. The results of the text categorizer were basically unchanged (please refer again to D5.2 for a discussion of these results). More advanced experiments led to an improvement over the initial multiple view approach by deriving new estimation formulas for the correlations between text and image categories. XRCE has also tested this approach on "true" multi-class, multi-label problems (the category system comprises more than two categories, and documents can be assigned to more than one category) by combining the multi-class, single label categorizers with calibration procedures. All these results are described in the deliverable D5.4 and are the subject of a paper[**] presented at ICML 2007.

On the engineering side, XRCE refined the APIs for its categorization systems (text, both mono- and multi-lingual, images and cross-media) so that they can be used smoothly in the overall REVEAL THIS processing chain. All the categorizers are integrated as a workflow which allows one to process automatically the concerned categorizers depending on the type of input. For instance, a web-text document corresponding to news does not involve image and cross-media categorization. For efficiency purposes, the workflow is distributed on a group of machines.

[**] J.-M. Renders, E. Gaussier, C. Goutte, F. Pacull, G. Csurka. Categorization in multiple category system. To appear in Proceedings of the 23rd International Conference on Machine Learning.

3.5 WP 6 Cross-media Summarisation Subsystem (CSS)

Objectives

In this work package, cross-media content analysis results produced in WP4 are further exploited in order to summarize audiovisual files. Hierarchical video content summarization using unified semantic and visual similarity is also explored. The Cross-media summarization subsystem should be linked with the translation subsystem in order to provide multilingual summaries. Multimedia presentations of the produced summaries will be designed and developed allowing the user to quickly digest information and interactively adapt and refine her/his queries within the current context. Recent international standardization endeavours (SMIL, SVG) on multimedia presentation should be adopted.

Status at the beginning of the project

Text summarization (TS) and visual matching know-how was available in ILSP at the project start. The MEAD summarization environment allowing for experimentation on salient feature extraction was also available to the consortium. Prior to the start of the project, KUL had experience with image matching and similarity search, but never studied the problem of visual summarization.

Progress towards objectives

ILSP built a finer-grained version of its text-summarization component for two languages (EN, EL) and two domains (politics/news, travel), which includes fact extraction technology. Furthermore, a fully working cross-media summarization component was developed by ILSP, aiming to combine visual findings (e.g. characteristic scenes and/or important faces extracted from video) with textual ones (summaries of the accompanied transcript along with facts, named entities) and pre-defined templates, which constitute the class of knowledge-intensive approaches that was believed to be more suitable for this task. Translated versions of the textual part of the summaries are accommodated, through interaction with the XRCE translation web-services.

The architecture was designed in a unified and adaptable fashion, bearing in mind that different channels across countries and different genres (TV news, European parliament sessions, travel programmes) have to be accounted for. Additionally, two types of users/audiences were targeted: (a) a generic audience who would like a summarized preview of the current stories (or the stories matching the user's profile), and, (b) specialized audiences (e.g. journalists) who usually seek for specific information, for instance a timeline of statements made by a particular politician.

In the final cross-media summarization prototype, one summary accommodating the needs of both generic and specialized audiences is being generated serving for both generic/light-weight summarization needs and interactivity. The latter, is a distinguishing characteristic of the summaries, which consist of the following layout *areas*:
The *Informative Area* holds generic information concerning the summary, displaying the name of the summarized file, date/time information and the selected category/topic. The *Windows Media Player area* contains the embedding Windows

Media player responsible for the playback of video segments. The *Textual Summary area* displays the summary text in scrolling mode, rolling parallel to the video. The *Index Area* (EbS and Travel Summaries) is the place where all available topics are presented and the user is able to interact and navigate within the summary. Finally the bottom, *Misc. Navigation area* holds characteristic images that, on selection, provide additional navigation paths, depending on the summary domain.

In order to display the context-rich summaries in a user-friendly way, ILSP developed visualization/presentation interfaces responsible for forwarding and displaying the summarised metadata into a suitable viewer installed in the user device. SMIL (Synchronized Multimedia Integration Language) and HTML+TIME (Timed Interactive Multimedia Extensions) markup languages were exploited as they provide a framework for this type of collaboration and presentation of different media that can be properly time coordinated and synchronized.

Summaries for three different domains/genres have been developed: TV news, European Parliament sessions and travel, in two languages (EN & EL) and with translated versions. ***For TV news***, textual summarisers and the scene categorisation module were integrated. In particular:

KUL developed a scene categorisation module, i.e. methods needed in order

1. to group consecutive shots into scenes and select the most representative keyframe for each scene,
2. to cluster scenes into a hierarchical tree structure, based on visual similarity,
3. to select the top N most salient images from a video, based on the hierarchical tree, and to label scenes into a fixed set of domain-dependent categories for the news broadcast scenario (distinguishing between anchor person, interview, reportage, and graphics), based on a belief propagation network.

First, a method was developed to **group video shots into scenes** and represent each scene by a single, representative keyframe or keyshot. Semantically meaningful scenes are often characterized by the use of (possibly interleaved) visually similar video fragments, and this phenomenon is exactly what is captured by our grouping tool. This structures the data, providing a higher level segmentation. At the same time, focussing on the selected keyframes or keyshots for each scene, this brings a drastic reduction in the amount of data (as compared to the original video data, or the whole set of keyframes extracted with the shotcut detection algorithm provided by KUL), while most of the relevant information is kept in the new representation.

Next, the scenes were clustered into a **hierarchical tree structure**, based on visual similarity over the entire video (so not just nearby shots). This again yields information about the structure of the video and can be used in an interactive setting to have the user explore the data or delve into a specific subtopic he may be particularly interested in.

Based on the hierarchical tree representation, a method was developed to automatically select the **top N most representative images** for the video, where N is not fixed but can be changed at runtime. A measure to determine how representative a keyframe is ('saliency') was developed, that takes into account how typical the keyframe is (i.e. many similar frames in the video), how much of the video can be

explained by it (taking into account the duration of the shots) and how complimentary it is to other selected keyframes.

Finally, a **Scene Labelling tool** was developed. This is an extension of the 'Face-based Categorizer' mentioned in the original project proposal. Based on the faces in the image (number of faces and size of faces) as well as more generic scene characteristics such as average shot duration, number of shots, color content of the keyframes, etc. a tool was developed to label each scene into a fixed set of domain-dependent categories (e.g. Anchor person, interview, reportage and graphics for news broadcasts). These scene labels are further improved by taking the temporal dimension into account. This is achieved by learning a belief propagation network, which models the probability of scene transitions.

The cross-media TV news summary presents summaries of what the "anchorman" said in the news story, of the "reportage(s)" in the specific story and of the "interviews" in the story; the scene labeller characterises the scenes of the story as belonging to one of the three categories and the textual summarizer creates the corresponding extract. By default, the user is presented with a summary of all the anchorman utterances (and the corresponding visual scenes) and the user is also given the choice to interact with the summary by choosing to view a summary of the story that is based on the interviews or reportages included in the story.

In the case of ***European Parliament sessions***, most of the time the camera remains static; focused on the face of the specific EP member who has taken the floor. Occasionally, a general long shot of the chamber might be shown. Therefore, a summary is focused on the major topic/issues raised during the story and the different views/argumentation of the political parties involved in the discussion. In this case, the user is presented with a list of all major topics discussed in the session (as determined through term extraction). By selecting a topic, the user can watch a short video that summarizes the different political views expressed in the chamber (a "per-topic" summary). For the selected topic, the list of speakers/MPs who talked about this topic are also shown; by clicking to a speakers image, the user can watch a summary of what the speaker said for the specified topic (a "per-speaker/per-topic" summary).

In the EU politics summaries, visual summarization is restricted to the footage that corresponds to the summarized utterances/text. It is mainly term extraction, speaker and face identification technologies that are being employed for constructing the summary. KUL experimented with variations on its purely vision-based summarization scheme and in particular the scene labelling tool for the travel documentary summaries, using novel image descriptors, using motion cues, using all keyframes rather than just one per shot, etc. Unfortunately, the results were not stable enough.

Therefore, for ***the travel programme cross-media summaries***, ILSP decided to exploit the XRCE image categorisation technology. In particular, detailed media analysis of the travel corpus showed that most of the travel programme scenes fall under three basic thematic categories:

(1) Historical data and archaeological places of interest

(2) Entertainment and lifestyle activities and
(3) Landscape (indoor, outdoor), environment and natural beauties

By clustering the XRCE image categories per keyframe into one of the above thematic categories, one gets a visual-scene segmentation of the file into these three thematic categories. The cross-media summarization subsystem focused on creating three discrete summaries, one for each basic thematic category per geographic location presented in each travel programme. Additionally, in order to enhance the quality of the textual summarizer for the travel documentaries, two human judges manually constructed a list of valuable and distinctive terms out of the corpus through (1) evaluation of all the potential terms produced by the term extraction module and (2) assignment of the validated terms to (one or possibly more) thematic categories. The resultant travel vocabulary terms for each category thus carry enough semantic load to boost the audio segments they appear in. In addition, each such term has an associated weight, analogous to its frequency appearance in the corpus; this weight is used in order to further differentiate the importance of each segment. Details and screenshots of the resulting summaries are available in D6.3 and its final updates.

3.6 WP7 Cross-lingual translation subsystem

Objective

The goal of this WP is to provide basic technologies for handling document content across different languages. In particular, tools for translating queries, extracting bilingual glossaries/lexicons, automatically computing cross-lingual document similarities, and partial document translation should be developed.

Status at the beginning of the project

At the beginning of the project, XRCE had a first version of tools for bilingual lexicon extraction (words and terms). In this version, the extraction of bilingual terms relied on terminology grammars for the two languages of the project (i.e. EN and EL). XRCE also had a first version of a statistical machine translation system working with chunks. This version was lacking: graphical user interface for demo purposes, an optimization procedure for NIST score evaluation and a method to force translation in a limited time. On the resources point of view, neither bilingual lexicons nor machine translation systems were available in the domains and languages of the project (Travel and Politics, EN and EL).

Progress towards objectives

Statistical machine translation
Concerning the statistical machine translation system the first task consisted in developing a first version of an interface for demonstration purposes. Then, machine translation systems for Politics were built. These systems have been built from the European parliament corpus. They aim at translating Greek documents into English, and English documents into Greek. A pre-processing of the corpus has been undertaken, which consisted in corpus cleaning, alignment at the sentence level and lemmatization for bilingual lexicon extraction. Training of statistical machine translation systems (EN->EL, EL->EN) has taken place by running Giza++ in both

directions, extracting the first 200 alignments so as to feed a word alignment matrix which is used as input to the system. This has led to a first library of chunks used both for the translation tool and the bilingual lexicons.

Subsequently, XRCE modified the statistical machine translation tool in order to take into account what it called "elastic chunks". The basic idea underlying the notion of elastic chunks is that the number of gaps (i.e. place holders for words) separating two words of an expression is not fixed. By replacing the rigid (even though non-contiguous) chunks with elastic ones one gains flexibility and generality in translation; this is an innovative approach in machine translation and has been patented by XRCE within the framework of the project.

Evaluation was conducted on a test set derived from the European Parliament corpus. It shows that the EL->EN machine translation system performs well (NIST score of 6.4), and is in the range of the performance obtained with similar systems (French-English language pairs). However, the initial performance for the EN->EL system was much lower, inasmuch as the NIST score only amounts to 4 in this case. This result was surprising and XRCE investigated it further. It examined the bi-chunks automatically extracted from the EuroParl corpus and noticed that many of these bi-chunks were empty on either side (one may refer to these bi-chunks as *orphan* chunks). Basically, the XRCE procedure to assign bi-chunks across languages got rid of some information. Suspecting that this information might be useful for translation (the proportion of orphan chunks was much higher in the English-Greek case than in other cases), XRCE modified the assignment process so as to avoid generating orphan chunks. In particular this forced each monolingual chunk to be assigned to a non-empty chunk in the other language.
The complete data-set was processed again for translation, while getting rid of *orphan* chunks. The new results have shown a significant improvement in the translation accuracy (as measured by the NIST and BLEU scores) from English to Greek. They are summarized in the following tables:

Iteration	**Greek-English**				**English-Greek**			
	NIST		**BLEU**		**NIST**		**BLEU**	
	dev	**test**	**dev**	**test**	**dev**	**test**	**dev**	**Test**
0	5.1484	5.5065	0.1571	0.1953	2.9039	3.7341	0.0846	0.1385
1	5.1380	5.3520	0.1968	0.2231	3.7988	3.8191	0.1203	0.1279
2	5.9552	6.1269	0.2354	0.2466	3.6219	3.6839	0.1140	0.1223
3	5.9798	6.1313	0.2296	0.2929	3.7812	3.8473	0.1236	0.1322

Table 6: Previous results

Iteration	**Greek-English**				**English-Greek**			
	NIST		**BLEU**		**NIST**		**BLEU**	
	dev	**test**	**dev**	**test**	**dev**	**test**	**dev**	**test**
0	5.1775	5.5273	0.1564	0.1995	3.7180	3.8847	0.0878	0.1206
1	5.3104	5.5516	0.2033	0.2274	3.1727	3.1667	0.0887	0.0967
2	5.8285	6.0138	0.2203	0.2356	4.6909	4.7075	0.1409	0.1579
3	5.9259	6.1474	0.2270	0.2447	4.5426	4.4435	0.1422	0.1413

Table 7: New results (without orphan chunks)

Once the results of the machine translation system were satisfying, XRCE translated and delivered to the consortium core data sets for the system evaluation and integration. At the same time, XRCE re-engineered the translation tool so that it can be used in the overall REVEAL THIS processing chain: it designed and hosted a web service that offers translation services to the consortium. This web service is able on the one hand to manage structured documents following formats introduced by the consortium for storing data and meta-data corresponding to videos. On the other hand, the web service is able to manage raw data (plain text, sentences or multi-words expression). Both directions (EL->EN) and (EN->EL) are available.

Bilingual lexicon extraction

With respect to bilingual lexicon extraction, XRCE developed lexicons for both directions of the EN-EL translation pair. These lexicons were used for cross-language information retrieval and cross-lingual categorization purposes. XRCE designed two new methods to extract bilingual lexicons of multi-word units:

The first method is based on the above mentioned “elastic chunks”, and the second is based on a weighted finite state transducer approach. In addition, experimentation with various optimization strategies to derive lexicons with good coverage (recall is here privileged over precision) has taken place, so as to fully exploit such lexicons in cross-language information retrieval and categorization scenarios.

The EL->EN bilingual lexicon automatically extracted from the politics domain has been integrated with another one derived from an existing dictionary for better coverage. This enriched lexicon has been used in particular in the cross-media, cross-lingual categorizer (See WP 5). For the creation of this lexicon, XRCE pursued two paths: asymmetric extraction, where only English candidate terms are known in advance, and symmetric extraction, where both English and Greek candidate terms are known. For asymmetric extraction, it developed a method that searches locally for the best contiguous sequence for each English term. For symmetric extraction, it developed a specific Greek grammar for terms (terms constitute a subset of noun phrases; grammars for noun phrases are thus too general for terminology purposes, and should be refined in order to more precisely focus on terms), and then applied its bilingual extraction tools.

From an experimentation point of view, XRCE implemented and tested the method for asymmetric extraction based on local search with a French-Arabic parallel corpus, tagged with named entities in French, which is available through ELRA. In order to get sufficient lexical information, it relied on English as a pivot language, by first aligning French and English words, then aligning English and Arabic words, and finally mapping French and Arabic words from the above alignments.

Last, XRCE produced and delivered to the partners the two lexicons in the form of translation tables for cross-language information retrieval purposes. The tables have been cleaned removing ill-formed entries (e.g. entries containing non alphabetic characters) and filtering the remaining entries on the basis of the association strength across languages

3.7 WP 8 System Integration

Objective

Based on the design of the system as specified in WP3, this work package integrates all components to give an integrated prototype of the REVEAL THIS system.

Status at the beginning of the project

SAIL LABS contributed a web application framework and the results of a previous project, CIMWOS, to get the project started and accommodate initial output from Speaker Identification, Speech Recognition.

Progress towards objectives

The goal of this work-package is the integration of all software subsystems (developed in WP4-7) into a functional prototype. SAIL set up a flexible component-based framework to provide a common infrastructure and control interface for all subsystems. The final prototype consists of two major components: the Multimedia Indexer and the Media Server:

The **Multimedia Indexer** consists of the components developed in WP4-7 i.e.

- Low-level and medium-specific components:
  - audio processing and speech-to-text,
  - video segmentation and image analysis,
  - text analysis
- Higher level components:
  - cross-media categorisation,
  - cross-media indexing,
  - textual and visual summarisation,
  - cross-lingual translation.

To allow for high flexibility during distributed development by the partners, a loose integration scenario has been adopted, enabling the exchange of metadata information even across different platforms (e.g. Windows vs. Linux). This scenario is shown in figure 5.

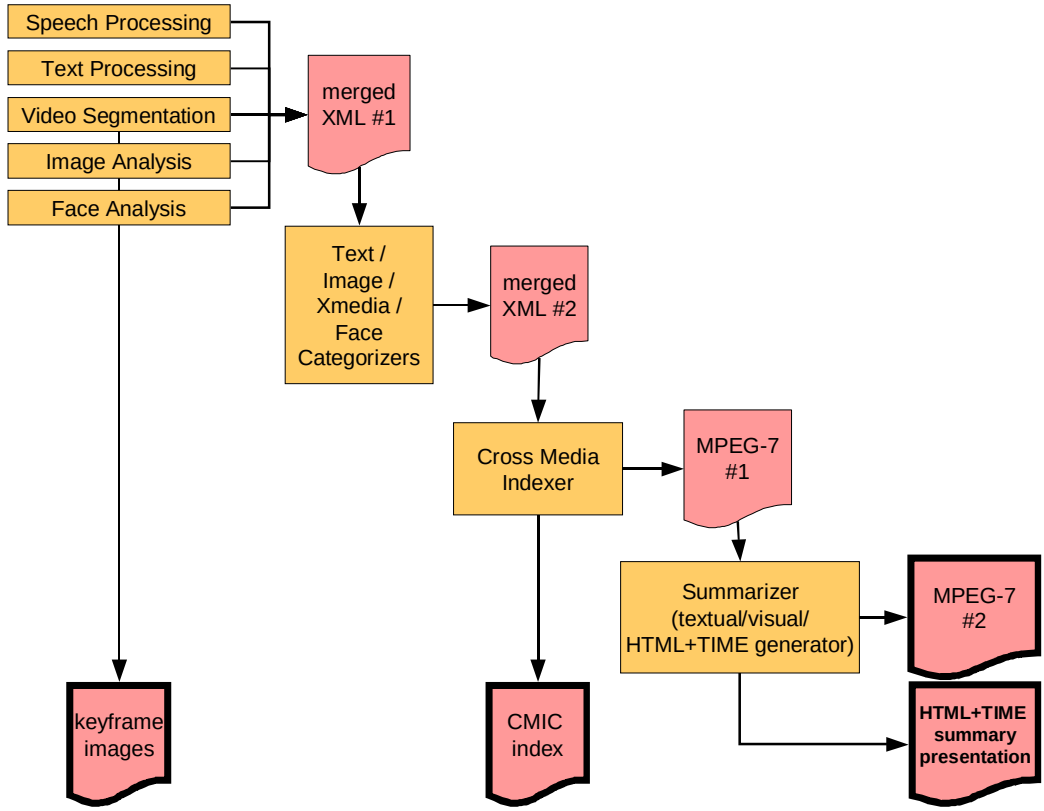

Fig. 7 : Loose integration of indexing components

Each low-level/medium-specific component delivers its results in XML form, and a merging component merges all results into one XML metadata file. This "merged XML #1" contains (among others) the following types of information (assuming a video file as input):

- content words (recognized by speech processing component)
- speaker change times
- named entities, terms, facts (found by text processing of content words)
- shotcut times (identified by video segmentation component)
- for each shotcut, a pointer to a keyframe image
- information about faces found in keyframe images
- segmentation of the file into stories according to topic changes identified by SPC

All these information items are interconnected since each item is associated with an exact time stamp indicating the time offset from the start of the media file.

"Merged XML #1" is then enhanced further by subsequent higher level components, which work on entire stories and add their results into the XML on a per-story basis: the various types of categories, and textual summaries. The final result is in MPEG-7 format. In addition, some extra items are generated which are not stored in the MPEG-7 file:

- keyframe images corresponding to keyframe timestamps mentioned in the metadata

- the CMIC index generated by the Cross-Media Indexer
- multimedia summary presentations in HTML+TIME format, both in original language and translated to an alternate language (by means of the Translation Server developed in WP7, which supports English-to-Greek and Greek-to-English)

All results of the Multimedia Indexer together with the original media file are then uploaded to the Media Server. Media files are uploaded in two formats: in Window Media format (.wmv/.wma) and in 3gp format (targeted at mobile devices).

The **Media Server** stores uploaded XML metadata together with the original media, multimedia summaries and selected keyframe images, and makes its content accessible to users in various forms. It supports searching for content or filtering content based on the metadata. It provides multi-lingual search, multi-media presentation of retrieved content, personalisation, summarisation, and delivery to different devices.

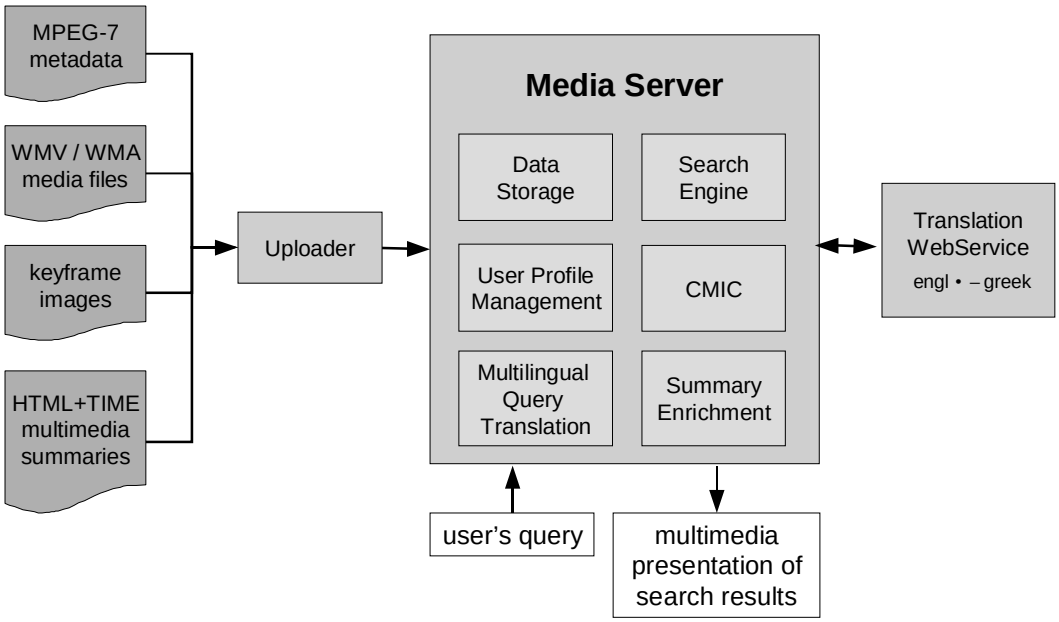

Fig. 8: Components in and around the Media Server

The underlying database system for storing metadata is Oracle Database. Oracle supports full-text searching as well as searching in XML data. Media files, multimedia summaries, and keyframe images are stored in the filesystem.

The Media Server runs on Windows XP or Windows Server 2003. The core of the Media Server is an Enterprise Java Beans application within the JBoss J2EE application server. It communicates with Oracle through the Oracle JDBC driver. The presentation layer is an Apache Tomcat web server. Media files are either in the Windows Media format (.wmv, .wma), played via Windows Media Services, or in the 3gp format, played via Darwin Streaming Server.

Search queries (and filters) are first performed in Oracle, and results are then sorted by relevance by the CMIC module. The CMIC module (Cross Media Indexing

Component) is a .jar file (Java archive) developed in WP4 and is integrated in the Media Server.

Integration of translation and summarization functionalities:

Query translation: Enables users to specify a query (or set up their profile) in their own language and retrieve matching results in other languages as well. This functionality uses translation tables developed in WP7, available for English and Greek..

Story/summary translation: The Media Server can translate stories and textual summaries either during file upload, or "live" upon request from the user while viewing a story. This is done by communicating with the Translation Webserver developed in WP7, which supports English-to-Greek and Greek-to-English.

Multimedia summaries: Multimedia summaries are uploaded to the Media Server already in two languages (English and Greek); they are translated during their creation by the same Translation Webserver mentioned above.

Personalization aspects:
a) User settings:
The Media Server stores various settings for each user and allows the user to change them:
- The user interface language of the web interface
- Whether query translation is to be applied, and from which language
- Possibly complicated queries that the user created and stored for quick access later
- Profiles for use with the mobile user interface

b) Summary Enrichment:
For a given story retrieved as result of a search query or a filter, the Summary Enrichment module creates a personalised summary for the user, based on the standard textual summary stored in the database, but enhanced depending on the query or the user's profile. The Summary Enrichment module is a .jar file (Java archive) integrated in the Media Server. The personalized summary can then also be translated.

User Interfaces:
The Media Server, being an Enterprise Java Beans application combined with a web server, can support various user interfaces. Two different user interfaces have been explored in this project, a classic web interface supporting the "Pull" scenario (for advanced users) and a simpler interface for mobile devices supporting the "Push" scenario.

3.8 WP 9 Assessment and Evaluation

Objectives

The goal of this WP is to carry out a user and task oriented evaluation of the demonstration system implemented in WP8 together with the users and tasks identified in WP2 which form the scenarios of the evaluation.

Progress towards objectives

WP9 consisted of a user-centred and task oriented evaluation of the pilot application, that is of the results of the integration of the different components developed in WP4-WP7 that was carried out in WP8. It should be noted that the evaluation carried out in this WP is concerned with the integrated system and not the single components, the technical evaluation of which was carried out in the corresponding development work-packages. As such, this evaluation took the form of a user and task oriented evaluation, which is deemed the most appropriate for the pilot system. In fact, it has long been recognised that only an evaluation that involves real users in real life tasks can assess the actual effectiveness of an information system. The evaluation relied heavily on the involvement of the user groups that participated in the definition of user requirements in WP2.
As the project developed two different pilot applications, both have been evaluated in a similar way (same tasks and same collection of documents), but with different users and different scenarios in order to check thoroughly all the functionalities of the R-T application.

The design of the evaluation, which was directed by USG but developed in conjunction with BeTV and TVeyes, followed a typical user and task-oriented approach, according to a formative process, so that evaluation data was gathered as part of the design and development process. Therefore, following this methodology, USG decided to carry out two different evaluations on the two pilot applications:

1. *Usability evaluation* of the interfaces;
2. *Usabitlity evaluation* of the pilot application in terms of its effectiveness for accomplishing the tasks for which it was designed.

The usability evaluation of the interfaces was carried out in the form of a focus group including news professionals, usability experts, and laymen. The effectiveness evaluation instead was carried out in the form of a user and task-oriented evaluation, with realistic tasks and a collection of real documents. However, the scope of the project made it difficult to design an evaluation strategy that enables to assess the pilot application in all its aspects and characteristics. In fact, such an evaluation would have too many parameters, as the project addresses:

2 languages (English and Greek)
2 types of users (professionals and laymen)
3 domain areas (politics, news, and travel)
2 types of information access tasks (pull and push)
2 types of interfaces (web and mobile)

The design of an evaluation that would account for all these variables would require a very large number of users and tasks and effort that goes beyond what the project allows for. So, the consortium agreed to simplify the evaluation by considering two distinct evaluation scenarios, a primary and a secondary. Most evaluation effort was directed to the primary evaluation scenario, with the secondary carried out with the dual purpose of a) validation of some of the results obtained from the primary and b) exploration of the validity of some of the results of the primary scenario into a different context.

Here is a description of the two scenarios:

Primary evaluation scenario: this scenario targeted professional users, engaged in pull/retrieval tasks on English and Greek news/politics and travel using the web interface;
Secondary evaluation scenario: this scenario targeted laymen, engaged in push/filtering tasks on English and Greek news/politics and travel using the mobile interface.

Given the above two scenarios and given the requirements of a sufficient amount of gathered data, it was considered as a minimum requirement to have at least 8-10 users for the primary evaluation and 6-8 users for the secondary evaluation. These numbers were also realistic, given the difficulty of having access to professionals.

As for the criteria and evaluation measures used, the evaluation aimed at looking at the overall usability and effectiveness of the integrated REVEAL THIS prototypes and given the style of evaluation, mainly qualitative data were gathered in order to judge overall levels of user satisfaction. In particular, the following general aspects of the system were considered even before engaging users with tasks:

- Ease of use, in terms of how comfortable users felt, and how intuitive it was to interact with the system;
- Ease of interaction in terms of browsing and finding one's way around it;
- Presentation looking at both graphical interface and legibility of text.

As these factors were considered of crucial importance in order to carry out any form of user and task-oriented evaluation, USG decided to assess them before running the full-scale evaluation with users. For this purpose it decided to carry out a small usability evaluation of both the web and the mobile interface in the form of a *focus group*. The focus group was in charge of identifying any possible major problems with the interfaces, so that these could be eliminated before the user and task-oriented evaluation. This would enable to analyse the evaluation results knowing that the interface was "taken out of the equation", since this is often a major factor affecting the evaluation results of interactive systems. The focus group involved professional users, HCI experts, and laymen and was moderated by one of the evaluation designers. The results of these focus groups indicated a number of recommendations that were passed to the respective implementation teams and that could improve the usability of the interfaces.

The main evaluation, being task-based, required the design of appropriate *tasks*. These tasks had to be as close as possible to real life push and pull tasks that professionals and laymen carry out as part of their every day activities. For this purpose, USG designed a questionnaire that users had to fill as part of their daily activity noting down search tasks they carried out. These would help USG in designing tasks that are as realistic as possible, given the constraints of the number of users and data available. The result of this task gathering activity enabled the evaluation team to draw 8 tasks that were used in the evaluation by 14 professional users (8 English speaking and 6 Greek speaking) and 10 laymen (4 English speaking and 6 Greek speaking).

Once users had engaged themselves with the system performing the suggested tasks, few more factors (both qualitative and qualitative) could be measured:

- How easy it was to find information;
- How fast it was compared with expectations and/or previous systems used;
- How satisfactory the overall experience was

In order to get a better understanding of users' point of views all these indicators had to be read in conjunction with users' opinion on how complex the tasks were and how successful they felt they were in accomplishing them.

Questionnaires were designed to gather user opinion on both scenarios (pull and push) and their overall effectiveness. They were designed with a majority of closed questions but also space for free comments on positive and negative features of the system. Both questionnaires aimed at gathering user opinions on the interface and its usability. The questionnaires for the primary evaluation scenario considered retrieval and the way users felt the system supported them in performing retrieval tasks, while the questionnaires for the secondary evaluation scenario considered filtering and the way users felt the system was supporting them in performing filtering tasks. Both questionnaires contained a common initial part aimed at gathering general and non-task specific evaluation data. Both questionnaires were circulated among members of the consortium prior to their use in order to assess their suitability for tasks and the availability of sufficient information in the collection to find potentially relevant documents.

Even with few hiccups due to time constrains and the difficulty of involving busy professionals in the evaluation plan, the evaluation study produced a corpus of useful data to be analysed in order to understand what role could a system like REVEAL THIS play in reality. Users, professionals and laymen, Greek and English speakers, got engaged working with REVEAL THIS on a variety of tasks of different complexity and provided a volume of relevant structured and unstructured feedback in the form of scores, comments and suggestions covering a number of aspects of the system from appearance to functionality, from usability to relevance of results.

Even though users seemed to consistently prefer alternative tools to REVEAL THIS, that is those they were used to before they experience REVEAL THIS, mainly Google, they appreciated the innovative aspects of REVEAL THIS, i.e. its multi-media and multi-lingual components as well as the simplicity and effectiveness of its mobile implementation. On the one hand, the Google effect (i.e. the change in users expectations induced by the Google phenomenon) implied users were quite demanding in terms of results being always available and relevant without considering the limitation of the data collection REVEAL THIS was being tested on. On the other hand, it made perhaps users more open to novelty and more willing to try new features, which worked in favour of R-T as users were quite enthusiastic about its original contributions.

Data analysis has produced a list of features ranging from minor interface changes in terms of look and feel, labels, buttons and icons, to more serious improvements to translation and video retrieval, that need to be addressed in order to make R-T a more usable and more effective tool, but nonetheless has highlighted the fact that ***there is a***

role for such a tool to play in the information world and users are quite ready for it. Something the design team had to bet on when the project started as little was then known in terms of trends and directions in this research area.

3.9 WP 10 Dissemination and Exploitation

The purpose of this work-package is to generate as much publicity about REVEAL THIS as possible both in academic forums and also to potential customers, to ensure that developments within the project are in step with current markets and technological changes and to develop a practical exploitation plan that maximizes the likelihood of an economically successful exploitation of the technology developed during the project.

In the thirty-month duration of the project, TVEyes has continuously watched the rapid market developments in multimedia (image, video and text) market documenting its observations and conclusions in corresponding deliverables (D10.1, D10.2, D10.3). With regard to exploitation and dissemination, a high number of activities have taken place throughout the duration of the project, and others are planned after the official end of the project; Section 5 provides full details. A quantification of most of these activities is presented in table 6.

Dissemination Activities	Totals
Journal papers	7
Conference and workshop papers	19
Posters	6
Demos	3
Patents filed	10
Events Organized	1
Participation in Competitions	1
Open Access Releases	4
Talks	6

Table 6: Quantification of dissemination activities

4 Contributions beyond the state of the art

The following table presents the scientific and technological contributions of the project per Workpackage, indicating the partner responsible for each contribution, the dissemination activities undertaken for each contribution and providing details on what has been contributed and why it goes beyond the state of the art:

WP	Contribution	How it goes beyond the state of the art	Dissemination	Partner
WP2: User Requirements and Data Collection	Market Survey and Technology Watch	Unique review of the rapid developments in the video and image search market, in content aggregators, and in new services related to digital media. Conclusions, examples and connection with related research prototypes, in terms of the new challenges in processing audiovisual data, and the role of REVEAL THIS in this setting were published.	*1) Journal Paper* in JVRB 2) *Conference Paper* in EuroITV 2006	**TVEyes** (responsible for continuous market watch along the duration of the project) and **ILSP** (complementary role in market watch, responsible for the technology watch; responsible for the publications)
WP2: User Requirements and Data Collection	User Requirements Study	Unique survey of user habits, interaction with existing content aggregators and multimedia search services, user expectations and preferences with regard to a new type of multi-functional system like the envisaged REVEAL THIS prototype. Part of this survey that focuses on news aggregators and the related user expectations has been published.	*1) Journal Paper* in the Online Information Review journal 2) Poster *paper* in ECDL 2006	**USG**
WP2: User Requirements and Data Collection	REVEAL THIS corpus	Unique corpus that is multimodal (speech, text, photographs, videos), multi-source (web, TV, radio), multi-lingual (EL, EN and FR - part of it parallel, most of it comparable), and covers multiple domains (EU politics, news, travel). It is also balanced with regard to idiosyncrasies of each medium, i.e. read vs. spontaneous speech, formal vs. colloquial language, face-rich vs. object-rich images etc. The corpus extends to more than 50 hours per language and domain.	*1) Conference paper* in LREC 2006 2) It is planned that the part of the corpus that is free for research purposes (i.e. all European Parliament Data: video recordings of plenary sessions and press-conferences in all EU languages and in different file formats) will be *released to the research community*	**ILSP**

WP	Contribution	How it goes beyond the state of the art	Dissemination	Partner
WP2: User Requirements and Data Collection	Annotated REVEAL THIS corpus	Greek Dependency Treebank (part of textual REVEAL THIS data annotated for part of speech and syntax) - unique for Greek	*1) Released to the research community* for the CONLL 2006 parsing challenge *2) Workshop papers: in RANLP 2005 and in TLT 2005.*	**ILSP**
WP4: Cross-Media Content Analysis and Indexing Sub-system	Automatic Speech Recognition (ASR) for EN and EL, for European Parliament and Travel	This is the first ASR for Greek trained on European Parliament data and the first ASR engine for EN and EL trained for the travel domain.	*1) Demonstration* of the engines in exhibitions such as NAB, IBC, FITEM. 2) *Exploitation* of the engines by TVEyes for its media monitoring and multimedia content access services. 3) *Exploitation* of the engines by Profit and MatrixMedia	**SAIL**
WP4: Cross-Media Content Analysis and Indexing Sub-system	Speaker Identification for EN and EL European Parliament MPs	First attempt to perform speaker identification on native speakers of English or Greek in the European Parliament	*1) Demonstration* of the engines in exhibitions such as NAB, IBC, FITEM. 2) *Research Paper*	**SAIL**
WP4: Cross-Media Content Analysis and Indexing Sub-system	Image Categorizer	The categorizer combines three strengths which are not found simultaneously in other state-of-the-art approaches: It exhibits state-of-the-art performance, It is fast to train and test, new categories can be added incrementally. In the PASCAL Visual Object Challenge (2006), XRCE ranked second institution and the system ranked third (the first two ones being those of QMUL), with only a very narrow margin separating it from the best system.	*1) Participation and distinction* in the PASCAL competition 2) T*alk* at the University of Genova 3) *Conference Paper* in ECCV 2006 and a *workshop paper* in 2006 4) Journal Paper in IEEE Transactions on Pattern Analysis and machine Intelligence 5) Three *Patents Filed*	**XRCE**

WP	Contribution	How it goes beyond the state of the art	Dissemination	Partner
WP4: Cross-Media Content Analysis and Indexing Sub-system	Face Identification	This is a 3DMM-based face recognition system, based on a tool for fitting a 3D morphable model to an imaged face. The fitting process works in a fully automatic manner, including initialization, and is robust to partial occlusions. The latter is a unique feature of the tool.	*1) Conference papers: two* in CVPR 2006, one in AMFG 2005, and one *workshop paper* *2) T*he 3D morphable generic face model will be made *availabe online for research purposes* *3) Exploitation* of the results in the EC-funded project CLASS	**KUL**
WP4: Cross-Media Content Analysis and Indexing Sub-system	Named Entity Recognition (NER) for EL and EN for EU politics and travel	Unique NER for EL for the two domains.	1) Conference paper in LREC 2006	**ILSP**
WP4: Cross-Media Content Analysis and Indexing Sub-system	Fact Extraction for EL for EU politics	Unique fact extraction software for EL; it relies on dependency parsing, exploits a lexical database with semantic roles per predicate and information from a domain-specific ontology that clusters predicates into qualitative categories (positive, negative etc.). Both the lexical database and the domain ontology are also unique for EL and they follow (in format and development methodology) the corresponding resources for EN available in the related literature.	1) Conference paper in LREC 2006	**ILSP**
WP4: Cross-Media Content Analysis and Indexing Sub-system	Cross-Media Indexer	Unique indexer that integrates information coming from different modalities (expressed in the text space) and decides on the piece of information to be used for indexing as the most representative one. The module uses a modified TF*IDF metric and the Dempster-Shafer multiple-evidence approach.	1) *Conference papers*: Axmedis 2006, ECIR 2007 short paper, ECDL 2006 poster paper 2) *Workshop papers*: LREC 2006 REVEAL THIS workshop	**USG**

WP	Contribution	How it goes beyond the state of the art	Dissemination	Partner
WP4: Cross-Media Content Analysis and Indexing Sub-system	Ground-truth data-set with relevance judgements for video retrieval	The data-set and its creation innovates wrt related video retrieval sets (e.g. in TRECVid) in many respects: the query formation process relies on information need scenarios, the queries are free natural language related to information need rather than descriptions of images the user seeks, the retrieval unit is not a shot but a segment that presents a story (i.e. crosses over a number of shots), the relevance judgement is based on a 3-point scale etc.	All documents related to building this ground-truth data set will be *released to the research community*, along with the relevance judgments related to the videos of the European Parliament. In particular, we will release: an overview PPT presentation of the task, its objectives and innovations, the scenarios written for invoking queries, the guidelines given to users for the query formation process, the guidelines given to the human subjects for the relevance judgment process, the list of queries generated, the interface for performing the relevance judgments, the actual judgments and the manually segmented (into stories) videos of part of the data set (the part that is copyright free, i.e. a European Parliament plenary session and a press-conference), as well as general information/statistics on the users/annotators that took part in this task.	**ILSP**

WP	Contribution	How it goes beyond the state of the art	Dissemination	Partner
WP5: Cross-Media Categorization Subsystem	Multilingual Text Categorizer	The categorizer incorporates mechanisms for multilingual categorization (use of category profiles in English and translation of EL documents into English prior to categorisation, for categorizing EL documents with a tool initially trained on EN documents). Unique tool in categorization of EL documents and in the method followed.	1) documented in the *publications* that present the cross-media categorizer - see below 2) *Four patents* related to the textual categorizer have been filed	**XRCE**
WP5: Cross-Media Categorization Subsystem	Cross-Media-Categorizer	Unique module in that it makes use of a "multiple view" approach, leveraging the results of an image categoriser with the outputs of a textual categoriser and vice versa; the approach exploits correlations between image and text labels.	1) *Conference paper* in ICML 2006 and a workshop paper 2) One *Patent filed* *3) One talk* *4) Used in the frame of the french project Infomagic* cap digital grouping industrial, academic and french institution like INA. See http://www.capdigital.com/xwiki/bin/view/Projet/Infomagic (french document)	**XRCE**
WP6: Cross-Media Summarization Subsystem	Text summarization (extract-generation)	This is a unique tool for EL; it is based on the MEAD summarization platform, but has been extended to include terms, named entities and facts in the scoring formula used for selecting sentences to include in the final extract.	detailed in publications related to the cross-media summarization system (see below)	**ILSP**
WP6: Cross-Media Summarization Subsystem	Incremental length summarization using novelty detection	A new method of text summarisation particularly suitable to mobile devices that combined incremental length query-biased summarisation with noverly detection in the context of a "show me more" framework.	1) A *journal paper* (Information Processing and Management) 2) *poster* at ECIR'07	**USG**

WP	Contribution	How it goes beyond the state of the art	Dissemination	Partner
WP6: Cross-Media Summarization Subsystem	Scene clustering and scene labelling for news programmes	This is a tool for hierarchically clustering shots into scenes, selecting the most salient keyframe(s) per scene and labelling the scene according to its contents; the labelling reveals the structure of the news story (i.e. anchor man - interview - reportage). It is a unique visual summarisation tool in that it attempts to reveal the structure of the news story following the clustering, keyframe extraction and labelling steps mentioned earlier.	1) detailed in *publications* related to the cross-media summarization system (see below) 2) *Technology transfer* between KUL and ILSP for the needs of the Greek national project TV++	**KUL**
WP6: Cross-Media Summarization Subsystem	Cross-media Summarization	Combines the visual summarizer and textual summarizer and presents a compact view of audiovisual data for news programmes per news story; for travel programmes it combines the text summarizer and the output of the image categoriser for summaries per geographic location mentioned in the programme; for EU Politics it combines the text summariser, the term extractor and the keyframe extraction output, for summaries per plenary session/press conference. The module interacts with the translation module for providing summaries in both EN and EL for each file. The domains covered and the approach followed for producing multimedia summaries in all cases mentioned above are innovative.	1) Workshop paper in LREC 2006 2) *Talk* 3) Joint (with KUL) *conference and journal papers* planned	**ILSP**

WP	Contribution	How it goes beyond the state of the art	Dissemination	Partner
WP7: Cross-lingual Translation Subsystem	Statistical Machine Translation module	The innovation of the module is that it incorporates methods that deal with "non-contiguous chunks" (e.g. phrasal verbs, negation etc.) introducing the notion of "elastic chunks".	1) *Conference paper* in HLT 2005 2) Two *patents* filed 3) *Used in the frame of the french project Infomagic* cap digital grouping industrial, academic and french institution like INA. See http://www.capdigital.com/xwiki/bin/view/Projet/Infomagic (french document)	**XRCE**
WP8: System Integration	Development of an integrated multimedia access prototype - with a number of functionalities accessible both through the web and through a mobile	Innovation lies in the fact that there is no other prototype that gathers all functionalities that REVEAL THIS has, i.e. categorization, translation, summarization, search and retrieval of audiovisual data, TV/radio programmes and web text.	1) Four *talks* (2 by ILSP, 2 by USG) 2) Two *poster papers* (ILSP) at SAMT 2005 and SAMT2006 3) Print *brochure* and online presentations (ILSP) 4) *Workshop* organised (ILSP) 5) *Demonstrations*: IBC 2006 and planned IBC 2007 (SAIL), SAMT 2006 (ILSP) 6) planned *press release* (ILSP) 7) planned *journal papers* to be submitted 8) *Discussions with potential users* conducted (BeTV, TVEyes)	**SAIL** (responsible for the integration of all modules, for setting up the web application, interacting with the mobile application and design/implementation of interfaces), **USG** responsible for development and integration of the mobile-application and corresponding interface, **ILSP** and all responsible for dissemination of information on the integrated prototype

WP	Contribution	How it goes beyond the state of the art	Dissemination	Partner
WP9: Assessment and Evaluation	Evaluation of a new type of system (REVEAL THIS) through the web and a mobile	Innovation in the object of the evaluation rather than on the methodology followed	1) Journal *paper* on the evaluation methods used in the Online Information Review journal; other publications on the whole evaluation process and analysis planned	**USG** responsible for the design and analysis of the evaluation, **ILSP** and **BeTV** responsible for conducting the evaluation with potential users

5 Dissemination of knowledge

This section includes all dissemination activities undertaken by the consortium throughout the lifecycle of the project.

Journal Papers

- K. Pastra and S. Piperidis (2006),"Video Search: New Challenges in the Pervasive Digital Video Era", Journal of Virtual Reality and Broadcasting, 3(11).
- S. Sweeney and F. Crestani (2006), "Effective research results summary size and device screen size: is there a relationship?", Journal of Information Processing and Management, 42(4):1056-1074.
- Sweeney S., Crestani F., Losada D. (in press), *"Show me more": incremental length summarisation using novelty detection*", Information Processing & Management Journal.
- Chowdhury S., Landoni M., and Gibb F. (2006), "A review of research on digital library evaluation", Online Information Review 30 (6).
- Chowdhury S., Landoni M. (2006), "News aggregator services: user expectations and experience", Online Information Review 30 (2), pp.100-115.

Papers accepted to Journals

- Perronnin F., "*Universal and Adapted Vocabularies for Generic Visual Categorization*", IEEE Transactions on Pattern Analysis and machine Intelligence, Paper accepted with revisions.
- Pastra K., "*COSMOROE: A Cross-Media Relations Framework for Modelling Multimedia Dialectics*", Multimedia Systems, Springer Verlag, Paper accepted with revisions.

Papers in conference/workshop Proceedings

- Csurka G., E. Gaussier, F. Pacull, F. Perronnin, J. Renders (2006), "Image and Multimedia Categorization", in Proceedings of the Workshop on Category-level Object Recognition, Siracusa
- De Smet M., R. Fransens, L. Van Gool (2006), "A generalised EM approach for 3D model based face recognition under occlusions" , in Proceedings of the Computer Vision and Pattern Recognition 2006 conference, New York, USA.
- Fransens R., C. Strecha, L. Van Gool (2006), "Robust Estimation in the Presence of Spatially Coherent Outliers", Proceedings of RANSAC25, New York, 2006.

- Fransens R., C. Strecha, L. Van Gool (2006), "A Mean Field EM-Algorithm for Coherent Occlusion Handling in MAP-Estimation Problems", Proceedings of CVPR, New York

- Georgantopoulos B., T. Goedeme, S. Lounis, H. Papageorgiou, T. Tuytelaars, L. Van Gool (2006), "Cross-media summarization in a retrieval setting", in Proceedings of the LREC 2006 workshop on "Crossing media for improved information access", Genoa, Italy.

- Giouli V., A. Konstandinidis, E. Desypri, X. Papageorgiou (2006), "Multi-domain, Multi-lingual Named Entity Recognition: revisiting and grounding the resources issue", in Proceedings of the 5th Language Resources and Evaluation Conference (LREC) 2006, Genoa, Italy

- Papageorgiou H., E. Desypri, M. Koutsombogera, K. Pouli, P. Prokopidis (2006), "Adding multi-layer semantics to the Greek Dependency Treebank", in Proceedings of the 5th Language Resources and Evaluation Conference (LREC) 2006, Genoa, Italy

- Pastra K. and S. Piperidis (2006), "Crossing Media for Video Search: enabling usability beyond traditional broadcast and TV", in Proceedings of the 4th European Conference on Interactive TV (EuroITV) 2006, Athens, Greece

- Pastra K. (2006), "Beyond Multimedia Integration: corpora and annotations for cross-media decision mechanisms" , in Proceedings of the 5th Language Resources and Evaluation Conference (LREC) 2006, Genoa, Italy

- Perronnin F., C. Dance, G. Csurka and M. Bressan (2006), "Adapted Vocabularies for Generic Visual Categorization", European Conference on Computer Vision, Graz, Austria, 2006

- Renders J., E. Gaussier, C. Goutte, F. Pacull, G. Csurka (2006), "Categorization in multiple category systems", in Proceedings of the 23rd International Conference on Machine Learning (ICML), Pittsburgh, Pennsylvania, USA.

- Riedler J. and Katsikas S. (2007), "Development of a Modern Greek Broadcast-News Corpus and Speech Recognition System", Proceedings of the 16th Nordic Conference of Computational Linguistics NODALIDA-2007, p. 380-383, Tartu, Estonia

- Sweeney S., Crestani F., Losada D. (2006), "Summarisation and novelty in mobile information access", Mobile and Ubiquitous Information Access 2006 Workshop (MUIA'2006), Finland.

- Yakici M., and F. Crestani (2006), "Cross-media Indexing in the Reveal This prototype", in Proceedings of the LREC 2006 workshop on "Crossing media for improved information access", Genoa, Italy.

- Yakici M., Crestani F. (2006), "Design and Implementation of a Cross-Media Indexing System for the Reveal-This System", in Proceedings of the Axmedis 2006 Conference.

- Desipri E., Prokopidis P., Koutsombogera M., Papageorgiou H. and Piperidis S. (2005), ""Towards a Greek Dependency Corpus", in proceedings of RANLP-2005

International Workshop on Language and Speech Infrastructure for Information Access in the Balkan Countries, Borovets.

- Fransens R., Christoph Strecha, Luc Van Gool (2005),"Parametric Stereo for Multi-Pose Face Recognition and 3D-Face Modeling", in Proceedings of AMFG05, Bejing, China.

- Prokopidis P., Desipri E., Koutsombogera M., Papageorgiou H. and Piperidis S. (2005), ""Theoretical and practical issues in the Construction of a Greek Dependency Corpus", in proceedings of the 4th Workshop on Treebanks and Linguistic Theories (TLT-2005), Barcelona.

- Simard, M.,N. Cancedda, B. Cavestro, M. Dymetman, E. Gaussier, C. Goutte, K. Yamada, P. Langlais and A. Mauser (2005), "Translating with Non-contiguous Phrases", In Proceedings of HLT/EMNLP, Vancouver, Canada.

Short-Papers and Posters

- Sweeney S., Crestani F., Losada D. (2007), "Summarisation and Novelty in Mobile Information Access: An experimental investigation", in Proceedings of the European Conference on Information Retrieval, (poster).
- Yakici M. and Crestani F. (2007), "Investigation of the effectiveness of Cross-Media Indexing", in Proceedings of the European Conference on Information Retrieval, (short-paper).
- Chowdhury S. and M. Landoni(2006), "Desired Features of a News Aggregator Service: an End-User Perspective", in Proceedings of the 10th European Conference on Digital Libraries (ECDL), Alicante, Spain (poster paper).

- Piperidis S., X. Papageorgiou, K. Pastra, T. Netousek, E. Gaussier, F. Crestani, T. Tuytelaars (2006), "Multimedia Content Processing and Retrieval in the REVEAL THIS setting", in Poster and Demo paper Proceedings of the first International Conference on Semantic and Digital Media Technologies (SAMT 2006), Athens, Greece (poster also available).

- Yakici M. and F. Crestani (2006), "Design of a cross-media indexing system", in Proceedings of the 10th European Conference on Digital Libraries (ECDL), Alicante, Spain (poster paper).

- Piperidis S. and X. Papageorgiou (2005), "REVEAL-THIS: Retrieval of Multimedia and Multilingual Content for the Home User in an Information Society", In Proceedings of the 2nd European Workshop on the Integration of Knowledge,Semantic and Digital Media Technologies, London, U.K. (poster available too)

Invited Talks

- K. Pastra (June 2007), "REVEAL THIS and the COSMOROE cross-media interaction relations framework", Department of Computer Science, University of Sheffield, U.K.

- H. Papageorgiou (March 2007), "Cross-media Analysis & Summarization of audiovisual content", Cost 2102 project workshop on "The Fundamentals of verbal and non-verbal communication", Vietri sul mare, Italy.
- G, Csurka (October 2006) "Image and Multimedia Categorization", Seminar, October 2006, Department of Computer and Information Sciences at the University of Genova, Italy.
- F. Crestani (July 2005)," Information retrieval and speech: some research at the foundations of mobile information retrieval", International Computer Science Institute in Berkeley, California, USA.
- F. Crestani (April 2005)," Mobile and Ubiquitous information access", University of Hawaii, USA
- S. Piperidis (November 2004) "REVEAL THIS : Retrieval of Video and Language for the Home User in the Information Society", IST Event 2004, 15-17 November 2004, The Hague, The Netherlands

Workshops organised

A half-day workshop entitled "***Crossing media for improved information access***" was held in the framework of the *5th Language Resources and Evaluation Conference* - LREC 2006, on 22-28 May 2006 in Genoa, Italy (http://www.lrec-conf.org/lrec2006) . The workshop was organised in collaboration with the IST PRESTOSPACE-Integrated Project, by Stelios Piperidis (ILSP), Hamish Cunningham and Valentin Tablan (University of Sheffield).

The workshop comprised nine invited talks from widely known researchers in multimedia processing and representatives of a number of related, past and ongoing EC-funded projects:

- "Multimedia Semantic Analysis in the PrestoSpace Project".
Valentin Tablan, Hamish Cunningham, Cristian Ursu

- "Cross-Document Coreference for Cross-media Film Indexing"
Eleftheria Tomadaki, Andrew Salway

- "Cross-media Indexing in the REVEAL-THIS system"
Murat Yakici, Fabio Crestani

- "The iFinder audio-visual indexing framework for cross media applications"
Joachim Koehler

- "Cross Media Aspects in the Areas of Media Monitoring and Content Production"
Herwig Rehatschek, Michael Hausenblas, Georg Thallinger, Werner Haas

- "Representation and Analysis of Multimedia Content: The BOEMIE Proposal"
D.I. Kosmopoulos, V. Karkaletsis, C.D. Spyropoulos

- "X-Media: Large Scale Knowledge Acquisition, Sharing and Reuse across media"
Fabio Ciravegna, Stephen Staab and X-media consortium

- "Cross-media summarisation in a retrieval setting"

Byron Georgantopoulos, Toon Goedeme, Stavros Lounis, Harris Papageorgiou, Tinne Tuytelaars, Luc Van Gool

- "From Media Crossing to Media Mining"

Franciska De Jong

The "Crossing Media for Improved Information Access" workshop explored the new tendencies in accessing multimedia content through cross-media decision mechanisms by bringing together researchers working on the development of indexing technologies for archived and contemporary multimedia content.

Tablan, Cunningham & Ursu presented a method of automatic semantic analysis in the process of creating analytical metadata for digitized audiovisual archives in the PrestoSpace project. Tomadaki & Salway dealt with the resolution of cross-document coreference in an attempt to generate representations of film content out of various texts, such as screenplays, audio descriptions and plot summaries, in order to improve video indexing. Yakici & Crestani presented the cross-media indexing component of the REVEAL THIS project, a component that leverages the individual potential of each indexing information generated by the analyzers of diverse modalities such as speech, text and image. The initial prototype utilises the multiple evidence approach by establishing links among the modality specific descriptions in order to depict topical similarity in the textual space. Koehler described the multimedia indexing system iFinder, a development of the Fraunhofer IMK, and its usage in several research and development projects and applications. The main idea of iFinder is to integrate different multimedia extraction methods for the automatic generation of metadata of audio-visual content and to support international meta data standards, like MPEG-7.

Rehatschek et al discussed cross media tools and multi-modal analysis and their role in automating media monitoring and advancing content production, by presenting relevant results from the DIRECT-INFO and NM2 projects. Kosmopoulos et al. proposed an approach to knowledge acquisition, which uses multimedia ontologies for fused extraction of semantics from multimedia content, and uses the extracted information to evolve the ontologies. Ciravegna & Staab presented the X-Media project which addresses the issue of knowledge management in complex distributed environments, by implementing large scale methodologies and techniques able to support sharing and reuse of knowledge that is distributed across different media (images, documents and data) and repositories (data bases, knowledge bases, document repositories, etc.). Georgantopoulos et al described the cross-media summarization component of the REVEAL-THIS project. They report different ways of synthesizing the most salient elements of the constituent parts of a cross-media object, visual, auditory or textual, and adapting the way in which these salient parts are fused in accordance with the users' interests, digital equipment and the typology and semantic characteristics of the original information. Last, DeJong reviewed how the concept of media crossing has contributed to the advancement of the application domain of information access and explored directions for a future research agenda. She discussed ways to incorporate the concept of medium crossing in a more general approach that not only uses combinations of medium-specific processing, but that also

exploits more abstract medium-independent representations, partly based on the foundational work on statistical language models for information retrieval.

More information on this workshop together with the full papers and presentations is available on www.reveal-this.org

Demonstrations

- Participation of the project with a demonstrator in the IBC2006 Conference and Exhibition, September 7-12, in Amsterdam:

IBC Conference and Exhibition is the biggest and most important event for the media industry in Europe. "IBC is committed to providing the world's best event for everyone involved in the creation, management and delivery of content for the entertainment industry." IBC2006 attracted more than 44.000 visitors according to the organisers' announcement on the IBC website (www.ibc.org). REVEAL THIS participated in IBC2006, hosted at the SAIL booth. The demonstrator consisted in version 3 (pre-final) of the prototype and a specially designed poster accompanied it. REVEAL THIS was well received and interested parties and contacts ranged from media organisations like BBC World Service and EbS, to companies like Thomson, Sysmedia and lots of SMEs operating in the areas of media monitoring, subtitling and other media business.

- Participation of the project with a demonstrator and a poster in the SAMT2006 Conference, December 6-8, in Athens, Greece.

SAMT2006 was the 1st International Conference on Semantic and Digital Media Technologies. It was a very well focused conference attracting some 120 people deeply involved in multimedia research and development, as well as EC officials running related areas of the IST work-programme. REVEAL THIS participated in SAMT2006, with a poster and demo paper. The demonstrator consisted in version 4 (final) of the prototype and a specially designed poster accompanied it. REVEAL THIS was well received and interested parties and contacts ranged from international journal editors like IEEE Multimedia Magazine to researchers, working on similar or adjacent areas, and EC officials.

Competitions

XRCE participated to the **PASCAL Visual Object Challenge 2006**, with its image categorizer. XRCE ranked second institution and the system ranked third (the first two ones being those of QMUL), with only a very narrow margin separating it from the best system. 11 institutions took part in this evaluation: Microsoft Cambridge along with the Cambridge University, the Carnegie Mellon University (CMU), the Helsinki University of Technology (HUT), the Institut National de Recherche en Informatique et Automatique (INRIA), the Leoben University, the Queen Mary University of London (QMUL), the RWTH Aachen University, the Siena University (UNISI), the Amsterdam University (UvA) and the Xerox Research Centre Europe (XRCE). Each institution had the possibility to submit results for several systems. XRCE submitted results only for the REVEAL THIS image categorizer and, in total, 19 systems competed with it.

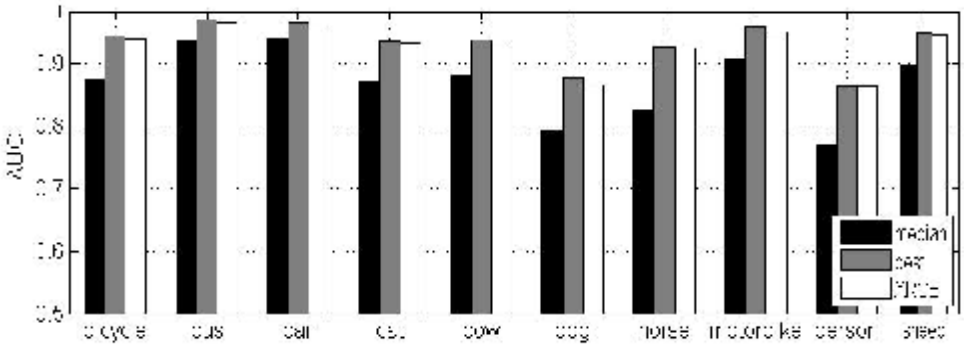

Figure: Per category results on the PASCAL VOC 2006 database.

"Open access" activities

- ILSP made available its Greek annotated data for dependency parsing, to the corresponding CONLL 2007 shared task (Computational Natural Language Learning Conference – Dependency parsing challenge). This was data that were created for the development of the Greek fact extractor in REVEAL THIS.
- ILSP plans to make available to the research community, part of the REVEAL THIS corpus and in particular all video recordings of the European Parliament sessions and the related press conferences. This is data that the EbS channel allows to be recorded and shown in public when the broadcasting channel is mentioned explicitly. Therefore, it will be straightforward for ILSP to release its recordings including: video files in various formats and audio files in various formats for all EU-languages transmitted.
- ILSP plans to make available to the research community, part of the ground-truth data-set with relevance judgements for video retrieval, that was created in the project; all documents related to building this ground-truth data set will be released to the research community, along with the relevance judgements related to the videos of the European Parliament. In particular, we will release: an overview PPT presentation of the task, its objectives and innovations, the scenarios written for invoking queries, the guidelines given to users for the query formation process, the guidelines given to the human subjects for the relevance judgement process, the list of queries generated, the interface for performing the relevance judgements, the actual judgements and the manually segmented (into stories) videos of part of the data set (the part that is copyright free, i.e. a European Parliament plenary session and a press-conference), as well as general information/statistics/demographics on the users/annotators that took part in this task.
- KUL cleaned up and refined its 3D morphable generic face model, and will soon make it available for research purposes online. A dedicated website will describe the model and how to use it. This will allow other researchers to build on the Reveal-This work and develop novel face processing algorithms without the need to repeat the labour-intensive process of collecting and aligning data, removing illumination effects, etc.

For more information, please keep visiting www.reveal-this.org .