Final Report Summary - CAPER (Collaborative information, Acquisition, Processing, Exploitation and Reporting for the prevention of organised crime)

Executive Summary:
CAPER project had 40 months duration and started in July 2011. The consortium was constituted of 17 partners from six different countries: Spain, Germany, Italy, Portugal, France and Israel.

The project was organized in nine individual work packages, out of which seven are related to research activity, one related to dissemination and exploitation activities and one related to management.

CAPER project envisaged to create a common platform for the prevention of organized crime through sharing, exploitation and analysis of open and private information sources. It was aimed to develop a strategic concept of organized crime that “includes measures designed for improving knowledge of the phenomenon and for strengthening prevention, investigation and cooperation” (Priority no. 8 of the European instruments in the field of Freedom, Security and Justice of The Hague Programme (COM/2005/0184)).

The technical work packages (WP), namely from WP2 to WP8, were focused on the creation of CAPER platform, WP7 handled the Ethical Issues, WP9 defined the exploitation of the results and disseminated the project outcomes and WP1 focused on project management.

CAPER platform captures a wide range of information types from open sources, analyzing and presenting it by a set of visualization tools. Specifically, the platform processes all kinds of information in different formats (texts, images, audios and videos) gathered from open sources, such as the Web and social media networks. It aims to produce a knowledge repository that can be exploited by a set of graphic tools.

In support to the technology development, the ethical analysis contributed to the creation and implementation of a regulatory model which is a new, practical and plausible ethical and legal framework for security projects and for users’ data protection.

The dissemination effort has been focused on reaching LEAs, members of the scientific community, EU institutions and possible end-users. With regard to exploitation, the partners have tackled the matter from all the possible angles: (i) exploitation of the single Foreground produced by each partner; and (ii) exploitation of the platform, that is the result of the combination of all the single Foregrounds produced, and that represents a Foreground itself.

Thanks to the legal and ethical procedures followed in the project, CAPER platform can foster the general trust in the activities of governments since it allows a safe storing and a controlled use of the information gathered.

Project Context and Objectives:
Context
Law Enforcement Agencies (LEAs) are increasingly more reliant on information and communication technologies and affected by a society shaped by the Internet and Social Media. The richness and quantity of information available from open sources, if properly gathered and processed, can provide valuable intelligence and help drawing inference from existing closed source intelligence.

The revolution in information technology is already in our lives and is making open sources more accessible, ubiquitous, and valuable. LEAs have seen open sources grow increasingly in recent years and most valuable intelligence information is often hidden in files which are neither structured nor classified. The process of accessing all these raw data, heterogeneous in terms of source, format and language, and transforming them into information is therefore strongly linked to multi-modal and multi-lingual data analysis and Visual Analytics technologies with powerful Human Computer Interfaces.

LEAs do use Open Sources in investigations to identify those who have the intent and capability to commit crime. These sources are however exploited by manual means with little automation. Tools such as search engines, meta engines, Internet browser add-ons are used to improve their process but state of the art analysis and information exploitation techniques are not being correctly applied. For this objective, the consortium had exploratory meetings with the participating LEAs for defining the requirements to be met by any system such as the one proposed in CAPER:

1. Information sharing and exploitation that integrates Open Source Intelligence (OSI) and handles structured and unstructured information.
2. Establish best practices for the use of ICT in criminal investigation and preventative indicators.
3. Homogenization and exploitation of information standards.
4. Integration of semantic technologies to both improve Open and Close Intelligence sources in multilingual environments and to improve cooperation and information sharing.
5. Ability to handle information from audio and video sources.
6. Compliance with current national and European legislation with a particular emphasis on data protection and privacy.
7. No dependencies on the agent that has introduced information into the system or the results that it produces.
8. Capability to generate a query or monitoring request in one language and have it performed over sources in several languages.

Objectives
The main goal of the project was to deliver to the LEAs an OSINT (Open Source INelligence) platform which prevents the organized crime leveraging. CAPER platform captures the widest range of information types and open sources, analyzing and exploiting it by a set of visualization tools. Specifically, the platform processes all kind of media (texts, images, audios and videos) gathered from open sources, such as the Web and social media networks. It aims to produce a knowledge repository that can be exploited by a set of graphic tools.

CAPER is a platform that can potentially handle big amounts of data. As a consequence, user privacy may be compromised. In order to counteract this risk, CAPER has implemented a regulatory model which is a new, practical and plausible ethical and legal framework for security projects and for users data protection.

The CAPER project has therefore been prepared with the following core aims:

User Defined/Focused Project: The project was structured specifically towards the needs of the participating Law Enforcement Agencies. The goal was to allow the LEAs analyze and define their requirements, and allow the participating LEAs perform integration tests and field trials of the CAPER platform in partnership with the wider consortium.

Cross and Multilingual Analysis: The platform aimed to integrate 14 languages to allow the LEAs analysis search in the widest range of information. The inclusion of these languages together with the semantic ontologies developed within the project
increased the accuracy of the results.

Video, Image, Biometrics, Speech and Audio Analysis Integration: Each analysis module was geared towards a specific content type, i.e. Text, Image, Video, Audio and Speech or biometric data. In addition, modules interact with the ‘Semantic.

Combine Open and Closed information sources: Documents in different formats, TV, Radio, and Information in closed legacy systems and in the open sources are the data sources to be mined and evaluated by CAPER. In addition to general Internet data sources, CAPER integrated mass media, internal LEA information systems and access to Semantic Web data collections.

Ethical and legal framework: Providing a legal and regulatory framework ensuring that the deployment is ethically and legally coherent. Other important objective regarding this point was to establish protocols and guidelines to ensure that the technological aspects of CAPER advance in line with necessary recommendations.

Active and real participation of LEAs: As an end-user defined and focused project, the participation of the LEAs was one of the priorities. The consortium organized meetings to carry out this objective and also encouraged LEAs to organize meetings between them.

Project Results:
The results of the CAPER project are various and, in order to understand them better, can be classified in technical and non-technical. Starting from the technical results, it is noticeable the fact that the CAPER platform integrates a large number of high demand technologies that the LEAs, governments and other similar organizations are researching. The CAPER consortium worked hard in defining a common and suitable architecture to integrate all of the components involved in the process of CAPER.

The CAPER platform is able to collect big amounts of data from publicly available sources analyzing these data, applying a semantic analysis and then transforming this data to be visualized by information analysts. These modules will be detailed in the following WP descriptions.

Regarding non-technical results, the project created a notable atmosphere in which technical people and LEAs were working in the same direction, organizing workshops, meeting between LEAs, etc. In these workshops LEAs transmitted their expertise and knowledge to the rest of the consortium and generated a fruitful networking with the results of a very good cooperation during the project.

Other non-technical aspects are the creation of a common regulatory model for ethical and legal issues raised on the project. Due to the fact that CAPER deals with big amounts of data also referring to people, a common framework was developed to cover it.

WP2 – Architecture and User Requirements Modeling

WP2 main goal is to clearly define the project requirements based on end-user functional specifications. In order to achieve this objective, other partial objectives must be met:

• Detailed definition of the end users requirements: collect all LEAs requirements proposing different ways to gather this information such as workshops, forms, video conferences, etc.
• Definition of the technical requirements: once gathered the minimal and desired functional requirements a further step should be carried out. The definition of the technical specifications and requirements for the functional requirements.
• Definition and implementation of the best suitable architecture for the CAPER system: study different architecture approaches, chose the best one for the system and design some tests to verify that the selected architecture is correct for the
platform.
• Definition of middleware services: the core of the platform is the layer that supports all the services involved in the system. A correct definition and design is important for the construction of the platform.
• Definition of the development lifecycle: decide and argument whether the selected development lifecycle is appropriate for the CAPER project.

End User Requirements
In order to define the best architecture for the system and gather appropriate user requirements, several peer-to-peer meetings have been held with LEAs with the aim of collecting user scenarios, functional requirements and non-functional requirements (architectural, interoperability, integration, performance, quality, security and legal requirements).

It has also been collected from LEAs the project acceptance criteria, exclusions, constraints and assumptions. For this objective we organized a LEAs Workshop in Pisa on October 5th 2011 with the following objectives:
• Give the opportunity to the LEAs to express their needs
• Share requirements between LEAs
• Define a user requirements road-map.

During March 2012 other meetings were organized together with LEAs with the aim to show different Visual Analytics tools that CAPER could integrate. A feedback from LEAs was obtained. Finally, a Web form was made with the objective to collect additional user requirements information and LEAs filled it. With all the information collected by different channels, a Project Scope Statement document was created, reviewed, evolved and approved.

System and Software Requirements and System Design
In order to select the best architectural approach, the S21sec team researched and analyzed a set of technologies:
- Open Pipeline http://www.openpipeline.com/ a multimedia data processing application integration infrastructure. We made a proof of concept using this tool focused on CAPER needs and an evaluation of the results
- Synthema’s architecture vision, based on the experience of Synthema in similar projects, adapted to KAF annotation format and exposed and discussed with the rest of partners
- Weblab – Petals http://weblab-project.org/ a multimedia data processing application integration infrastructure based on Petals Enterprise Service Bus. We made a proof of concept using WebLab platform focused on CAPER needs and an evaluation of the results. We also made a research of the best ways to integrate this platform with KAF annotation format.
- Virtuoso framework http://www.virtuoso.eu/. In collaboration with Virtuoso project team, we shared information about our projects. The possibility of integrating Virtuoso framework into CAPER system was studied. Collaboration between both teams has been cordial, fluid and successful.
- Talend ESB. In collaboration with Altic, S21SECstudiedTalend tools with the objective to find the ESB that better fits with CAPER technological needs.

In order to finally choose the most suitable architecture for the platform the following topics were approached: results of studies of different approaches, LEA’s physical architecture needs and the introduction to Virtuoso approach, in which Virtuoso representatives showed all CAPER partners a technical vision of their framework

With all the information collected, technical tools were analyzed and according with the projects cope defined an architecture document was created and approved.

Plan Development Lifecycle
CAPER lifecycle is based on PMBOK (Project Management Body Of Knowledge) lifecycle specification edited by PMI (Project Management Institute – http://www.pmi.org).
The lifecycle of the CAPER project was divided into the processes that were managed by the project team to achieve one or more deliverables: knowledge areas and Caper processes related to initiating phase, planning phase, executing phase, monitoring and controlling phase and closing phase.

Regarding the approach of development, it was proposed to be an iterative development that facilitated the user to validate progressively the system together with other proved advantages such as:

- Allow an early and progressively evaluation of components and systems capabilities
- Early visibility to give LEA’s users an idea of what the final system looks like and supports the validation of the system through experimentations and demonstrations
- Encourage active participation among final users and technical partners
- Early risk discovery and mitigation. Helps to refine the potential risks associated with the delivery of the system being developed
- Better progress tracking and predictability

Conclusions for WP2
WP2 is a work package that establishes the bases of the overall system. The work performed in the tasks of this work package together with the overall view of the results of the platform show that the work was successful. At the very beginning of the project we sat down with the LEAs and defined the requirements and main functionalities. Then, with the architecture proposed and the design of the common interfaces all the acquisition, analysis and exploitation modules started to work actively to meet the specifications.

Although the lifecycle proposed has some things to be improved it has been proved that has been beneficiary to the project. We have seen that, regarding the technical phase, an agile approach would have been beneficial because the development can react earlier to changes in the specifications. Thanks LEAs tests and feedback protocol the technical team has been able to receive properly the end-users feedback and apply corrective measures.

WP3 – Interoperability and Management Application

The main goal of this work package was to create the Caper Management Application which is the central point of supervision, configuration and management for the overall system. Other objectives have also been defined:

- Define the interfaces to be used to transfer information from and to LEAs legacy systems and the CAPER platform.
- Develop a secure system by applying the recommendations provided by S21sec in relation to security aspects of software development.
- Describe and define how the final user interacts with the system and how the user input data will be processed in a semantic way in order to retrieve the best information from the CAPER system.

Interfaces Sets
In the first meetings of the work package Syntheema, VI, Technion and S21sec were assigned to contact their closest LEAs (PCPS, P, MOPS, PGME and GC) in order to prepare a document with the information of their legacy systems and their vision about the integration of CAPER and their systems. Those meetings were held with every LEA and finally in after six months after the beginning of the project it was discussed about releasing a document with the conclusions. A release of an entities’ data-model was proposed to all the LEAs and the result was a draft for the D3.1 deliverable “LEA Database API”.

During the second semester of 2012, a final entity data model was proposed, which was then described in the D3.1 deliverable. The second part of the D3.1 deliverable was the definition of a set of Talend Jobs which extract the information from the closed data sources, the LEA legacy systems.
Central Management Application (CMA)
This task consisted in the implementation of a central application that manages and controls the Caper system. The CMA was already running at the beginning of 2012 with some graphical user interfaces such as the definition of new research lines, user management and configurations. Several meetings were held with Synthema and Ikusi in order to integrate their crawling systems with the CMA.

Similar to the versions of the CAPER platform four versions of the CMA were released. The first two versions already had key functionalities such as definition of research lines and all the processes related to the access control management. Version 2 of the CMA already called the Orchestrator but the performance was not good enough. In successive versions the main work dealt with the improvement of this performance once all modules were already integrated within the platform.

Version 3 and 4 included other functionalities such as the definition of reference image and audio sets, services status monitoring, link with Visual Analytics tools, etc.

Secure systems procedures
From the beginning of the project, S21Sec was already thinking about the secure procedures to implement when building web applications. In this context, experts in security issues were consulted and several drafts were delivered to the rest of the CAPER participants.

Security requirements for CMA were implemented and integrated into the CAPER Platform. Access control, authentication, authorization and user/roles management have been included into the scope of first version, but audits have been postponed for the V3 due to deviations in the first two versions.

A security audit was successfully passed due to some problems found inside the servers of the CAPER Platform. This audit raised some issues that had to be fixed prioritized according to dangerousness. These issues were then solved following S21sec recommendations and the audit was successfully passed.

Semantic interoperability
The goal to enrich the CAPER platform with organized crimes ontologies was reached. Several meetings took place to discuss the definition of ontologies for CAPER - both the methodology of the process and the selected fields of crimes to be developed have been discussed in detail. The multilingual aspect and the highly dynamic ontology (organized crime changes terms frequently to outsmart LEAs) are a major challenge to the ontology development.

The four main uses of the ontology were clearly defined:
• Help define the crawl related to a specific area of crime in the Console UI
• Support content (text) analysis (e.g. suspicious terms)
• Support improved VA results
• Support ethical and legal regulatory requirements.

During the last year of the CAPER project the Multilingual Caper Ontology (MCO) tool was developed. The tool enables LEAs either local or global updating of the ontology, according to authorization and specification. The tool enables ontology development by further detailing the scope or expanding the ontology by adding additional synonyms and slang terms to each node. The ontology has been incorporated in the CAPER platform and enables the analysts to search for specific crime phrases (specifying only the general term in local language, but in fact enabling the multilingual, synonym and slang term phrases as well). It has also been incorporated in the text analysis and SAM modules to enhance the collected knowledge.

Conclusions for WP3
Most of the work done in the work package has been carried out by the team assigned to define and build the CMA (Central
Management Application). The reason is that the CMA should connect to the orchestrator and to the repositories to gather the information of the configuration for a better definition of the research lines.

Besides from this work, the definition of the interfaces for the communication between CAPER and LEAs legacy systems has also been an important task in this WP. As not all the LEAs have the same legacy systems, a hard work has been made to create a single and robust way to import and export data from and to CAPER.

Regarding the semantic interoperability, an organized crime ontology has been created and integrated to the system with the corresponding improvement of the results in the platform.

WP4 - Data Acquisition

Data acquisition main objectives were to develop and tune tools and applications related to the acquisition and preprocessing of raw information, and its subsequent collection and facilitate the integration between information coming from open sources and users’ legacy datasources, from different media and in different languages.

The Multimedia Capture Platform was delivered. The main elements of the platform are capture boxes, Linux based systems capable of capturing several A/V streams. The main components of this CAPER Content Acquisition System (CAS) are as follows:

- **Capture Box**: the Box is capable of capturing analogue and digital TV channels. The Box breaks the overall stream into a sequence of small clips, and video is saved in a normalized format (MPEG-4).
- **CAS application**: This is the user interface for monitoring and controlling the capture elements of the platform. CAS application is integrated within Nagios, an open source computer-system monitor, network monitoring and infrastructure monitoring software application. In the final CAPER system this application is used to monitor the status of components and processes that are part of the multimedia platform.
- **Storage unit**: It has been selected a NAS server to provide the necessary flexibility and scalability required. The Storage unit is a FTP server that contains the recorded streams in a normalized format. This system can be centralized or distributed using one or more FTP servers.

The Open Source Capture Platform was designed and implemented either as a standalone system (for CAPER prototype TIGER) and as a web service for CAPER platforms 1, 2 and 3.

The main components of the CAPER Open Source Capture Platform are as follows:

1. **Crawler module** for collecting Internet and Social Web sources. It collects documents, adding them to the document base;
2. **Conversion module** for converting textual sources to standard open formats, so that the analysis module can process the content;

We released 2 versions for the CAPER prototype, TIGER v1.0 and v2.0. Then, for CAPER platform 1, we designed web service interfaces and integrated:

- new features in order to extract from crawled data multimedia data, like images.
- new features in order to extract from crawled data multimedia data, like audio and videos.

For the final CAPER platform 3, we designed web service interfaces and integrated:

- new features in order to crawl web pages linked to the seed url only if they contain a specified keyword
- new features in order to upload into the CAPER Control and Monitor Application (CMA) files already stored somewhere, by specifying the file system.
The final Open Source Capture Platform was successfully integrated into CAPER Control and Monitor Application (CMA), and is able to extract and process different kind of multimedia content detected into the web page, namely text, images, audio and videos.

The Social Media Capture (SMC) was delivered. SMC is able to crawl data from social networks, namely Facebook and LinkedIn. The SMC Facebook crawling module uses the Graph Application Programming Interface (API) and the Facebook Query Language FQL, provided by Facebook itself, and performs three tasks:

- Crawling of single Pages/Groups/Events/Users.
- Search and crawling of Pages/Groups/Events based on specified key-words.
- Search of public posts content and crawling of the related Pages.

The SMC was implemented either as a standalone system (for CAPER prototype TIGER) and as a web service for CAPER platforms 1, 2 and 3.

Conclusions for WP4

As a conclusion, the main results of WP4 are the capture and crawling web services for the CAPER system, able to manage mass media data, multimedia data (text audio images and video) and social network data.

WP5 -Information Analysis

Data consists on facts which can be something simple and seemingly random and useless until it is organized, processed and interpreted, in a given context, to produce valuable information. Open-source Intelligence (OSINT) solutions collect publicly available data from open sources (e.g. the Web, Social Media networks, TV, radio) and then they store it in normalized formats that are ready to be processed by the system. It is at this stage that data is ready to be analysed. But the analysis of the data can take many different forms depending on what information is aimed to be generated. The same data can be analysed in different ways and therefore different information can be obtained from the same data set. The results produced by a contextualized (focused) analysis of a set of data produces focused knowledge. In the CAPER project an OSINT platform has been built with the aim of producing valuable knowledge for the prevention of organized crime from publicly available data collected from open as well as Law Enforcement Agencies’ private sources.

In the context of the CAPER project we refer to information analysis to the contextualized analysis with a given purpose of the collected multimedia content so that information about it can be extracted and we call the result of this contextualized analysis “knowledge”. The CAPER project has built an OSINT solution concept oriented to the prevention of organised crime. In the CAPER platform, once the end-user has configured a research line through the CAPER Management Application (CMA) and the orchestrator has launched the appropriate data acquisition modules, crawled data will start to be ready to be processed by CAPER information analysis modules. CAPER includes the following six analysis modules: (1) Image analysis, basically comparing crawled images and video frames with a set of reference images and/or classes of objects (images); (2) Multilingual text analysis in 14 different languages, identifying entities and relationships among them in the content crawled for each research line; (3) Multilingual analysis of audio content so that it can be reduced to its base components for deeper analysis (i.e. text transcripts of voice); (4) Analysis of videos so that it can be reduced to its base components for deeper analysis (i.e. frames, audio); (5) Integration of semantic-Web technologies and data to improve and relate analysis results (e.g. in the Named Entity Recognition process) and analysis of data coming from Social Media; (6) Biometric face recognition and speaker identification.

CAPER includes the following six analysis modules:

- Image analysis, which compares crawled images and video frames with a set of reference images and/or classes of objects (images) and provides a similarity score.
- Multilingual text analysis, which cover 14 different languages including Arabic, Basque, Catalan, Chinese, Hebrew, English,
French, German, Italian, Japanese, Portuguese, Romanian, Russian and Spanish; and which uses natural language processing techniques to identify entities and relationships among them.

- Multilingual audio analysis, which reduces audio files to their base components for deeper analysis (i.e. text transcripts of voice, speaker recognition and tracking, gender and age identification).
- Video analysis, which reduces video files to their base components for deeper analysis (i.e. scenes, frames, audio).
- Social Media Analysis, which integrates semantic-Web technologies and data (e.g. in the Named Entity Recognition process) and which analyses data coming from Facebook.
- Biometric module, which includes two sub-components: face recognition and speaker identification.

5.1 – Image analysis

The objective of the image analysis module in the CAPER project is to provide the ability to detect a defined image or a set of images in an acquired content. The acquired content is the data extracted from open sources (OSINT) or internal data sources.

The module has two main functions that provide similarity scores:
1. Image comparison function that assigns the numerical visual similarity value between two pictures or a list of pictures,
2. 1-class classification function that gives a yes/no decision whether the image refers to a class previously specified by an analyst.

In the first phase of the project, the work was done in a study and by a comparison of algorithms for detection of points of interest and their descriptor algorithms. We tested available implementations of these algorithms under the same scenarios and testing data. We defined and developed the architecture solution with a PostgreSQL database, which had to be revised during the project due to data protection issues. The second part of the project was the development and the implementation of the appropriate architecture of comparison and 1-class classification functionalities. These functionalities together with configuration abilities have been implemented into the CAPER platforms via web services.

The comparison of images is one of the strong tools that can give information to the end-user. For instance if a LEA has a set of relevant images with content of interest, the image comparison functionality in the CAPER system together with specially obtained knowledge can be connected in the CAPER system and thus deliver additional information to the analytics tool. This may lead to connections that are not obvious, difficult or time-consuming to detect. The advantage of the 1-class classification in CAPER is the versatility of the image analysis module, since it gives flexibility to end users to train the system with all possible different classes.

5.2 – Multilingual Analysis

The main goal of multilingual analysis was to develop and integrate fourteen language processors able to analyse text and highlight the main Named Entities and their relationships.

As a result of the aforementioned tasks, a whole linguistic pipeline for fourteen languages (English, Spanish, Catalan, Basque, French, Italian, Portuguese, German, Romanian, Japanese, Chinese, Arabic, Hebrew and Russian) has been developed as web services providing CAPER of multilingual linguistic analysis capabilities.

The partners contributing to this task have had different responsibilities. Vicomtech has been in charge of developing the pipeline for English, Spanish, Catalan, Basque, French, Arabic and Russian, as well as the whole management architecture. Synthema has been focused on Italian, German, Rumanian, Chinese and Japanese pipelines. Voice Interaction integrated Portuguese linguistic analyser and Technion worked on Hebrew and Arabic.

Regarding the management and architecture of the whole module, a set of extra functionalities were developed in order to be able to be integrated in CAPER framework. These functionalities were developed inside a component named Top Level analyser. Top Level analyser is the main web service that is integrated in the CAPER platform. This web service is the responsible for the detection of the language identification of the texts coming along and of sending them to the web service in charge of processing the language detected. Moreover this web service manages the input and the output from the
multilingual analysis and validates the output to make sure that is with the correct format (in this case, KAF). Additionally, there is a new layer created in these last ten months, that uses MCO (Multilingual Caper Ontology) in order to see if there are domain terms in the documents and add them in this layer to be processed in WP6 (Visual Analytics). Finally there is a post-processing that adds a layer to all the KAF files called entity-relation and which gives back weighted relations based on linguistics features for entities useful for WP6.

The main results regarding multilingual analysis have been the development and integration of an architecture able to manage requests for analysing texts. Those texts are analysed and returned for fourteen language processors. Those language processors analyse text and highlight the main named entities and their relationships as a result of a set of tasks. Additionally an ontology is applied and domain terms are included in the delivered results. The results are showed in a rich xml-based format, named KAF.

CAPER is able to visualize relevant entities through Visual Analytics interface making use of linguistic analysis of WP5.2. In this way, an important functionality is integrated and usable for end-users in order to be able to quickly obtain a quick summary of a certain research line.

Linguistic analysis is available in fourteen languages.

5.3 - Audio and Speech Analysis
In a project where the focus is related to the development of technologies to the prevention of organized crime, the use of audio information extraction plays an important role. Our goal is to develop different tools that help LEAs to enter inside video and audio contents and transform that content in information, to be searched and retrieved.

Our tools intend to fulfill all the chain of information extraction. We started by audio segmentation and classification, language ID of the segments, speaker clustering and tracking and automatic speech recognition. At this stage, the information generated is post processed by text analysis modules that generate a full linguistic characterization. Also spoken terms detection techniques were developed in order to facilitate the search and discovery of information.

Audio segmentation and classification
Audio pre-processing algorithms were developed in order to segment the audio signals into manageable chunks and classify them in terms of their acoustic features.

This module is composed of three components: audio segmentation based on acoustic change detection, audio classification of speech/non-speech (SNS) and background (clean/noise/music), and speaker classification. All classifiers share the same architecture: an MLP with 9 input context frames of 26 coefficients (12th order PLP coefficients plus deltas), two hidden layers with 250 sigmoidal units each and the appropriate number of softmax output units (one for each class).

Speaker clustering and tracking
A speaker clustering and classification system was developed using hierarchical agglomerative clustering techniques to associate all the segments spoken by the same speaker together, helpful for searching and tracking speakers along the audio streams.

Our algorithm combines a BIC based segmentation before the BIC clustering to overcome this problem. It starts by detecting speaker turns using BIC, where change points are detected through generalized likelihood ratio (GLR), using Gaussians with full covariance matrices. SNS segments are also modelled with Full Gaussian and compared with the current speaker Full Gaussian. Based on BIC score the algorithm decides between the merging of the current SNS Full Gaussian and the current speaker Full Gaussian or creating a new speaker Full Gaussian. When creating a new cluster a hierarchical clustering is performed for the previous speaker cluster. In our hierarchical clustering algorithm, the current speaker cluster, provided by turn detection and modelled with Full Gaussian, is compared with the clusters obtained so far.
Language ID
We developed a module able to identify the language of a spoken piece of audio. This Language ID module was based on Total Variability, which has emerged as one of the most powerful approaches to this problem and in similar way also applied to Speaker Identification. We use very similar approaches to both problems.

This technique jointly models speaker and channel variability as a single low rank space. Our language identification component uses the low-dimensionality total variability factors (i-vector) produced by the Total Variability technique to model known languages. This component works after the speaker clustering. Since this component works on-line, every time an unseen speaker starts talking, the component is incapable of knowing the speaker language immediately. To overcome this problem, we produce a first estimate for language identity (if it is a known speaker) after 10 seconds of speech and a final identity estimation after 30 seconds. Since the zero and first-order sufficient statistics (and the respective i-vectors) from Total Variability are associated with the cluster, the speaker information is immediately available whenever a cluster with a known identity appears.

The language ID module is able to discriminate between the following languages: Arabic, Basque, Catalan, English American, French, German, Hebrew, Italian, Portuguese, Russian, Spanish and Others languages, which are the languages of the project.

Automatic Speech Recognition
One of the goals of the project is to enable the linguistic processing of the spoken information embedded in the audio data. In order to provide speech-to-text transcription capabilities for all the languages for which the LEAs have shown interest, a mix of proprietary system development and commercial system integration was implemented.

There was a development of proprietary Large Vocabulary Continuous Speech Recognition (LVCSR) system, based on Voice Interaction's AUDIMUS technology. This is a hybrid Automatic Speech Recognizer (ASR) that combines the temporal modelling capabilities of Hidden Markov Models with the pattern discriminative classification capabilities of Multi-Layer Perceptrons in terms of acoustic models and a decoder based on the Weighted Finite-State Transducer (WFST) approach encompassing vocabulary, lexicon and language modelling.

We developed systems for Portuguese, English, Spanish, Basque, Italian, French and German, and are available on the project system. Broadcast news data for those languages has been collected and annotated in order to train the baseline systems. For Catalan, Romanian, Russian, Arabic and Hebrew commercial systems were integrated.

Spoken term detection
Spoken term detection (STD) technology was developed to allow open vocabulary search of specific terms over spoken contents. A two stage process was implemented. First, the generic LVCSR system was used to generate word and phone lattices. Then, the lattices are searched to determine the likelihood of the searched terms’ occurrence.

The main goal of the Spoken Term Detection (STD) module is to freely search of specific terms over spoken contents. The development of the Spoken Term Detection (STD) module has been done based on the output of the LVCSR system, using specific information from LEAs in order to update the vocabulary accordingly, and using search over phone lattices to determine the likelihood of specific out-of-vocabulary terms’ occurrence. The vocabulary and language model of the LVCSR systems were tuned to include key named entities and content words defined by the LEAs. Different search strategies, based on words or sequences of phones, are combined. Since commercial ASR systems do not provide functionalities for lattice and phone level search full STD was only implemented for the languages with proprietary ASR engines.

5.4 – Video Pre-Processing and Information Extraction
Video content may contain multiple images, audios and text tracks encoded in several formats. Processing all these data
enriches CAPER, therefore the main goals are to identify the container and the type of content that is included in each video file, to demultiplex and decode it to formats more suitable for CAPER platform and finally to extract video frames in order to analyse them as images.

These are the performed tasks:
• Format Detection. Extraction of video information such as, container type, coded type, frames per second, image resolution, etc. that will be useful for the following analysis.
• Demultiplexing and normalization. This module extracts audio tracks, normalizes them to the format agreed for audio files and saves them in Normalised Repository in order to have it available in the system.
• Scene boundary detection. Two different methods have been implemented: one based on histogram differences and the other based on DCT coefficients.
• Best Frame detection. In order to reduce the high data redundancy of videos and to speed up the content analysis processes, shots are represented with single frames that will be processed by the image analysis process. This module will select the best frame considering the significant information within each shot.
• Component Packaging and Interface. All the modules developed have been presented as a unique API with different input parameters. The integration with CAPER platform has been done using WS as designed in the platform architecture.

The main contribution lies in the shot boundary detection and best frame extraction tasks as they nourish the system with new images to be processed in the image analysis module, empowering the analysis capability of CAPER and broadening the type of multimedia content that can be analysed. Even more, the demultiplexation of audio and text tracks permits obtaining a deeper knowledge about video files, permitting to add this information to the system.

5.5 – Mash-up with Semantic Web data collections
The main objective of Mash-up with Semantic Web data collections was the exploitation of social web content and semantic web data currently available over the Web in order to provide the possibility to collect and integrate these different kinds and sources of online information.

As a result, two main objectives were achieved: a service to support the named entity recognition, based on open source and linked data archives and a module to analyse social media contents has been developed as web services providing CAPER of semantic and social media analysis capabilities.

According to the activities performed under the task Study and Analysis of Social and Semantic web data, a state of the art about both semantic web and social media analysis tools for open source intelligence purposes was performed. Moreover, a state of the art about social media useful in the investigations has been carried out.

The result of the task Semantic Web Linker is a module that performs the integration between web data, coming from different sources, including Semantic Web Data. As proof of concept was used the illegal drugs domains. In this context information coming from open source archives and Linked Data collection (i.e. Dbpedia and GeoNames) was collected and integrated. The result is a service that helps the disambiguation of named entities related to the illegal drugs domain.

In the Social Web analysis, activities were focused on the analysis of social media data. In particular, a module named Social Media Analyzer has been developed. The module is able to analyse information collected from Facebook pages, groups and events and extract the graph of relations between users.

5.6 – Biometrics Analysis
The objective of this task was to develop a biometric analysis module including face and speaker identification that can deal with images and videos taken under uncontrolled conditions (low resolution, poor illumination, etc.)

The functionality of the biometric module aims at predicting the probability of a certain subject existing in an input search image based on his/her face. The subject is defined by multiple images (one or more) that contain his/her face (as the main face in the image).
In the first part of the project, the proposed approach of 2D to 3D face recognition was developed and tested. This approach showed unsatisfying results for images taken under uncontrolled conditions, which caused the decision to move to the feature-Level face fusion techniques.

The main work in this part was done in developing a high performance solution for the efficient search for a certain subject face within given search images. The work included the development and combination of algorithms for image normalization, pose correction and feature-level fusion that succeeded in the end in a reliable and efficient approach. The second part of the work was the development and the implementation of the appropriate architecture of developed functionalities. These functionalities together with configuration abilities have been successfully implemented into the CAPER platforms via web services.

Another functionality that was part of the biometric analysis module is speaker identification: In the workflow of the proposed solution, speaker identification is performed on the audio crawled or extracted from a video and saved in the shared repository.

The algorithm creates models for a set of speakers associated to the specific RL and searches for those speakers in all audio files processed by the system on that RL. It is in the responsibility of the system operator to define the speakers of interest and to request the creation of speaker models to the system. When an audio is processed, the algorithm loads the current speaker models, provides audio analysis and outputs an XML file per audio file with the speaker segments belonging to the speaker models.

WP6- Visual analytics

The goal of WP 6 was the creation of the CAPER Visual Analytics (VA) Framework: as a software application for the visual analysis of large amounts of data. It constitutes for LEAs the main gateway to access, visualize and exploit the intelligence delivered by the CAPER text, picture, voice-analysis and social-web-capture tools.

Through visual analysis, the CAPER VA Framework can reveal new insights into organisational workings for the benefit of crime control. Furthermore, the framework can reflect the fluidity and adaptiveness of criminal organisations by evolving in step with their activities. It enables easy identification of central entities and peripheral or isolated entities, which, in turn, will enable the identification of key players as well as potential weak links.

WP6 contained four tasks dividing the work into four main parts.

Framework Specification: Within the initial task we developed the basic architecture for the Visual Analytics Engine as well as the system architecture for the integration of the other components to form the Knowledge Base and the Workbench of CAPER.

In close collaboration with the participating LEAs, WP6 selected appropriate data-mining and visualization tools to provide the users with the most appropriate solutions. To gather necessary input from the users separate Workshops on WP6/Visual Analytics have been held with the LEAs by the partners involved in the WP. LEAs current practices and requirements have been collected by using a standardised questionnaire by all partners. The deliverable D6.1 “Visualisation & datamining tools for visual analytics framework” gives an overview of current LEAs work as well as their basic requirements on Visual Analytics. Based on this valuable input we selected appropriate data-mining and visualization techniques. The deliverable D6.1 includes a specification for the Semantic Analysis and the Knowledge Base Browser as well.

Visual Analytics Knowledge Base: Primary objective was the implementation of the database for the extensible Knowledge Network. The Knowledge Network is the basis for the Visual Analytics Workbench and the Knowledge Base Browser and Editor
The VA Workbench. The CAPER Visual Analytics Knowledge Base supports knowledge derived from all media types supported by the CAPER data retrieval modules. Therefore the knowledge base can handle information extracted from text, audio, social media and image knowledge files received from WP5 via the Enterprise Service Bus (ESB).

Visual Analytics Engine: The main part of the Visual Analytics Engine is the Visual Analytics Server developed as data, data mining/transformation and visualization provider. The server can be used to connect to multiple data sources of different formats simultaneously. It provides a set of pre-defined data transformations and data mining algorithms. It is also possible to extend the functionalities of the server with 3rd-party or custom software components, which provide additional data transformations or data mining techniques. Finally the server allows the definition and reuse of web based visualizations utilising the (transformed) data provided by the server. The CAPER VA engine also consists of the semantic analysis module (SAM) which receives analysed data (the output from WP5) from the central repository. The SAM extracts those parts of the analysed data which can be used to create the Knowledge Network. Primarily entities and relations provided by the text analysis. But it considers also information from the social media analysis and the image modules.

The processing for all data types has been implemented (see Task Visual Analytics Knowledge Base). The Module has been successfully connected to the central CAPER repositories. Updates are performed using a push paradigm which allows for fast, frequent and efficient updates of the Knowledge Base whenever new analytical results are available in the central repositories. The first implementation pulled the repository on a schedule which proved less efficient and create a larger time delay. The VA engine was implemented with a persistence layer supporting relational (MySQL and PostgreSQL) and non-relational databases (MongoDB, OrientDB, and others). Support for Graph Databases was developed during the project but did not show sufficient stability and performance to be actually used within the project due to some limitations of the database software. Support for the creation of arbitrary processing workflows is provided. Workflows can be created in an administrative interface for testing as well as interactively via a JavaScript based programming API. The REST interface and the JavaScript client side library have been described in the deliverable D6.3 “Pre-processing, Data-exchange modules and Semantic Web browser” (Chapter 3). The full architecture of the Visual Analytics Server has been described in D6.4 “Data-mining and visualisation toolset”.

Visual Analytics Workbench: The VA Workbench consists of three basic parts, the KBBE as exploration based interface for the LEA analyst. Here the user can access all entities and relations provided in the Knowledge Base and explore the network node by node. The other two parts are strongly connected to the VA Engine. One is focused around the data exchange / data standardization provided by the VA Engine, the other part are the CAPER specific data access, data mining and visualization components.

The KBBE has been implemented as a web application for the analysts to access the knowledge network. The KBBE provides also an editing capability to enhance the underlying Network with custom nodes and relations. This has been implemented to allow analysts to annotate the network and add information which was originally not present. Additionally an interface has been developed to allow access to the original (raw), normalized and knowledge (analysed) data in the central CAPER repositories. This interface contains specific views for text, audio, video and image data, each of them focusing on the characteristics of their respective extracted knowledge. A graph motive miner has been implemented and allows, together with the respective visualization to perform visual queries against the Knowledge Base, to search for specific patterns of entities and relationships in the underlying network.

Conclusions for WP6
We have shown that Visual Analytics could provide useful access to an entity relation network for LEA analysts. We have shown that it is possible to aggregate the results of the automatic analysis of various data types into a single coherent entity
relation network. The KBBE and the Graph Query interface have shown that we can enable access to the entity relation network using two distinct paradigms: exploration and search. There is further potential to strengthen the integration of these two paradigms into a single interface for the LEA analyst. Another interesting enhancement would be the integration of more means to manually correct the automatically created network, like the possibility to split and merge entities and their relations. This would be useful in situations where the automatic detection cannot distinguish between different entities of the same name or unique entities known under more than one name. We have created a Visual Analytics Framework in the form of an “Information Visualization and Visual Analytics Server” (IVA-Server) which has already been used in other projects and contexts.

WP7- Embedding legal and ethical norms and standards into System Design, Development and Deployment

The objectives of WP 7 where as follows:
- To provide the relevant legal and regulatory framework for CAPER at different levels (EU, national, and local), including the technical standard dimension: ISO norms and European Standards (ENs).
- To provide the conceptual framework to ensure the semantic interoperability of the system.
- To establish protocols and guidelines to ensure that the technological aspects of CAPER advance in line with necessary recommendations
- To ensure that Caper is deployed in a manner that is both ethically and legally coherent.

These objectives allocated to IDT-UAB and BAK in the CAPER project can be classified in three main subjects: i) review of regulatory framework and ethical standards at EU, national and local levels (WP7); ii) the implementation of an ontology to improve interoperability between European LEAs (WP7); iii) the design of an ethical and legal strategy to protect individuals rights in the use of the CAPER platform. In this way we can see that objective 3 and 4 are review together as part of the third goal of the project as one lead to another.

Review of the ethical standards and regulatory framework of CAPER at the extra-European, EU, national and local levels.
The first goal was covered at the beginning of the project in close collaboration with the Ethical Committee Board. The work was performed buy the team of the IDT-UAB in the cases of EU and Spain, and by the team of BAK in the cases of Italy, France, Germany and Israel. Deliverables from D7.1 “Report on Regulatory and Ethical Framework in CAPER” to D7.4 “Mid-term Ethical Audit on System Development and Deployment” report the results of this work aimed at supporting technical partners by defining the framework that current ethical standards and legal provision set for a platform with the capabilities of CAPER.

Reusable models of shared legal policy concepts: OWL ontologies for semantic interoperability.
In order to achieve the second goal, the implementation of an ontology to improve interoperability between European LEAs was performed. The work done consisted, first, in the definition of an Organized Crime Structure (OCS) based on Europol Annual Reviews and the International LEAs cooperation literature. This structure is devised to provide a common supra national structure in order to provide interoperability for European LEAs. Once the structure was defined, meetings with the LEAs involved in the project were scheduled in order to obtain its validation. The feedback from these meetings was taken into account and the resulting final structure was used to build the concrete ontology, named as European LEAs Interoperability Ontology (ELIO), which models the OCS, the relationships among its concepts, the attributes and all the knowledge directly gathered from LEAs. The main idea is to ease the sharing of information related to organized crime among LEAs within the CAPER platform. The result of the tasks presented was the design, development and implementation of the ontology within the platform. Deliverable D7.5 “CAPER OWL semantic models” reports all the process carried out in this second goal.

Formats and policies to store and manage CAPER multimedia files and Legal and standard compliance
As for the third goal, the work done consisted on a deep analysis of the legal requirements and legal implications of the CAPER platform. Since the beginning of the project it was clear that the nature of the technology being developed and the
implications for individual rights that, a general overview of the legal instruments applicable was not enough and a specific Regulatory Model for CAPER was needed. Following this line of though the work performed was directed to define this CRM (CAPER Regulatory Model) and to be able to answer two questions by the end of the project: i) Is the CAPER Regulatory Model acceptable within the European Union community on Data Protection?; and ii) How can these elements be integrated not only into the CAPER model (CRM) —which constitutes a single privileged case triggering the reflection— but into a more general framework, to be shared and reused in other Web 2.0 and Web 3.0 scenarios, Security projects and Data Privacy developments? The answers to these questions were included in D7.7 “Final Report on Systems Review and Approval” and D7.8 “EAG Ethical Code. Final Ethical Audition System Development and Deployment” as the final results of the work performed within this second subject.

WP8- System Integration and End-User Testing

The main goal of this work package is the validation of the results of the CAPER platform by defining and executing a set of functional trials and demonstrations at component, sub-systems, system and central management application levels in order to validate the solution and demonstrate the operational benefits to LEA users.

The result of this task was meant to be the delivery of a useful, user friendly, final-user validated and compliant system to support LEA users in the mining of crime trends and in the prevention of the organized crime. This task implies to correctly integrate all the components developed and improved in the project, performing and exhaustive testing and being validated by the end-user.

Overall Systems Integration

A first prototype was created by Synthema, called TIGER, including parsers for Italian, German, Romanian and Japanese and producing KAF (version 1) as output. After some months of continuous development the parsers for English and Spanish were integrated by Vicomtech. Synthema and Voice Interaction worked on the web service functionality to integrate the parser for Portuguese. The parsers of the second version of the TIGER prototype produce KAF(version 2, with enhanced named-entity recognition) as output.

For the final platform specification, S21sec performed an extensive work in defining the general specifications to integrate all CAPER modules into the service-oriented architecture of Platform 1. Platform version 1 plan was agreed by all partners including a Work Breakdown Structure, modules description, schedule and governance information.

First months of integration were hard due to incompatibilities of the different development frameworks. Although common and, in theory, interoperable interfaces were designed not all the components were integrated at first try. Time was dedicated for debugging and testing the interfaces to reach the integration objective for the first version. Full deployment of CAPER platform 1 was completed for end-user testing at the beginning of the third year of the project.

In parallel to the integration work, the CAPER Orchestrator – based on Talend ESB – was deployed and an orchestration workflow was configured. This workflow included all CAPER modules and common repositories working together in a distributed SOA environment.

Once CAPER platform 1 was integrated, a technical team discussed and agreed the scope to be included into CAPER platform 2, which was developed and finally integrated two years and a half after the beginning of the project. This new version included new functionalities that provided the end-users with the opportunity to test the full CAPER workflow.

The integration was done in the development environment thus the next step was to update the virtualized platform (pre-production) with the new changes and let the end-users test. This virtualization was released to the LEAs at the end of the third year of the project.
After 30 months of project, two more iterations of the platform were decided to be carried out with the objective of increasing the quality of the already integrated components and improvement of the performance. Version 3 and version 4 were done as well as virtualizing these versions. The last step of the integration process was to deploy an instance of the CAPER platform in the LEAs premises. In order to achieve this goal, a first “pilot” deploy was done in a LEA. Then the rest of the LEAs will have their own private version.

Cyclical user field trials and validation against requirements
During the system integration activities, taking advantage of the user scenarios description and prioritization included in D2.2 “System, Hardware, Software and User requirements specification”, a CAPER platform test plan was created in order to ease the cyclical end-user testing and validation. Because of the above mentioned delay in the deployment of CAPER platform 1, the end-user testing plan was also postponed at the beginning of the third year of the project. A complete testing plan and testing execution templates were created and used by end-users, obtaining a better/more objective validation of the CAPER platform against identified user scenarios. In the third year of the project LEAs were testing the platform providing their impressions about the work done between V1 and V2. We added their feedback as improvements in the scope of V3 and V4 (V3 wasn't released to the LEAs). Finally after release of V4 in mid July, LEAs had the possibility of testing the platform till October 2014 (end of the project).

Conclusions for WP8
This Work Package has been a key component because it has a very important general objective to reach, which is the integration of all the components involved in the project. It is also strongly connected to WP2, a correct and detailed definition of the system architecture and common interfaces for all the services have had a big influence in the overall system. Having this premise, the system is easier to be implemented. The result of the CAPER platform, which is implicitly the result of this work package, shows a platform in which all the components are running in an integrated way. Further improvements should be made in order to achieve all the components smoothly running.

WP9- Dissemination and Exploitation

WP9- Dissemination and Exploitation

Work Package No. 9 tackles dissemination and exploitation of the projects activities and results.
- Project Dissemination. The goal of this task is to achieve an extensive dissemination, to promote awareness of the project activities and results.
- Exploitation plan. This task aims at developing a plan that shall help the partners to exploit the results of the project.
- Intellectual Property Management. This task tackles the management of intellectual property rights (“IPRs”) possibly arising from the activities of partners, performed in accordance with their respective WPs’ descriptions.

The main activities performed under this WP are the following:

Project Dissemination
The first step has been the elaboration of a dissemination plan. In the plan, the partners pinpoint the instruments they want to resort to achieve the desired level of dissemination. Namely, they identify what to disseminate (i.e. the project itself or rather the single pieces of knowledge produced), and associate it to specific dissemination tools (e.g. journals, conferences, workshops, fairs, etc.). During the project lifecycle, the partners have undergone intensive dissemination activities in accordance with the dissemination plan, by promoting the project and their results to the addresses identified for this task.

Exploitation plan
The partners have extensively worked to identify what could be subject to fruitful exploitation and the ways of exploitations. In this regard, the partners have elaborated a specific plan according to which: they wish to exploit specific pieces of knowledge
they have produced (i.e. Foregrounds), at the same time they wish to grant a certain use of the final platform to LEAs that have been partners to the project, and which contributions have been helpful in the development of a business plan to further develop Caper after the end of the project.

The relevant partners’ Foregrounds are listed in section B, par. 5 of this report, along with the intended exploitation form. With regard to use of the platform by partners LEAs, the partners, under the leadership of S21sec, have agreed on specific terms of use for the same. In the end, always under coordination of S21sec, the partners have deepened, under the business and financial points of view, the possibility to continue to cooperate to bring the platform from a prototype stage to a market stage.

Intellectual Property Management
The activities occurred under this task can be summarized in an on-going assistance provided, during the entire project, to partners in connection with their rights on the results they have produced. The partners have always been assisted, individually and jointly, to understand whether the relevant Foregrounds produced remain subject to sole ownership or joint ownership (depending on the level of contribution of each partner to development of the same Foreground). At the same time, the partners have been assisted to understand the ownership structure of the platform, it being the result of the cooperation of the relevant partners. All the rules applicable in this field are reported in a dedicated deliverable (i.e. D9.4 “IPR Management Guide”).

In connection with the activities illustrated above, the main results achieved can be summarized as follow:

Project Dissemination
Since the beginning of the project, it has been created a dedicated website, where internet users could find information on the project. Partners have produced a relevant number of scientific papers, on both technical aspect of the platform and legal/ethical implications, which have seen publication not only in the EU. At the same time the partners have spread information on the project through a huge number of articles, events, conferences and fairs, which have seen as target, in a proper balance, the scientific community and the LEAs and other operators of the security fields. CAPER has participated in several events such as the CyberSecurity EXPO in London (October 2014), the IEEE JISIC 2014 in The Hague and has cooperated with several European projects such as EPOOLICE, Virtuoso and INSEC. Regarding publications, CAPER has produced 9 scientific papers, 13 Poster/presentations, a video and a scientific book of the Springer editorial. The partners have also brought the platform to the attention of several LEAs - not part of the project - and to EU institutions like Europol and Eurojust. In order to maximise the dissemination effort, the partners have produced a dedicated video, to be used on a case-by-case basis to showcase the platform capabilities to interested end-users. In the end, the partners have organized a final dissemination event which took place during the last Cyber Security Expo in London.

Exploitation plan
The first goal achieved by partners in this field is a list of Foregrounds that will be subject to future exploitation by the same partners in several ways (e.g. in other research project, on the market, or provided under open source license to the scientific community). As second goal, the partners have achieved a preliminary business plan that is the ground of discussion for their possible future partnering in the further development of the platform.

Intellectual Property Management
The main result achieved in this field is the legal framework applicable to the ownership of Foregrounds. As a remark, it must be said that it has not been possible to seek for particular protection of the same Foregrounds in ways other than keeping them confidential.

Conclusions for WP9
In light of the above, it can be concluded that the approach adopted by the partners, in the fields of dissemination and exploitation, has permitted to successfully complete the tasks of this WP. The dissemination effort put on the field by the
partners has achieved the results envisaged in the relevant dissemination plan. LEAs, member of the scientific community, EU institutions, possible end-users, have all been informed of what Caper is able to do. With regard to exploitation, the partners have tackled the matter from all the angles possible: (i) exploitation of the single Foreground produced by each partner; and (ii) exploitation of the platform, that is the result of the combination of all the single Foregrounds produced, and that represents a Foreground itself.

Conclusion
The CAPER project was born as an ambitious project formed by an international consortium composed of 17 organizations in 7 countries. The main objective was to create a common platform for the Law Enforcements Agencies in Europe for fighting against all types of organized crime among other goals such as fostering the cooperation between LEAs or establishing a common regulatory model for CAPER.

During the 40 months of the CAPER project many things have been done. Starting from the technical results, CAPER brings together a set of acquisition, analysis and exploitation modules that makes a unique platform able to handle different types of media formats. In the multilingual aspect, the platform integrates 14 languages in both crawling and text analysis. Biometrics (including face and speaker recognition) adds special value to the platform as there is no other software in the current market that incorporates these two kind of analysis with other 6 types of analysis (Text, Video, Image Comparison/Classification, Social and Speech). Regarding the acquisition modules, CAPER is able to gather content from public sources such as the Web, blogs, forums and also can download social content from the Facebook social network. In addition, CAPER has a multimedia module that can capture public TV channels and radio stations.

The exploitation module, “Visual Analytics”, gives to the end-user an intuitive and powerful tool to explore and query the results obtained in the process. With this tool the LEA analyst can access the information that was automatically gathered. Time savings – compared to traditional web search - can be achieved through the automatic data collection and analysis as well as through the usage of specialized visual representations and data mining components.

The fact that both crawling and analysis modules perform their activities from a neutral point of view is a very important aspect to bear in mind. That means that the information is downloaded and processed directly from the Web, with no human interaction or correlation with databases is applied discarding also cultural biases or tagging.

Regarding non-technical objectives and results in the project, it is important to highlight the constant effort to define a common regulatory model able to manage interoperability among LEAs and their compliance with EU Directives, regulations and national statutes. In addition, the positive result of the ethical audit thanks to the recommendations made by different European Data Protection agencies.

From the LEAs perspective, the participation in the project has been a positive experience. The construction of a platform that integrates high demanded technologies would allow them to have a complimentary system that may help in their daily investigations. Currently, CAPER is not a market-ready product, it needs to have a further development to improve the quality of some modules and implement functionalities that were not added to the scope of the project such as "deep web search". These improvements are already detected and are to be evaluated by the consortium on how to carry them out once the project and the consortium framework has finished.

Potential Impact:
Potential impact of the project

Raise awareness of the EU political stakeholder
CAPER is the result of an ambitious project that aims integrating a large number of technologies into a single platform. These technologies are in high demand from organizations such as governments or LEAs that have as objective the preservation of
citizens’ security. Some of these technologies such as crawlers and analyzers may suppose a stopper in the citizen's trust to this kind of systems.

One of the main points of this project regarding the citizens’ privacy that has been carried out is the safe storing and a controlled use of the information gathered. As it is mentioned below in the section “Impact on the citizen” several ethical and legal audits have been passed to check that these principles are applied in the CAPER platform.

Other and not less important aspect to remark is the "neutral" point of view from a linguistic standpoint. The collection of information resulting from implementation of the technical solutions envisaged by this project guarantees a set of information gathered and offered for use that is free of bias and furthermore that may be used by different persons speaking different languages.

CAPER, as explained in Dissemination and Exploitation work package results, make use of industry standard and open technologies that have beneficial impact on further exploitation of the technology designed as output result of this project since inter-operability and integration of said technology is always taken care of. Furthermore, open and standards based approaches help guarantee the long term development and use of the system after the project has concluded. From the point of view of the end-users of the project, CAPER represents a project whose technologies are in the right direction of the current market. In this sense, the CAPER consortium has produced a business plan to continue developing the platform.

Demonstrate to the LEAs the added value to collaborate
LEAs are optimistic about the possibilities of using such a platform that integrates a set of high demanding technologies because it can help them fight against organized crime with technologies such as image and audio analysis, which are time consuming activities and require significant human working investment.

Organized crime groups, known as being geographically distributed, use Internet to enhance communication between them. Internet is also a means to deliver information about criminal activities or ways to execute them. In this sense, CAPER platform, as an automatic gathering and analysis software, would have a great impact on these activities since LEAs would save time detecting threats and preventing organized crime activities.

During the CAPER project all LEAs have been closely and actively working with the technical group and among them. They have shared their expertise and knowledge in every meeting the consortium has organized. Firstly defining their needs, specifying their requirements for the platform and then telling us their expectations about what CAPER should do. Every LEA had the opportunity to share their vision about the current organized crime situation in their territories rising up debates and constructive discussions among other LEAs. Observer LEAs have also supported the consortium with their expertise sharing their situations in their countries enriching the project.

Special mention has the fact that their feedback after months of testing over different versions of the platform was very appreciated and important for the continuous development of the prototype and was added to the business plan of the platform. After the finalization of the project the LEAs are very happy to have collaborated in such a project because it also provided them with an opportunity to add more contact with other LEAs which may be useful for daily work. They have also created contacts with other Agencies that may be useful for their daily work.

Economic impact
In order to evaluate the economic impact that a system such CAPER would have after its finalization a set of metrics (technical metrics) were defined and designed. These metrics and a benchmark document were filled by a representative LEA with the objective of extract conclusions about the whole project and the platform.

The results are promising if the platform use can make a difference by reducing time spent using an automatic tool such as
CAPER compared to traditional way of searching was too high. It means that the power of a tool that can automatically download thousands of documents and analyze their contents in some minutes or hours is a very positive point for the tool. The results of the test weren’t so promising when it was seen that some of the results were false positives and that there were instabilities in the platform.

After these tests the consortium realized that a further effort should be made to put this prototype in the market developing a market-ready product. The possibilities of such a product, in terms of potential clients, is wide and transnational as the platform contains modules that crawls and analyzes documents in 14 languages. In the business plan there are no numbers yet as the consortium is negotiating how this further development will be but in estimations of S21sec company together with other LEAs 30% of Digital Surveillance manpower can be saved using a system of CAPER.

Impact on the citizen
Generally, the citizens accept that systems such as CAPER should be implemented to preserve their security. These kind of systems for analyzing the virtual, digital world, uncovering underlying networking connections (both direct and intrinsic connections) between entities (organizations, individuals, etc.) that compromise the general standard of living and social safety may also cause citizens to fear that their privacy and liberty will be undermined, or worse, violated. In this project, from the beginning, a team composed by ethical and legal professors, lawyers and other experts in the area have been working in assessing the technical team to avoid CAPER overstep the citizens privacy.

Several workshops have been maintained with European data protection agencies such as Eurojust and Spanish Data Protection Agency to check that the system comply with the current law of data privacy.

Thanks to the legal and ethical procedures followed in the project, CAPER platform can foster the general trust in the activities of governments since it allows a safe storing and a controlled use of the information gathered.

Impact on the competitiveness of the proposers
The CAPER project has been a very good framework to build a system for the detection and prevention of the organized crime in collaboration with technical and non-technical partners. The consortium has done a strong work in disseminating the project since the beginning and at different stages of the development of the platform that allowed other companies know about the companies and technological institutes that are working in the project. This is clearly beneficial for these organizations as they are seen as institutions collaborating with a common objective that is to build a system that should be useful for the LEAs and thus to the security of citizens.

The project has been clearly beneficial to foster relations and communications between partners, for instance, S21sec and Vicomtech are working in more than one project besides CAPER, Fraunhofer IGD is working with Vicomtech as well in other projects, facts in which CAPER has a lot do with it.

In the case of S21sec, there are commercial actions and open talks with some of the other partners to maintain a strong link of collaboration or to acquire some technology to be appended to the proprietary systems.

Main dissemination activities and exploitation of results

As a first step for an efficient dissemination, partners have realized the CAPER official website, available at www.fp7-caper.eu. There internet users may easily retrieve information on the Caper Projects' features. The website illustrates in depth the functionalities to be implemented for assistance to LEAs activities against any kind of crime, the technical challenges to be faced when dealing with processing of data gathered through open and closed sources, and the methodology adopted in the project to overcome the current status of art in the relevant fields.

Going more in details into dissemination, for the purpose of this section it is necessary to clarify that the dissemination effort has been split in two separate phases. Phase 1 is represented by the “Communication” of the project: it represents the basic part of the dissemination and it consists in raising awareness on the project’s aims and objectives. The main targets of this
exercise have been identified in members of the research community, LEAs and, possibly, policy makers. Phase 2 is represented by the “Dissemination” of the project’s results: it consists in producing and spreading, in the most extensive way possible, any publication concerning the specific knowledge, technical results and information generated within the Caper Project. The main targets of this second phase can be searched in a wide range of stakeholders, having different natures, as well as in potential customers of the final product of the project.

The partition illustrated above inevitably follows, and it is therefore due to, the developments achieved from time to time by the partners according to the relevant tasks as scheduled in the DoW. During the first half of the project (from July 2011 to December 2012) the partners have focused on the Communication, as described above. While during the second half of the project (from January 2013 to October 2014), the partners, having produced material results on the basis of the preparatory work carried our during the first 18 months, focused more on the Dissemination.

Many of the results above have been illustrated and explained, by partners, in several scientific publications addressed to an audience of peers and scientific researchers active in their same fields. The publications have presented both single, specific technical operative aspects of the Caper Platform and, when close to the end of the project, the platform as a whole. The entire project has been illustrated in a dedicated book, yet to be published in the “Law, Governance and Technology” series of the editor “Springer”. The proposal was submitted and accepted for publication by Springer and is currently in the phase of finalization of the final draft to be published in the next months. The main asset of this publication is the contribution of all the partners involved in the project: technical, academic and LEAs. The book has the potential to become a key reference in the field of security as it contains chapters on technological development, LEAs activities and best practices, data protection in the field of police works and ethical and societal issues in such an important area as open source intelligence in the fight against organized crime. In a way the book summarizes all the aspects and disciplines involved in the Caper project and, at the same time, presents the results obtained in each one of those fields in a coherent way.

Always within the framework of the scientific dissemination, partners have also participated to a large number of conferences, events, exhibitions, presentations and workshops, where they have had the chance to illustrate in detail either the technical results they have produced within their respective work packages, or the final platform as a whole. In compliance with the dissemination plan, and in particular with the idea to disseminate particular pieces of knowledge, partners have meticulously selected important events where they have had the chance to present:

• The technologies developed to produce the single module(s) they were in charge to develop, according to the work package descriptions of the DoW, and
• The Caper platform.

In order to participate to such events, the partners have prepared and submitted before the relevant organizations, proper scientific papers. The overall figures of events attended amount up to 47. With regard to the audience addressed from time to time, the partners have selected the relevant events in accordance with the distinction between Communication and Dissemination illustrated. In particular, the dissemination effort, in compliance with the mentioned distinction between Communication and Dissemination, has always kept an eye to maintain a proper balance between dissemination toward the scientific community, on the one side, and possible end users (i.e. operators of the security field like LEAs), on the other side.

Since the very beginning of the project, partners, either technical partners or LEAs taking active part to the project, have presented the project to other LEAs outside the project. At the very beginning, the project has been presented to the Europol in order to get feedbacks on how to develop the platform. The Europol is a EU multi-disciplinary agency founded under the Treaty on the European Union, comprising not only regular police officers but staff members from the member states’ various law enforcement agencies: customs, immigration services, border and financial police, etc. Europol’s aim is to improve the effectiveness and co-operation between the competent authorities of the EU Member States primarily by sharing and pooling intelligence to prevent and combat serious international organized crime. Its mission is to make a significant contribution to the European Union’s law enforcement efforts targeting organized crime. Time after, and after having developed more concrete results, the partners have presented the project and different releases of the platform – through 1 to 1 sessions – to police authorities coming from Germany, Spain, Estonia, Holland, Ireland and UK, thus directly reaching a broad range of
authorities coming from countries not directly involved in the Caper Project. In addition, partners have approached the audience of LEAs also through fairs and events seeing the attendance of representative of police authorities coming from all over the EU.

In order to get the widest acceptance possible of the platform, the partners have organized two ad-hoc meetings, with representatives coming from several data protection offices of several EU Member States. The first meeting was hosted by the University of Barcelona on the 29th-30th November 2013 in Paris and the second one on the 16th May 2014 in Barcelona. They have seen the participation of representatives coming from Europol Data Protection Office, Eurojust Data Protection Office and the Spanish National Data Protection Agency.

Particular attention has been paid to events and meetings with other research projects funded under the FP7. In particular, the partners have sought for synergies and exchanges of knowledge produced with the projects (i) VIRTUOSO (Versatile Information Toolkit for end-Users oriented Open-Sources exploitation), and (ii) INSEC (Innovation and Research within Security Organizations). Technical partners, as well as LEAs partners have met several time representatives of both the projects, by attending events organized by latter or by inviting them to dedicated workshops.

In order to maximize the impact of possible presentation of the Caper platform, the partners have decide to invest resources in producing a video footage of the project. The video showcases the goals of the project, how the platform works and how it can be used by end-users (i.e. LEAs) in their day-by-day activities. It is intended to be a presentation of the main technical capabilities of the platform, so to let the audience understand the value added by the platform in the prevention of organized crime.

In the end, partners have invested resources and energies in a final dissemination event. As the project's final dissemination event, the partners have selected the Cyber Security Expo in London, a fair where the major sector-specific operators assembled to exhibit their solutions and products, as well as to facilitate networking and partnerships.

The event took place on the 8th and 9th of October 2014, at the ExCel Center London where it was co-located with two other top-level technology fairs, namely the IP Expo and the Data Centre Expo, having a slightly different focus but evident connection with the cyber security sector. In addition to the stands run by the single exhibitors, one of which entirely dedicated to the Caper project for the two days, the fair hosted a tight schedule of speeches and demonstrations, managed by the single exhibitors, and also plenary sessions, covering all the hot topics and latest technology developments in the field of cyber security, internet business and information technology in general. The event was therefore designed and perceived not only as a showcase for commercial solutions in the relevant sectors, but also as a high-quality overview on the state of the art. In this sense, it certainly was a suitable and prestigious venue to raise awareness on the cutting-edge technology developed in the scope of the project.

The stand raised a great deal of interest at the Cyber Security Expo. From a logistic standpoint, the stand the team chose for Caper was very well positioned, it being located just in front of the entrance of the presentation room where the plenary sessions were hosted. Furthermore, almost the entirety of exhibitors were cyber-security service providers, thus presenting a quite aligned range of solutions, though of the highest standard and sharing the same positioning in the supply chain. In this context, the Caper project, which was the only FP7 project represented at the fair, quite easily stood out for the originality of its concept, objectives and implementation.

Further to the presentation and communication activities at the stand, the Caper team also secured a slot for a 1-hour demonstration of the product capabilities on the second day of the event. The presentation was attended by a satisfactorily large audience, and was very well received, thus contributing to raising interest and awareness in Caper. The professionals who attended the stand and demonstration encompass a huge variety of figures, for example: private consultants, cloud business development managers, innovation service providers, cyber advisors in the banking sectors, private intelligence
service providers, information security officers, R&D officers, sales executives, as well as university researchers and lecturers.

The fair mostly being a convention of commercial operators, it is comprehensible how the majority of the questions posed to the Caper team related to a possible practical exploitation of Caper. Most of times, the team was asked whether a commercial release of the software as a service for public and/or private operators was envisaged. In addition, the certain attendees belonging to the academic community have expressed their interest in the concept of open source intelligence and proposed to host the CAPER team for speeches and/or demonstrations in the institutions they worked for, in the context of academic projects regarding cyber-security.

With regard to exploitation, this report covers only the so called “Foreground” that are the tangible and intangible results, generated during the project lifecycle of the project. In fact, it is clear that the project required a strict cooperation and interaction among the partners involved in order to successfully reach the desired results and benefits. This cooperation inevitably leads to the joint achievement of commercially profitable results (i.e. Foreground). In this connection, the partners have in first place elaborated a specific guidance on how to exploit and protect Foreground.

According to such guidance, in general terms, the partner that generates the Foreground remains the owner, and it is therefore free to exploit it to its best, without involving any other partner. However, it may happen that Foreground is the result of the participation of different partners. In case it is possible to define the specific piece of Foreground pertaining to each partner, participating to the creation of Foreground, each participant remains the owner of that specific piece. On the contrary, if it is impossible to distinguish the individual contribution provided by the partners, then the Foreground is subject to joint ownership of all the partners who provided contribution in generating the Foreground. In some cases, the ownership percentage might be defined taking into consideration the criteria of the efforts deployed by the relevant partners (e.g. Person Months) for the specific task giving rise to the Foreground.

According to this default joint ownership regime, each partner that participates to the creation of joint Foreground is entitled to use the jointly owned Foreground on a royalty-free basis, without the need to obtain prior consent of the other participating partners. In addition, each participating partner is entitled to grant non-exclusive licenses to third parties, without any right to sub-license, provided the following conditions are fulfilled:

• The participant Project Partner shall provide the others with at least 45 days prior notice; and
• The participant Project Partner shall provide the others with fair and reasonable compensation.

Each partner has produced, individually and/or jointly, a variable number of Foregrounds. Such Foregrounds remain of their exclusive or non-exclusive availability in accordance with the rules illustrated above. It means each partner is in the position, with regard to the Foreground on which it can claim any proprietary right, to re-use the same in other research project, license it on the market or make it available to the research community, for example, through open source licenses. In this respect, the exercise carried out by partners has been to identify the single Foreground they are willing/capable to re-use or otherwise exploit. Then, in respect of each of them, they have elaborated their own plan on what to do with the relevant piece of Foreground.

With particular regard to the platform, the latter being the final Foreground stemming from the project, the partners have also decided to grant the use of it to LEAs that have been part of the project as partners or observers. Even if LEAs have not technologically contributed to the creation of specific platform's modules have contributed by providing the operational requirements upon which the whole project is based, as well as they had an important role in the decision-making process which led to the optimal architecture of the system and LEAs have done testing of the different prototypes and provided feedback about the functioning of the different modules and needs of integration. The goal of the licensing is to compensate them, in the best way possible, for their efforts taking into due consideration their not-for-profit nature and the impossibility to benefit from any revenues upon commercialization of the project’s final product or single Foregrounds.
In accordance with the project’s goals, the technological partners have also elaborated a plan on how to further develop the platform after the end of the project. In this case it is important to bear in mind that the platform is considered as a single element (or Foreground) which is jointly developed and held by all the technical partners. They, all together, have contributed to its development and they are therefore the joint owners. The goal of such a plan is to bring the platform from a prototype stage to a commercial stage. Even if the platform matches the planned requirements it needs further improvements to become a product ready for the market.

In order to achieve a fruitful business plan, the partners have decided to assign the role of leader to S21sec. The latter has proved to be the entity having the resources needed to (i) study the market, (ii) elaborate a proper strategy to advance the platform to a more mature product, and, in the end (iii) face the challenges that the marketing of such a complex product could lead to. S21sec has than produced then a business plan. It has been elaborated in accordance with specific elements to be leveraged in order to understand which position the platform can get on the market in respect of potential competitors. Namely, it deepens the following aspects:

- key partners toward whom promote the platform,
- key activities to bring the platform to a more mature stage,
- key resources to achieve the desired result,
- the possible cost structure,
- best features of the platform to become an effective player on the market (i.e. value proposition),
- customer relationship to offer,
- channels through which promote the platform,
- revenue streams, and
- potential customers’

List of Websites:
More information is available on the project’s website: [http://www.fp7-caper.eu/](http://www.fp7-caper.eu/)

Coordinator’s Contact detail:
Felipe Melero
fmelero@s21sec.com - www.s21sec.com

Related information

| Result In Brief | A prototype for cyberspace security |

Reported by
S21SEC INFORMATION SECURITY LABS S.L.
Spain

Subjects
Security - Social sciences and humanities

Last updated on 2015-07-21
Retrieved on 2019-08-17
