Fighting Organized Crime Through Open Source Intelligence: Regulatory Strategies of the CAPER Project

Pompeu CASANOVAS, Juan ARRAIZ, Felipe MELERO, Jorge GONZÁLEZ-CONEJERO, Gila MOLCHO, Montse CUADROS

Abstract. OSINT stands for Open Source Intelligence. The CAPER project has built an OSINT solution oriented to the prevention of organised crime. We offer in this paper an overall view of some results, embedding into the system legal and ethical issues raised by the General Data Reform Package (GDPR) in Europe. We briefly describe CAPER architecture, workflow, functionalities, modules and ontologies (European LEAs Interoperability ELIO, and Multi-Lingual Crime Ontology MCO). This paper is focused on the indirect strategy to flesh out Privacy by Design principles (PbD) through the Caper Regulatory Model (CRM).

Keywords. OSINT systems, Security, Social Intelligence, Organized Crime, Privacy by Design, Ontologies, Regulatory Models, Ethics

Introduction

Security, Privacy and Data Protection are hot topics in EU legislation. As it is well known, we are now experiencing a transitional state of relevant European Directives, Regulations and Decisions on this issue. A General Data Protection Regulation is expected to replace the EU Data Protection Directive 95/46/EC, and the Police and Criminal Justice Data Protection Directive to replace the Framework Decision 2008/977/JHA. However, the General Data Reform Package (DPRP) issued in 2012 has not been fully adopted yet. Political discussions in the EU Parliament have been dragging on for a long time.

Although the legal structure of data protection and privacy rights are not equivalent [11], the main objective of the new regulations is putting citizens back in control of their data [10]. Under EU laws and Human Rights court rulings, this subject belongs to the so-called area of Freedom, Security and Justice [4]. One of the main features of GDPR concerns the relationship between Data Protection and Security.
According to the old regulations, security was considered an exception to suspend or reallocate the extension of individual rights, while GDPR resituates it into the same normative and ethical framework of fundamental rights. Human dignity and human freedom are considered to be above the balance between rights and threats: regulations should follow the same path, as recently stressed by Opinion 28, EGE [9]. But the question is how a balance must be struck between competing values while respecting these core principles, as organized crime networks have become more global in nature mimicking ethics, governance and good practices beneath a legal look.

We will present some results and some problems raised in the EU Project CAPER (CAPER).² Section 1 describes the preliminaries and objectives. Section 2 provides an overview of the architecture, workflow and modules of the platform. Section 3 focuses on the strategies to cope with these new regulatory trends. Section 4 provides some preliminary conclusions. We will try to clarify why ethical and legal concerns must be taken into account since the beginning, and how they can be handled.

1. Preliminaries: Open Source Intelligence (OSINT)

The goal of the CAPER project is to create a common platform for the detection and prevention of organized crime through sharing, exploitation and analysis of Open Source Intelligence (OSI). OSI or OSINT, is usually defined as intelligence collected from open sources. There is no homogeneous approach to this concept.

Some authors offer a narrow description tailored for intelligence services: “unclassified information that has been deliberately discovered, discriminated, distilled, and disseminated to a select audience in order to address a specific question” [18, p. 129]. This is a use of the concept that is linked to Intelligence Services and military uses [3]. But at least for the last five years, OSINT has been intended to mean a cluster of tools to browse the web, aggregate information, and get reliable profiles from websites, blogs, social networks, and other public or at least non-private settings [12]. From this broader point of view, it may be defined synthetically as “the retrieval, extraction and analysis of information from publicly available sources” [2]. The approach is applied to get, structure and manage information in a broad array of social domains — media, education, business, or disaster management.

But this does not occur without ethical and legal problems, even in the emergent humanitarian and most needed fields [16] [17]. Eg. Backfried et al. set a platform for disaster management whose architecture is strikingly similar to CAPER’s design to retrieve and process data from blogs, Twitter, Facebook and other sites with the aim of fostering more effective actions. They also state that one fundamental assumption underlying OSINT “is that information is indeed available in publicly accessible sources and just has to be gathered and provided to the right people at the right time” [1, p. 254].

We contend that in the light of GDPR the statement that information is out there, ready to be just gathered from all kind of sources because it is available, summarizes the pragmatic view we are trying to balance with principles and values in the CAPER Project.

² Collaborative Information, Acquisition, Processing and Reporting for the Prevention of Organized Crime TFP7-SECURITY-2010-1.2-1. http://www.fp7-caper.eu/
2. **CAPER architecture, workflow and modules**

2.1. **CAPER architecture**

The CAPER architecture design has four main components: (i) Data harvesting (knowledge acquisition: data gathering), (ii) analysis (content processing), (iii) semantic storage and retrieval, and (iv) visual analytics of data. They were planned according to the following criteria: (i) Logical architecture, in which a common data-exchange model and common module interfaces were designed enabling the construction of isolated modules that could be easily integrated in the overall system, (ii) physical architecture, with different scenarios in which alternative versions of the CAPER platform would be deployed (distributed, virtualized, and local scenarios), (iii) information architecture, which describes and implements CAPER common repositories and the way data is processed and stored in these common repositories (original, normalized, knowledge, and visual analytics repositories), (iv) Orchestration design, based in a Service Oriented Architecture (SOA), in which the data flow between all modules can be orchestrated by a core module.

2.2. **CAPER workflow**

CAPER workflow can be summarized in several steps, as acting through the Databases interaction plotted in Fig. 2:

1. LEA stands for Law Enforcement Agencies. LEA’s Analyst (LEA-A) can design a Research Line (RL) using a wizard in the CAPER Management Application (CMA). Each RL will launch a different job into the Orchestrator (O). This job will manage the workflow according to the specific configuration of the RL set by the LEA-A in the CMA. The LEA-A can check at any time the state of a RL. If necessary, some data can be obtained from LEA’s DB.

2. When a RL includes some reference elements against which the crawled data will have to be matched, the corresponding analysis modules (AM) might need to carry out some configuration or training before the real processing of a crawled (and normalized) file arrives (e.g. image comparison). The configuration process of a module can take different inputs, such as a set of images or numbers, and it returns a model to be used later on as a reference to be matched against.

3. The Orchestrator will launch the crawler modules using the configuration defined for RL. Original and normalized documents are stored in the Original Repository (OR) and Normalized Repository (NR) respectively. Within a RL context, the O can ask through a SOA proxy module for collected documents not processed yet. Normalized Proxy (NP) comes up with the universal resource identifier (URI) of the next document to be processed.

4. Then, the O calls the related AM and, if the AM requires configuration, the call will take two arguments: the input URI and the configuration object. To get the document, the AM can obtain it calling the Normalized Proxy Repository (NPR), which retrieves the document.
5. The AM produces an output (knowledge) that will be stored in the Knowledge Repository (KR) using the Knowledge Repository Proxy (KRP). This holds both for data and metadata. The AM will return a composed resource object to the O as a result. Collection Modules will inform the O when the crawling of the data for that RL is completed. The event takes place when the configuration of collection reaches a defined depth level in the crawling.

6. When a RL comes to an end, the Enterprise Service Bus (ESB) informs the CMA that crawling and AM processes are over. It is worthwhile to notice that CAPER does not perform a semantic analysis. The human intervention of a LEA-Analyst is required at this stage. LEA-A uses Visual Analytics (VA) module to obtain and carry out the analysis of collected documents. VA module feeds its own data repository (VAR) from the Knowledge Repository (KR). When necessary, VA module obtains a normalized document for LEA-A.

7. Data from the VAR can eventually be exported to the LEA’s DB.

Figure 2. CAPER Architecture and Workflow.

3. Discussion

CAPER includes the following six Analysis Modules: (1) Image analysis, (2) Multilingual text analysis in 13 different languages, (3) Multilingual analysis of audio content in 11 different languages, (4) Analysis of videos, (5) Integration of semantic-Web technologies and data to improve and relate analysis results and analysis of data
coming from Social Media; (6) face recognition and speaker identification biometrics. How can CAPER be handled from the ethical and legal point of view? What kind of guarantees have to be put in place to protect civil rights and comply with the new provisions and guidelines of the EU General Data Reform Package? And how to convince LEAs that personal information is an ethical good as well, and that it is not “available in publicly accessible sources and just has to be gathered and provided to the right people at the right time”?

3.1. Some preliminary answers

The CAPER Project started from the EU LEA’s impulse to have better tools to fight organized crime, complementing OSINT systems widely used in the security field (such as Maltego forensics\(^3\) e.g.). They have shown all along the Project a real interest in acting properly, under national and EU laws. This means not only complying with their regular controls, but also developing protocols, best practices and patterns of behavior that could save time and effort. Regulations are not seen as a limitation, then, but as a common tool at hand to be implemented into the CAPER lifecycle.

The ethical issues at stake are high. Therefore, we set up an independent Ethical Committee (EC), performed a Privacy Impact Assessment (PIA) with tailored questions to LEAs, and monitored (i) research developments and (ii) LEAs performances throughout the project as separate processes. We focused on collection, processing, storage, use and management of the information, and results were discussed in several Workshops and meetings with data protection experts from the national Spanish Agency, Eurojust and Europol. As a result, we look up traceability of logs on the platform, as suggested by DP Agencies, and we refined the principle of accountability, as suggested by the EC. At the same time we created links to other EU Security Projects, such as SAVASA and VIRTUOSO\(^4\).

For some time now, the main concepts of Privacy by Design (PbD) —and Data Protection by Design and by Default— have attracted a lot of attention to be used as guidelines for ethical committees and legal advisors. Ann Cavoukian recently asserted that “it is not true that privacy and security are mutually opposing” and that big and smart data “proactively builds privacy and security in” [10]. It might be true, but it is not evident in the fight against organized crime. In order to be effective, OSINT tools have been designed just for what they should be controlled: spotting as many things as possible and getting personal information about individuals and organizations.

The results recently published from the VIRTUOSO Project laid down a conceptual framework to link PbD principles to OSINT security issues [8] [14] [15]. The analytical framework to design PbD strategies in OSINT is specifically set —minimize, separate, aggregate, hide, inform, control, enforce and demonstrate— to explore two main approaches for embedding legal compliance: (i) the concept of revocable privacy (RP), and (ii) policy markup language to define Enterprise Privacy Policies. We will focus only on the former, which seems appropriate for the CAPER regulatory objectives. “A system implements revocable privacy if the architecture of the system guarantees that personal data is revealed only if a predefined condition has been met” [15, p. 681]. Authors distinguish between two variants of RP as well: (i)

---

\(^3\) https://www.paterva.com/web6/products/maltego.php
\(^4\) http://www.savasa.eu/, http://www.virtuoso.eu/
spread responsibility (one or more trusted third parties verify whether all conditions for releasing personal data have been met, and grant access or release the data if this is the case), (ii) self-enforcing architecture.

3.2. CAPER PbD strategies

Stemming from this point of view, CAPER could be considered compliant with legal and ethical provisions as a revocable privacy system with spread responsibility. It could apply to LEA’s investigators, internal controllers, and external authorities performing regular audits, such as Privacy and Data Officers and judges.

We explored several related strategies that constitute an indirect approach to PbD principles in security and surveillance matters. It could also be considered a hybrid approach between spread responsibility and self-enforcing architecture, to use Koops et al. terms [15], because we maintained PbD principles both flexible (able to be discussed by LEAs) and embedded into the system—especially use and storage limitation, and accountability—through traceability of logs and ontologies. Design means institutional design as well (not only executable programming). The notion of institutional-Semantic Web Regulatory Model (i-SWRM) leans on this assumption: establishing a self-regulatory model entails including the dimension of PbD into a social environment than can be represented as an ecosystem [7].

3.2.1. Distributed end-users’ roles

CAPER workflow is addressed in fact to four different LEA’s analysts: (i) Generic Analyst (GA) is involved in investigations that are related to any type of crimes, (ii) Advanced Analyst (LEA-AA) is intended to be an ITC expert knowing Internet threats; (iii) System Administrator is able of performing configuration tasks and monitoring the system, but without permission to act as analyst at the same time; (iv) LEA’s External User (LEU), belonging to a similar EU organization and the receiver of outbound or unbound information from LEAs’ users (Fig. 3). Information is not directly spread or shared among all the stakeholders and possible users.

3.2.2. Well-defined scenarios for specific types of crimes

LEAs’ intervention is focused on four distinct areas/objectives, which are standard among criminal investigation units. Six scenarios have been defined following this schedule: (i) monitoring in selected sources, (ii) known vs. unknown networks, (iii) open/closed data integration, (iv) focused target, (v) social web analysis, (vi) focused investigation. Sometimes a matching between open and closed information can be necessary. For example, a Generic Analyst (GA) has some evidences that there is a criminal group dedicated to sale expired medicines. There are 31 pharmacies and 20 labs and the GA tries to get detailed information about which employees own cars, houses, swag... The analyst also wants to know any kind of connections that could be established between them. Stemming from this information, new research lines can be launched for each criminal suspect. But qualifying a person as suspect is up to the GA, not to the system. CAPER tools operate only within the investigation conducted by
LEA, helping them to better define the lines of research, but avoiding any automated policing, criminological or legal conceptualization.

3.2.3. Well-defined module interdependencies

CAPER strength lies more on its capability for integrating already existing and mature technologies —i.e. on its research strategy— than on innovative (and not yet tested) ones. For example, audio analysis module processes audio files starting by audio segmentation and classification, language ID of the segments, speaker clustering and tracking and automatic speech recognition for English, Spanish, Basque, Portuguese, Italian, German and French. Besides commercial systems for Catalan, Romanian, Russian, Arabic and Hebrew have been integrated (all the languages for which the LEA have shown interest).

Besides, as we will immediately show, a multi-lingual ontology has been developed for drug crimes, following INTERPOL classification patterns.

The CAPER crawling system has three modules: (i) crawler by keyboard, (ii) by URL, (iii) by URL focusing on keyboards. But the crawler is able to convert metadata of images and videos into allowed mimetypes required by the Visual Analytics module (VA). There are just two modules in the platform that offer human interaction through an interface. One in the VA and the other is the CMA. The CMA is designed to permit LEA analysts to manage the overall system (create and configure RLs, manage security issues, configure alerts...). The alert system is closely linked with the crawling component. Alerts allow users to define events and associate tasks or actions to their occurrence. E.g. if a certain condition is fulfilled —an event pops up in a social network whose name matches a word included in a pre-defined dictionary, e.g. pedophilia— it brings about an action —sending an e-mail to LEA’s analysts with relevant information.

The task of determining the meaning of a specific term in a specific context is separated from the task of creating the ontology, and is dealt within the CAPER system by the text analysis modules, Semantic Analysis Module (SAM) and the Visual Analytics database design. Fig. 4 depicts the processing steps.

3.2.3.1. European LEAs Interoperability Ontology (ELIO)

There are two main ontologies in CAPER. The first one is ELIO, to ensure interoperability among LEAs. Both ontologies are not connected, because they address two different problems: information exchanges, and analytics. ELIO aims at create a secure reliable environment for LEAs. But it is worthwhile to note that both ontologies have a regulatory side, enabling or preventing them to perform in a certain way their investigations. There are no neutral ontologies. They have a purpose and a particular shape, and need to be regularly updated. This is a kind of warrant for citizens’ rights.
E.g. possession of drugs is a crime in UK, but not in Spain. Only trading is. Therefore, this has to be specified.

In a situation of non-harmonized criminal law systems, such as that of EU countries, the structure used in a platform built with the purpose of improving the acquirement and sharing of information between European LEAs cannot be based on a particular national structure. In addition, differences among legal frameworks, languages and police and judicial culture may create interoperability issues. Therefore, National structures need to be embedded in a supranational structure. We referred to Europol and its definition of organized crime published in its “Annual Reports”. Figure 3 shows the Organized Crime Structure (OCS) and the ELIO taxonomy.

![Figure 3. Basic ELIO taxonomy based on the OCS. Source: [13].](image)

The taxonomy has four main concepts represented as classes into the ontology structure: “Crimes”, “Techniques”, “Essential Conditions” and “Countries”. Moreover, object properties present in ELIO that connect elements among these classes are also defined: “hasTechnique”, “hasEssentialCondition”, “hasCrime” and “hasCountry”. This knowledge representation enables LEAs interoperability through OCS.

### 3.2.3.2. Multi-lingual Crime Ontology (MCO)

MCO was specifically required by LEAs because of (i) their use of criminal specific language, (ii) their collection of multilingual documents used to decipher a case, (iii) and the need to reduce data complexity in data visualization tools. As a proof of concept, a specific use case has been defined. The drugs branch from Europol classification has been enriched adding some terms and concepts in several CAPER languages (including synonyms and criminal slang terminologies). A questionnaire was circulated among LEAs to both enrich the ontology and specify the different concepts to be translated from the original Europol conceptual classification. LEAs’ cooperation in the knowledge acquisition process was crucial to translating and adding synonyms to concepts. An interoperability database was built taking into account the original Europol OCS, ELIO and the new acquired concepts. In this way, the multilingual crime ontology can automatically map concepts between different languages and be exactly
in the same crime classification branch. The tree currently includes 346 nodes. There is an ongoing work at present to complete the collection of translations, synonyms and slang terminologies for Italian, Spanish, English and Hebrew.

3.3. CAPER Regulatory Model (CRM)

We can advance now what we think is required to regulate the different parts of the system we have described. What kind of tool is needed? More than a code of best practices or a generic ethical code, structured recommendations should be able to handle together the different parts of the system, linking them not only vertically with national and European legal provisions, but horizontally too with all relevant players and organizations. This is a matter of institutional design, regulating the interface of the system both with interoperable machines and with human behavior. We call it CAPER Regulatory Model (CRM). CRM is the conceptual mechanism underlying protocols and recommendations. It accounts for hard law, soft law, multilayered governance, and cooperative behavior (best practices). Ethics stands on top of them. In purity, as it is built through the Semantic Web approach, CRM is the first step for an institutional SW regulatory model (i-SWRM) [5].

Compliance with legal norms is important, but it is not enough. Dialogue with experts and ethnographic work involving LEA’s units at the same time are crucial to understand where the problems are, and to let LEA’s investigators participate into the regulatory process. But control is exerted because binding norms apply as well. So, CRM model bridges the dialogue and negotiations of social agents with the normative requirements and conditions of the rule of law. This is why it can be implemented among LEA’s organizations and embedded into the CAPER system to regulate the use of the platform. And this is why it is a model able to regulate OSINT as an open ecosystem.

4. Conclusions and future work

We have described so far Open Source Intelligence (OSINT) and some aspects of FIPs and PhD principles, focusing onto the architecture, modules and workflow of CAPER. We have shown that implementing them into security tools cannot be taken for granted and it is not an easy task, because OSINT tools have been designed just to be effective at easing the exploratory and investigative work of LEA. From the legal point of view, interpreting security issues not as exception but as covered by the privacy and data protection rights envisaged by the EU GDPR requires some innovative ways to figure out and build self-regulatory institutions able to be embedded alike into technological environments and organizations.

We can advance at least three preliminary conclusions: (i) even in surveillance toolkits and security there are feasible ways to bridge PhD principles and citizens’ rights; (ii) PhD can be broadly understood as a form of institutional design; (iii) ethics can and should play a major role in such a modeling task.

CAPER is an ongoing Project, not finished yet. Therefore, it is still being tested by LEAs. A final ethical audit must be carried out as well. A final list of full specific recommendations on data storage, processing and management will be laid down. Metrics is an issue that must be addressed, and especially the possibility of setting legal composite indicators should be explored [7]. Stressing the difference between models
and meta-models is crucial. Comparative work on corporate governance models (such as COSO and COBIT) and related ISO standards will be taken in the next future.

References


