UNIVERSITÀ DEGLI STUDI DI NAPOLI FEDERICO II

PH.D. THESIS

IN

INFORMATION TECHNOLOGY AND ELECTRICAL ENGINEERING

A SEMANTIC METHODOLOGY FOR (UN)STRUCTURED DIGITAL EVIDENCES ANALYSIS

GIOVANNI COZZOLINO

TUTOR: PROF. ANTONINO MAZZEO
PROF. FLORA AMATO

COORDINATOR: PROF. DANIELE RICCIO

XXXI CICLO

SCUOLA POLITECNICA E DELLE SCIENZE DI BASE
DIPARTIMENTO DI INGEGNERIA ELETTRICA E TECNOLOGIE DELL'INFORMAZIONE

A semantic methodology for (un)structured digital evidences analysis

Thesis for the Degree of Doctor of Philosophy

Giovanni Cozzolino

The PhD School of Science

Academic advisors:
prof. Antonino Mazzeo[1], prof. Flora Amato[1]

[1]Department of Information Technology and Electrical Engineering
Scuola Politecnica e delle Scienze di Base - Faculty of Engineering
University of Naples "Federico II"
Italy

Submitted: 30/10/2018

A semantic methodology for (un)structured digital evidences analysis

Short abstract:

Nowadays, more than ever, digital forensics activities are involved in any criminal, civil or military investigation and represent a fundamental tool to support cyber-security. Investigators use a variety of techniques and proprietary software forensic applications to examine the copy of digital devices, searching hidden, deleted, encrypted, or damaged files or folders. Any evidence found is carefully analysed and documented in a "finding report" in preparation for legal proceedings that involve discovery, depositions, or actual litigation. The aim is to discover and analyse patterns of fraudulent activities. In this work, a new methodology is proposed to support investigators during the analysis process, correlating evidences found through different forensic tools. The methodology was implemented through a system able to add semantic assertion to data generated by forensics tools during extraction processes. These assertions enable more effective access to relevant information and enhanced retrieval and reasoning capabilities.

Keywords: Forensics, Semantic, Analysis, Event, Logging, NLP

Contents

Part I

Description of Thesis

1

Background and Motivation

Contents

1.1 Introduction

In recent years, seeing any news program, listening to a radio news report, reading a (online) newspaper, we can see how much the number of investigations and trials that have seen the use of digital evidences is growing.

People doing activities in the digital world leave pieces or echoes of themselves, just like they leave traces of themselves – fingerprints, hairs, DNA, etc. – when they move and interact with other people, places, and objects, in the real world. So, activities conducted on individual computer systems and networks routinely leave some kind of *digital fingerprint*. These may range from web browser history caches and cookies, through to deleted file fragments, email headers, document metadata, process logs, and backup files.

As Information Technology (IT) systems are involved in almost all daily activities (related to business and industrial purposes, safety systems, education or entertainment, etc.[1]), these virtual or digital traces may be deemed to be of value, for any number of reasons. They may be useful as evidence in establishing the origins of a document or piece of software, for legal purposes in determining the activities of the parties involved in a criminal case, or even as a resource for cyber-criminals looking to reconstruct information or identifying credentials on their victims. Whatever the motivation, the examination, interpretation, or reconstruction of trace evidence in the computing environment falls within the realm of Computer Forensics.

Cybercrime is the term used by the Anglo-Saxons to indicate crimes involving computer systems; more precisely, it is possible to identify two categories related to cybercrime:

1. crimes for which IT or telematic systems[i] are the object of the offence[ii];
2. crimes for which information or telecommunications systems are the tools of criminal action; more precisely, the technological tool is used to "deceive" the user[iii] or to "facilitate" the commission of an offence[iv].

In the last twenty years there has been a real explosion of computer crimes,[2] although the matter is still difficult for many legal operators. IT crimes are difficult to reconstruct because it is not clear how many authors are and where they act, often they are difficult to trace and, even when this happens, it can be difficult to understand how many times an action has been committed and what are the victims.[3]

More generally, computer evidences now come into play in almost all processes (both criminal and civil) due to the fact that an information system is, trivially, a mere container of digital data (and therefore of potential evidences in digital format): in all cases of homicide the phone records are used for the geo-localization of people, in all cases of robbery CCTV camera records are used, and so on.
A research conducted by IISFA[v] in 2012 has shown that more than

[i] The term "computer or electronic system" is used to indicate in general any kind of computer tool (computers, mobile phones, storage devices, washing machines ...). In the Budapest Convention, art. 1 computer system is defined as "any equipment or group of interconnected or connected devices, one or more of which, on the basis of a program, perform the automatic processing of data".

[ii] An example of a crime for which the offence is a computer or electronic system is the infection with viruses or worms

[iii] An example is the phishing, a phenomenon of social engineering that, through the sending of unknown scammers by fraudulent emails, encourages victims to voluntarily provide personal information, such as credit card data.

[iv] An example is the case of the copying of confidential data protected by intellectual property from the company servers to carry out unfair competition practices

[v] The IISFA (International Information System Forensics Association) is an association of professionals in the field of computer forensics (lawyers, law enforcement agencies and technicians) internationally recognized (http://www.iisfa.org/), whose primary purpose is the promotion of the study, the formulation of methods and standards concerning the Computer Forensics activities. In this regard, the Italian section of the association (http://www.iisfa.net/) is particularly active, organizing numerous refresher courses and seminars for both lawyers and technicians: among them the annual ISISFA event stands out. Forum, during which the latest results of studies on the topic are presented.

half of the computer crimes pertain to abusive access to an information system, corporate infidelity (which often results in abusive access to an IT system in order to steal industrial secrets) and offences of child pornography committed through the use of file sharing clients on a peer-to-peer network.[4]

The youngest of forensic sciences, Computer Forensics, comes into play, which deals with the preservation, identification, extraction and documentation of the digital evidences. Like any other forensics science, Computer Forensics is about the use of sophisticated technological tools and procedures that must be followed to ensure the preservation of the computer evidence and the accuracy of the results regarding its elaboration.[5] Generally speaking, it is a question of identifying the better ways to acquire the evidences without altering or modifying the computer system on which they are located and ensuring that the evidences acquired on another medium is identical to the original ones.[6]
Unlike *Computer Security*, that deals with the protection of an IT system, Computer Forensics informs after the violation has occurred or, in general, that the system has been actively or passively involved in a crime; the aim is the examination and documentation of the data contained within the informatics findings to reconstruct the events that have occurred: a detailed analysis allows to know the activities, tastes, thoughts of the user in order to conduct the investigations in the right direction and acquire evidences related to the life of its user.

After analysing in detail the methodological and applicative aspects of forensics, this work will focus on the definition of a specific methodology for forensic analysis of information systems. Moreover, an application of semantic techniques to enrich the analysis and correlation process of a forensics investigation is presented. The addition of semantic assertion to data generated during various analysis phases, should improve the presentation results, enabling more accurate correlation of traces and more powerful searches.

1.2 Computer Forensics

Computer Forensics is a branch of forensic sciences, ie those of medical, biological, electrical, mechanical, electronic, computer science, etc. that can provide "objective" techno-scientific evidences as elements of judgement both in the case of civil and criminal proceedings. It comprehends the practices of collecting, analysing and reporting relevant information, called *digital evidences*, found on computers and networks, in such a way that this process is deemed admissible in a legal context – whether that be as evidence in a criminal or civil investigation, or as documentary proof in a commercial or private setting.

Since computers, mobile phones, and the internet represent the largest growing resource for criminals, Computer Forensics has assumed a key role in the law enforcement sector. With cyber-crimes offering a high-yield and relatively low risk opportunity that doesn't require physical violence, law enforcement agencies are now continually engaged in digital forensic activities to curb the exploits of fraudsters, identity thieves, ransomware distributors, and others in the cyber-criminal ecosystem.

Therefore, the aim of the forensic sciences is to analyse facts inherent in digital systems that result in violations of civil, criminal and/or internal regulations and to produce scientific evidences. These evidences can be introduced into a judicial process in the form of means of proof, on which the judge will base his conviction and therefore his decision. As this definition suggests, digital forensic operations may be applied in commercial, private, or institutional applications, and in the context of cyber-security.

The analysis must lead to obtaining digital evidence sources, ie data that can testify, as absolutely as possible, the facts related to the violation to be ascertained, highlighting the times, the events, the systems used, the lines and the means of communication, as well as identifying elements such as user names, IP, MAC address, password, biometric codes, etc. The evidence sources acquired in this way will be used as support for decisions through their admission and discussion in court. The essential characteristics of conducting a Computer Forensics analysis are essentially three:

1. use of scientifically accepted procedures;
2. determinism and correct timing of the activities carried out;
3. repeatability of technical assessments.

Computer Forensics must satisfy both technical/methodological and legal/juridical needs. The use of scientifically accepted procedures is of paramount importance during the analysis; in fact, following procedures that are not scientifically contested (and difficult to contest) will facilitate the use of the means of proof in court, protecting the evidence, and the results of the analysis, from possible disputes. At present, there are no codified procedures for conducting a forensic analysis, whereby each operator and each police force and/or private entity have developed internal protocols to be followed, which are constantly evolving and improving.

1.2.1 Definition

The first complete definition in Italian literature was made by Maioli[6] that qualifies the Digital Forensic science as "*the discipline that studies the set of activities that are aimed at analysing and solving cases related to cyber-crime, including the crimes realized with the use of a computer, directed to a computer or in which the computer can still represent a source of proof*".

The purposes of Computer Forensics are the preservation, identification, acquisition, documentation and interpretation of data on a computer. Generally speaking, it cares about identifying the best ways to:

- acquire the tests without altering or modifying the computer system on which they are located;

- ensure that the evidence acquired on another medium is identical to the original ones,
- analyse data without altering them.[7]

The forensic computer science includes the verification activities of the data storage media and of the computer components, of the images, audio and video generated by computers, of the contents of archives and databases and of the actions carried out in the telematic networks.

In reality, this definition, albeit very detailed, would seem to relegate forensics to the criminal field only, while it is now common to use Computer Forensics techniques also in the civil process, in the field of labor law, in administrative law or in internal corporate investigations.

Another definition elaborated in doctrine that has been widely disseminated sees forensic as "the science that disciplines the methods for the preservation, identification and study of information contained in computers or information systems in general, in order to to highlight the existence of useful evidence to carry out the investigative activity"[8] and that "*studies the value that a data related to an information or telematic system can have in the juridical field*".[9]

During the years, several American Computer Forensics researchers, proposed different definitions of the topic of study of the "Computer Forensics", which converged substantially on some aspects constituting the common denominator: the treatment of digital data for forensic use, and therefore investigative and judicial, had to be carried out according to scientific principles, certain technical methodologies and in accordance with the procedural rules.
For Caloyannides, the Computer forensic "*... is the collection of techniques and tools used to search for a means of proof in a computer*".[10]
For Vacca, "*Computer forensics involves the preservation, identification, extraction and documentation of computer evidence stored as data or magnetically encoded information (...) Computer forensics also referred to as computer forensic analysis, electronic discovery, electronic evidence discovery, digital discovery, data recovery, data discovery, computer analysis, and computer examination, is the process of methodically examining computer media (hard disk, diskettes, tapes, etc.) for evidence*".[11]

For Kruse II and Heiser, "*Computer forensics (...) involves the preservation, identification, extraction, documentation and interpretation of computer data*".[7]
For Marcella and Greenfield, "*Computer Forensics (...) deals with the preservation, identification, extraction, and documentation of computer evidence. (...) Like any other forensics science, computer forensics involves the use of sophisticated technology tools and procedures that must be followed to guarantee the accuracy of the preservation of evidence and the accuracy of results concerning computer evidence processing*".[5]
Finally, Casey focused the study of "digital evidence", defined as "*Any data filed or transmitted through a computer to support or deny a theory about how a crime was committed or which directs crucial elements of the crime as the end or the alibi*",[12] overcoming the category of Computer Forensics.

1.2.2 History

The birth and development of Computer Forensics as an independent branch is closely related to the evolution of the information society (Information and Communication Technology, ICT). Until the late 1990s, what became known as Computer Forensics was commonly termed "*Computer Forensics*". Computer Forensics was born in the United States of America as the discipline that deals with the study of information technology for judicial purposes, and since the 70s has known a significant theoretical and practical development. In those years, the American social and productive landscape had radical changes as a result of the growing and widespread diffusion of information technologies in the most disparate social contexts: from the communicative, to the productive, to the personal. At the same time, the diffusion of computer technology also determined an increase in criminal activities in which the IT devices, at the time mostly computer sometimes connected to the network, constituted the object of a crime (*computer crimes*), or an instrument for the commission of a crime (so-called *computer related* *crimes*), or, more frequently, an instrument storing relevant data relating to the facts under investigation (*computer evidence*).

Computer Forensics assumed a great relevance in the American investigative panorama, even more after the proposal of a standard in the "*Proposed Standards for the Exchange of Digital Evidence*", made by the Scientific Working Group on Digital Evidence (SWEDGE) and by Organization on Digital Evidence (IOCE), published in 1998. Thereafter, the FBI introduced a list of techniques for processing digital devices for justice purposes, as part of the "*Handbook of Forensic Services*", published in 1999 on her own website. In the section of the Manual concerning the insurance and the examination of the means of proof, in addition to the traditional investigation techniques, a new series of services and techniques for the treatment and examination of IT findings, formalized as "Computer Forensics" services, was listed, organized according to standardized methods and tools.

Related to the new requirements, the FBI made available to authorities and agencies, new specific competences for the treatment of digital devices, such as data seizure, data duplication, data preservation, data recovery, document searches, media conversion and expert witness services. In particular, in the "*Manual of Forensic Services*" an articulated series of services relating to data and digital devices was exposed: from the "Computer Analysis", for which technicians could perform operations of analysis of the device in relation to the content (*"the exams can determine what types of data are contained in a computer*"), to the comparison of data ("*the exams can compare files to know the contents of documents*"), to the reconstruction of data creation timeline ("*the exams can determine the time and sequence with which the files were created*"), to the extraction and recovery of files deleted from the computer ("*files can be extracted from the computer*"), to the file format conversion ("*files can be converted from one format to another*"), to keyword searches ("*files can be searched by word or phrase and the results can be stored*").

In 2001, a research on computer crimes stated the need to spread the awareness among the local state law enforcement agencies of the problems and practices of Computer Forensics. So, the National Institute of Justice (NIJ), the Research and Development agency of the US Department of Justice, published a guide for the inspection activities. In 2002, the Department of Justice published the second version of the Manual "*Searching and Seizing Computers and Obtaining Electronic Evidence in Criminal Investigations*"[vi], updated to the provisions of the Patriot Act approved after the events of 11 September 2001, while the IOCE presented to the G8 a first set of principles on procedures related to digital means of proof.

In the same time, Computer Forensics principles have been widely applied in large corporate sectors (industries, banking, insurance, healthcare, ICT), professionals (lawyers, private investigators and consultants) and in non-governmental organizations, within which studies were performed for the enhancement of collection and storage of digital evidence during an event of a security breach.

1.2.3 Topics, goals and challenges

The wide diffusion of computer technology has favoured the development of a whole specific branch of the investigative sciences, concerning the study of the computer as an archive of information useful for the reconstruction of the facts of the relevant process. A growing demand for the analysis of digitized data for investigative purposes has been derived, thus determining the development of the Computer Forensics techniques, characterized both by the need to proceed by adopting techniques and tools that allow compliance with the principles recognized by the American process system, as well as the application of information technology to the investigation activities. In particular, the origin of Computer Forensics derives from the investigative practice developed on computer devices, properly considered as crime tools, but also formidable archives of useful information for the ex post reconstruction of a crime scene. The attention to computer systems analysis has been facilitated by the

[vi] Searching and Seizing Computers ad Obtaining Electronic Evidence in Criminal Investigations, 2002, `http://www.finer-bering.com/GULAW_PDFs/s&smanual2002.pdf`.

typical culture and tradition of US investigative reality, historically oriented to the application of scientific principles, tools and methodologies to the activity of acquisition of information useful for the reconstruction of the facts relevant to judicial purposes.

In this regard, the fundamental principles of the American (criminal) process, such as the *legal and procedural fairness*, the *timely notice of the charges*, the *guiltness to be proven by legally obtained evidence*, and the *verdict must be supported by the evidence legally presented*, must be recalled. From these principles derive some corollaries.
First of all, the investigator must observe with rigour scientific and methodological methods during the acquisition of the means of proof to be presented for trial, to avoid that the trial counterpart undermines their reliability and probative value.
Secondly, the part proceeding (Plaintiff) has the responsibility to exhibit means of proof against the Defendant and, therefore, to effectively prove the assumptions in support of his own procedural position, or to prove that the counterpart's assertions are wrong, thereby undermining the representative effectiveness of the adversarial means of proof.
Thirdly, it is the duty of the defendant to exhibit the evidence for his own discharge and then to effectively prove the assumptions in support of his own procedural position, or by highlighting the erroneous assumptions of the counterpart undermining the representative effectiveness of the opposing means of proof.

After the development of networks, Internet and telecommunications technologies, this approach has also been extended to data transmission systems. This has led to the development of new techniques and tools of data capturing for legal purposes, according to the procedures established by the procedural rules and other fundamental rights such as, for example, the freedom and secrecy of electronic correspondence, telematic and telephone communications, as well as confidentiality in general.
This new area is called *Network forensics* and studies the rules, methods and tools for investigative and judicial manipulation of data transmitted on digital networks.

The affirmation of Computer Forensics over the years has been progressive and constant, investing all the devices produced by the technological evolution, and articulating itself in other branches such as *Mobile forensics*, concerning the investigation of mobile phone devices, tablets, and other mobile devices like GPS and, overall, smartphones. The last ones, thanks to their mode of operation, sum up the characteristics and the problems of all the other digital devices and systems, and, in relation to their diffusion, now far superior to that of computers, are affirming as a reference compendium of all the Computer Forensics problems.

The incessant progress of information and communication technologies promotes many changes in methodologies adopted to notice, collect, manage e analyse evidences during an investigation.[13] The goal of Computer Forensics is not only collection, acquisition and documentation of data stored on digital devices, but, above all, it is the interpretation of evidences. Information correlation, through a logical-inferential reasoning process, is a crucial phase in forensics analyses, because examiner must take into account of all kinds of information (considering the broadest meaning of term), such as context of investigation, acquired clues, investigative hypothesis, and many other elements, not necessarily in digital form. Correlation of these information is the only mean to allow for the contextualization of digital evidences, promoting them as clues.

In the procedural stage it often happens that the operations of collecting and storing digital data are challenged and these problems are due to the fact of having to work with something not tangible and invisible to people who do not have specific knowledge on the subject (an example can be a trace of unlawful activity found in a log file). Moreover, it should be borne in mind that if once the data were saved exclusively in digital data storage devices connected to the computer physically used by the end user, today a lot of information is de-localized from a single location and stored in a network. Therefore, if a crime is committed even partially through the use of a network, the tests are distributed on different computers and in many cases it is not physically possible to simultaneously collect all the hardware. At the same time, the presence of a network can induce redundancy, a phenomenon for which a data not available at one point in the network could be traced to another.

Taking into account the previous considerations, experts have to face many challenges during the different phases of forensic investigations:

- the admissibility of acquired data must be preserved;
- the volume and heterogeneity of data requires information extraction;
- the lack of standard format for file produced by different practices and tools leads to integration and correlation problems;
- the unification events from multiple sources or devices.

1.2.4 Branches

The excursus of the various definitions, made in Section 1.2.1, confirms the strictly interdisciplinary nature of Computer Forensics, whose theories and practices could not leave aside the profound inter-correlations between Information and Communication Technology (ICT) and law, in particular criminal law and criminal procedure. In the meantime, Computer Forensics has evolved and diversified, turning its attention to the various sectors offered by technological evolution.

In fact, depending on the peculiarities of the devices being analysed or how data is collected, forensics can be better classified in a certain number of branches, whose composition may change over time based on technological development.

Disk forensics Disk forensics is the main and oldest branch of Computer Forensics. For this reason it is common for the term "disk forensics" to be replaced by "Computer Forensics" or by the more general "forensic computer science", and its variants in English. The disk forensics deals with the analysis of computer media (hard disk, solid state disk, USB sticks, CD-ROM ...) that can follow different purposes depending on the role assumed by the same.
The first cases of disk forensics date back to the Seventies, mainly for

the fight against financial crimes. In 1978 in Florida the issue of the Computer Crime Act[vii] formalized its birth with copyright and child pornography crimes.
Only since the 2000s has emerged the need to define standards and guidelines that can achieve an adequate level of reliability: to this end, the National Institute of Justice has published several guides[viii], mainly intended for law enforcement agencies, aimed at defining the correct procedures for forensic analysis of digital evidence and for the protection of the crime scene.

Network forensics The information systems are not confined to the now small space they occupy but, having a network connection, they communicate with other systems: especially in recent times and thanks to the availability of high-speed connection, the possibilities offered by cloud computing are considerable. For example, it allows remote data storage. The forensic analysis therefore undergoes all the effects brought about by the development of technologies based on computer networks, since the information power provided by the single computer tends to decrease. Therefore, attention must be extended to remote or even virtual systems. Without network forensics, there is now the risk of omitting fundamental data for an investigation.
The network forensics has as its object the forensic analysis techniques typical of a data transmission system. This discipline presents a considerable complexity as it is forced to follow the data path on a multitude of systems: for example, in the case of e-mail, the e-mail sent by the user A to the user B can be found not only on the IT systems in use to A and B, but also on the intermediate servers crossed.

Cloud forensics The term cloud computing refers to a set of technologies that allow a Cloud Service Provider to provide to its customers a variety of data storage and processing services through the use of virtualized hardware and software resources distributed on the network. According to the definition of the National Institute of Standards

[vii] `http://www.clas.ufl.edu/docs/flcrimes/chapter2_1.html.`

[viii] `http://www.nij.gov/topics/forensics/evidence/digital/pages/welcome.aspx`

and Technologies (NIST)[14][ix] the main features of the cloud computing paradigm are: the availability of broadband network connections, the flexibility in the provision of services, the supply of consumer self-service services and the sharing of resources among multiple users.

Cloud forensics is one of the most popular and debated topics of the last years:[15,16] the distributed processing of data and the lack of physical access to the mass memories, sited on the servers distributed in the cloud, represent a serious problem for the investigator who must review the techniques of disk and network forensics from the point of view of the new phenomenon. An example of a cloud service is the email: users does not need to install and configure any software on their computer, having the ability to access data remotely from other systems through an access interface such as a browser.

Mobile forensics Mobile forensics has as its object of investigation mobile devices, and more in detail, it can be divided into:

- SIM card forensics, in which the analysed object is the content of the SIM card;
- Mobile Handset Forensics (or Cell Phone Forensics), in which the analysed object is the content of the mobile device (smartphone, tablet, mobile phone ...);
- Memory Card Forensics (or Removable Media Forensics), in which the object analysed is a memory card that could be used to extend the memory capacity of the memory device; in this case, the analysis techniques are borrowed from the disk forensics;
- Cellsite Forensics, in which the survey object are telephone records that contain the traces left by mobile devices along the network, allowing geo-localization of devices in time and space.

[ix] `http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf`

Embedded forensics Embedded forensics does not have its own circumstantial dimension due to the extreme diversification of systems: this branch deals with the forensic analysis of digital systems that can not be classified in the previous categories. In fact, there are more and more specialized digital tools that may contain useful tracks for the purpose of an investigation: think for example in video game consoles (PlayStation, Xbox ...), intrusion detection systems, computerized systems for car management, black boxes and so on. Among all, this is probably the branch that has the greatest criticality because it is extremely difficult to retrieve information on the acquisition techniques and forensic analysis of the same.

Multimedia forensics Following the growing popularity of images and videos in digital format, the need to analyse such data to extract useful information to be produced as evidence in a process becomes increasingly frequent: the typical case is that of video surveillance videos recording the criminal event but they are difficult to use because of poor quality (resolution, noise ...)[17] The multimedia forensics is responsible for extracting information useful for the purpose of a survey by images and video, also verifying the authenticity and provenance of the digital evidence. The use of image and video forensics techniques is necessary when it is necessary to determine the counterfeiting of a multimedia data, the evaluation of the reliability of the data (in support of an accusation or an alibi), to improve the quality or to derive information (for example, the speed of a car in a video rather than the equipment used for capture[18]).

1.3 Computer Forensics as forensic science

The scientific method is the way in which science proceeds to achieve a knowledge of objective, reliable, verifiable and shareable reality. On the one hand, the collection of empirical evidence is carried out through experimental observation, on the other hand it has to be done the formulation of hypotheses and theories to be submitted to the examination of the experiment to test its effectiveness.

Computer Forensics, as a branch of forensic science, requires the knowledge of legal norms and of different operational protocols; the results are not just for research, but there are jurisprudential constraints and effects that reflect on personal freedom. Therefore, those who deal with Computer Forensics should mainly be born as a "man of science", with appropriate knowledge of some legislative aspects, who must be able to explain - both in his own reports and in court depositions - the results of the tests he carried out on findings, in order to communicate effectively also with the interlocutors who ignore these scientific aspects.

The forensic scientist contributes to the investigative phase by suggesting and confirming hypotheses with the aim of achieving a faithful reconstruction of what happened; in the trial phase he intervenes expressing opinions and evaluations and helping the judge, who remains the *peritus peritorum* anyway, to take the final decision.

1.4 Scientific evidence

In recent years, scientific evidence has been overwhelmingly entering the courtrooms (not only criminal). The scientist uses his knowledge, his studies, his experiences to produce statements of a general nature. Instead, jurisdiction operates in exactly the opposite direction because the judge uses the general knowledge and statements to state something about specific facts.

This observation has very important implications because no scientific law, no universal utterance, no matter how certain and cogent in its implications it is, can tell us everything about the particular case that we are asked to solve in a court of justice. In fact, that particular case is indeed a *unicum*, while the scientific law speaks of a class of facts.

At the time of admission, recruitment and evaluation, knowledge tools drawn from science and technology are used, ie scientific principles and methodologies, technological methods and technical tools whose use requires expert skills.[19]

The problem is the verification of how the use of scientific laws, always changing due to continuous technological progress, can occur in compliance with the principles of due process and in particular the right of defence, which can not be ignored, pointing out that the conviction can only be imposed if the accused is guilty beyond reasonable doubt. The adjective "reasonable" means "understandable by a rational person" and therefore through a motivation that makes reference to logical arguments and that respects the principle of non-contradiction.[20]

In terms of scientific proof, therefore, the judge should give an account of the critical evaluation of the degree of control and reliability of the scientific method, the existence of critical reviews by experts in the sector, the indication of the margins of known error.

1.4.1 Digital evidence

Digital evidence is the main element of any forensics process. Data is elementary facts, encoded information that needs an interpretation to take on meaning and provide knowledge. The computer data is a representation in a binary system of sequences of bits not immediately comprehensible to humans, so it requires a series of operations through which a transformation is performed that can lead to different results (shown on the monitor in textual representation or as a video, but also as an image printed on a piece of paper).
By its nature, the digital data is:

- immaterial, for which it needs a suitable support to contain it such as CD-ROM, hard disk, USB sticks;
- volatile, as it can be dispersed quite easily;
- deteriorable, modifiable even in an anonymous and / or involuntary manner;
- reproducible in a potentially infinite number of copies.

Digital evidence can be considered any data allocated to a particular device or transmitted by IT and telematic systems that may have some procedural relevance.[21] Any data used to support or refute a thesis in order to define how an offence has occurred, or to establish the intention or the alibi is to be called scientific evidence in digital format. According to ISO/IEC 27037:2012[x], a digital evidence is every kind of information that can be stored or transmitted in a digital form and that can be considered an evidence.

Literature offers many definitions of digital evidence, such as:

- any data stored or transmitted using a computer that support or refute a theory of how an offence occurred or that address critical elements of the offence such as intent or alibi;[12]
- information of probative value stored or transmitted in digital form;[22]
- any digital data that contain reliable information that supports or refutes a hypothesis about an incident.[23]

Casey proposes a definition of digital evidence according to which the digital proof is "any data stored or transmitted using a computer that support or refute a theory of how an offence occurred or that address critical elements of the offence such as intent or alibi".[12] Next to this it puts three other definitions:

1. any data that can establish that a crime has been committed or that can provide a link between a crime and its victim or between a crime and the person who committed it;

[x] See Section 3.2.1

2. any information with a probative value that is stored or transmitted in digital form[xi];

3. information transmitted or stored in binary format that can be used in court[xii].

From a purely technical point of view it would be appropriate to distinguish digital evidence from electronic evidence in order to put the emphasis on data rather than on electronic devices that contain them; this distinction is acknowledged and explained in the ISO/IEC 27037:2012 standard where are defined separately:

- the "digital data storage device" is defined separately, as an object containing the digital data of interest;

- the "digital device" and the "peripherals", the latter tools that allow the digital data to be processed, received or sent to the external world, but still lacking long-term storage capacity.

The high risk of deterioration makes the digital proofs easily alterable, damageable or destructible, sometimes even by the same investigators if not properly prepared to carry out a survey of computer forensics. The intrinsic weakness of digital evidences make them easy prone to alterations or modifications, even from examiners that, if not experienced, may compromise and contaminate the *scena criminis* status.[24] Thus, its rigorous and authenticated management is of extremely importance.

In[25] author has identified three basic properties of digital evidence:

Fidelity refers to Chain of Custody forensic principle, and involves the adoption of techniques that grant the integrity of evidences through the process;

Volatility is related to the nature of the support were evidences are stored (disk, memory, registers, etc.): this property affects considerably the acquisition and analysis of evidences;

[xi] Definition proposed by Standard Working Group on Digital Evidence (SWGDE).

[xii] Definition proposed by International Organization Computer Evidence (IOCE)

Latency involves the presence of additional information to contextualize and interpret a digital encoding. The focus of current work is on this property.

1.5 The "Repeatability" of the investigations

With the expression *Repeatability* of the investigations we mean the possibility to completely re-run the analyses on models absolutely identical to the original; therefore repeatability is configured as a key factor, especially when working on crimes, and results must be presented in trial for an indictment. In other words, we assume the possibility of carrying out a post-mortem study of the systems, completely limiting the interactions with the outside world. The digital evidence sources validly obtained from digital forensics procedures, from the technical-scientific point of view, should possess the following characteristics: **integrity**, **consistency** and **documentation**.

Integrity For *integrity* of the source we mean the possibility of freezing the evidence found so as not to allow alterability over time. Once the source has been collected, this procedure entails the proper preservation of the data.

Consistency The *consistency* of the source consists in demonstrating that the information taken during the forensics analyses highlight facts logically correlated with each other in a temporal sequence that is coherent with the other procedural information also coming out of the digital environment.

Documentation The availability of the *documentation* details and certifies the adopted procedures, describing all the steps involved in the extraction and identification of the sources of evidence, including the methods and procedures used, the problems encountered, the ad-hoc solutions adopted, the tools with related version and license, the personnel who conducted the analysis with the relevant scientific qualifications, the chain of custody of the finds from the crime scene to the laboratory, etc.

1.6 Differences between Forensics and Security

Finally, it is intended to highlight how there is a difference between Computer Forensics and IT security, although the two areas of activity are linked. IT security is concerned with protecting data and the functionality of a system, thus representing an obstacle for computer forensic because the difficulty in accessing data is high both for those who intend to commit illegal activities and for those who intend to use the same data to investigative and trial purposes. The acquisition of the data contained within the system to be analysed will therefore require the use of hacking techniques that actually entail the violation of the system. Computer Forensics intervenes after the system has been violated and is responsible for preserving the evidences to document violations of security, as well as identifying potential evidences for crimes that did not result in the violation of an IT system.

2

Legal Aspects of Digital Forensics

Contents

2.1 Information, data, bit

If the actors of the process want to collect digital data, that have a high representative capacity and from which is possible to gather information useful for the purpose of the procedure, then their suitability to constitute an evidence and the related techniques for their treatment must be subjected to scientific methodologies and techniques studied by Computer Forensics.
Nevertheless, the reconstruction of the juridical aspects characterizing the procedural activity as IT object can not ignore the preliminary definition of the fundamental concepts of informatics which constitute the indefeasible premise for a correct scientifically based approach to the issues that we intend to tackle in the continuation of the discussion.

In European and Italian legislation, we often meet the terms of *data*, *information*, *digital document*, but rarely the term *bit*. These are the terms used in the jurisprudence invested in its application of the law, as well as by the authors who treated the subject in the doctrinal context. In truth, these terms are used in a non-univocal way and often as synonymous, causing confusion in less serious cases and antinomies in some striking cases. The origin of the nonchalant use of these terms is undoubtedly identified in the national legislation which, from the way it approaches or uses them, reveals to ignore the profound ontological differences between the terms, provoking a heterogenesis of purposes with respect to the *ratio legis*. For this reason, the attempt to differentiate the scope of the terms is essential. The first issue to be addressed concerns the postulate of forensic informatics, related to the identification of the object of the study and the techniques of computer forensics, ie the data.

Data By the term "*data*" we mean the original, uninterpreted representation of a fact, phenomenon or event, carried out through symbols

(numbers, letters, signs). A datum can be analogical, that is represented by distinct and continuous or variable symbols with continuity (for example the temperature of a body detected by a mercury thermometer), or digital, that is represented according to a code of well defined and discontinuous symbols or signals that do not change continuously and can be read by a computer.
Data must be kept separate from the information: the first is an "*original, that is uninterpreted, representation of a phenomenon, event or fact, performed through symbols or combinations of symbols (numbers, letters, signs) or any another form of expression (vocal, visual, etc.) linked to any medium (paper, magnetic or optical disks, photographic film, etc.)*". As such, the data does not constitute itself an information, which, for be inferred, needs a further activity that relates to the data.

Information The information "*... derives from a datum, or more likely from a set of data, that have been subjected to a process of interpretation, deriving from knowledge oriented by a subject, which has made them meaningful for the recipient, and really important for the intended purpose*".
The previous definitions make it possible to identify a series of differences between data and information:

- the data is the input of an information system, information is the output of the data;
- the data are facts and figures not processed, information is a data processed;
- the data does not depend on the information, the information depends on the data;
- the data is not specific, the information is specific;
- the data does not carry meaning, the information must carry a logical meaning;
- the data is raw material, information is its product.

However, the definition of data contains three important elements:

1. is a representation that is not interpreted, and therefore objective, of phenomena, events or facts;
2. is realized through symbols or combinations of symbols of another sensorially appreciable expression form;
3. is sometimes linked to a support (called *medium*) that can be of different materials, suitable to fix the data or to transmit them to a recipient.

Therefore, the peculiarity of the relationship between data and the media lies in the fact that the data can be contained on an appropriate medium and must be composed of symbols that the recipient can understand. The usefulness of such data construction, when linked to a support that store it, derives from the possibility of reading it from the support thanks to a device that can process it.

Information, on the other hand, is characterized by the fact that its usefulness is revealed only to those who deal with the specific subject and can interpret the symbols constituting the data. Therefore, the information, precisely because it is the result of an interpretative process that presupposes an intellectual activity, disregards the support and instead depends on:

- the recipient or the person performing the interpretative activity, in fact the same sequence of data can be interpreted by different people in different ways;
- the context and the time in which it is created;
- the place of creation or destination;
- the source of original data.

Therefore, information is closely linked to the context in which it is created, used and enjoyed.

Bit In the Information Theory field, the term *bit* indicates the smallest existing and imaginable information unit that a system can handle. The term also has a double notation: the first one refers to the smallest unit which, according to the binary system, can have a value of 0 or 1;[26] the second, indicates *(a) smallest unit in binary number notation , wich can have the value 0 or 1 (b) smallest unit of data that a system can handle (...).*

In computer science the use of the term, derived from BInary digiT,[26] identifies the minimum amount of information and the elementary choice between only two possibilities such as, for example true/false, on/off, present/absent voltage and any other dual dimension, such as yes/no, white/black, full/empty, left/right. A set of bits is called a *string*. An eight-bit string is a *byte*, which is a sequence used to encode a single alphanumeric character on a computer. The set of conventions and rules with which pre-set configurations of bits are used to represent univocally the numbers, letters, symbols, is called *encoding*, while coding is called the operation of transforming letters, numbers, symbols into strings of bits that a computer (or other device) can store or process. A set of bytes that constitute a logical agglomeration (text, sounds, images, movies ...) is called a file. Digitalization is defined as the process of transforming an analogical data into a digital, that is its representation through binary coding.

2.1.1 The material nature of the bit

The relation between bits and the medium on which they are stored rises a question of considerable importance for the definition of the technical methods and the rules for the correct processing of data for legal and procedural purposes: are bits material or immaterial? From a strictly legal point of view, the question is of broad relevance because the implications deriving from the definition of the "substance" of which the bits are made constitutes the precondition for the correct juridical qualification of IT facts. In fact for the jurist called to apply the rules to concrete cases, it is important to establish whether an object has a material or immaterial dimension. In fact, some rules, for their application to the concrete case, presuppose the materiality of the object as an indefectible element; other rules, on the other hand, when applied to intangible objects, imply the direct or indirect adoption of initiatives aimed at safeguarding the completeness and reliability of information deriving from the bits. The problem of the materiality or immateriality of the bit moves above all on the ontological-philosophical plane, so much so as to constitute for some a fundamental element to elevate the bit to a paradigm and an instrument of knowledge of reality.

In a fundamental essay on digital themes, to the question "*What is a bit?*", the author gave an unequivocal answer: "*A bit has no color, size or weight, and can travel at the speed of light. It is the smallest atomic element of information' DNA. It is a way of being: yes or no, true or false, up or down, inside or outside, black or white. For practicality we say that a bit is 1 or 0. What the 1 or 0 means is another matter. In the early days of the computer era, a string of bits generally represented numeric information (...)*".[27]

However, on a more practical level, the question of the materiality of the bits imposes to consider the technical aspects concerning the physical dimension of the bits in relation to the medium on which they are stored. The types of memories are classified in primary (or central) and secondary (or mass). At the state of technological progress, a bit, in its static phase, can not disregard the memory on which it is recorded. Therefore, the physical size of a bit also depends on the memory technology, primary or secondary that is, and on the type of material of which it is composed.

The bits, besides being statically stored on memory, can also be transmitted by packet switching techniques. In this situation the bits are transmitted through a network by associating each bit with a physical phenomenon that can be reproduced remotely through a transmission means. In this case, the bits are subjected to a further rather complex coding. Depending on the type of physical phenomenon used, the transmission media used in the networks are currently divided into three categories:

1. electric means: these are the means used in the past and which, exploiting the properties of the metals to conduct electrical energy, allow data transmission by associating the bits with particular voltage or current values, or certain variations of these quantities;

2. radio waves (e.g. wireless means): in these technologies, the physical instrument used to transmit the data associated with the bits is the electromagnetic wave, that is the combination of an electric field and a variable magnetic field, which has the property of propagating in space and reproducing at a distance a electrical current of a receiving device (antenna);

3. optical means: optical fibers and last lasers, ie very recent transmission technologies that exploit the physical phenomenon of light.

Such transmission means are all based on the transport of some form of energy which encodes the bits (so called *signal*) to which the physical system crossed opposes, thus determining an attenuation of the transmitted energy. This attenuation is also different depending on the frequency, so that for each physical system there will be a bandwidth, that is the set of frequencies that can be transmitted without excessive attenuation.

Returning to the initial question, that is, if the bits are material or immaterial, the discussion just carried out leads us to believe that, at the current state of technology:

1. the bits, and therefore the encoded data, are read from the memory and stored on the same (or different) by modifying the material of which the support is made of;

2. the bits processed by the system are subsequently represented by the system;

3. the bits transmitted, regardless of the technology and the technique used, are disgusted by any material support and therefore are immaterial.

2.1.2 Preserving the integrity of bits

Related to the material or immaterial nature of the bits, there are some considerations that must always be kept in mind for a correct setting of the method of data processing for procedural purposes:

need for a memory (hard disk, floppy disk, flash memory): in cases where the bits are in the static phase, their physical size requires that you can not assume documents that for the usability do not need a memory; when instead they are in a dynamic phase, the correct formation of the information deriving from the bits imposes that all the bits (or most of them) that are part of the flow are processed or acquired;

reproducibility in infinite number of copies : thanks to their digital coding, the strings of bits can be replicated equal to themselves in an infinite number of strings always perfectly equal to each other, as verifiable by calculation and comparison of the respective hash[i].

volatility : the technology on which primary memories are based, also called volatile memories, requires continuous power supply, so that the bits processed by them are not permanently archived, but are lost if the power supply is stop;

deterioration : the bits stored on the secondary memories only apparently seem to be destined to remain available for a long time; in reality, the ability to remain legible over time depends on the technology used for the construction of secondary memories. On a strictly physical plane, the secondary memories are affected by the intrinsic limits of the material which they are composed. These parameters condition the degradation of the bits and therefore of the data and of the information that can be deduced, so the bits stored on them are subject to deterioration and illegibility phenomena that go under the name of "digital obsolescence". This phenomenon manifests itself for all digitized documents, until they are unusable;

[i] https://it.wikipedia.org/wiki/Hash

(almost) anonymous modifiability : by virtue of the mediation of the hardware and software in the process of creating and storing bits, they do not have physical elements that make it possible to trace unequivocally the operator who created, modified, transferred or deleted them, or to trace the characteristics that establish a relationship between bit and operator, a relationship that can be inferred on the basis of information that can be deduced from the specific physical characteristics of the bits.

These considerations not only constitute the minimal ontological and physical basis for a correct juridical qualification of the procedural facts relevant to the matter, but allow the following considerations to be made regarding the best techniques for the acquisition of the bits and therefore of the data both in the static phase (Computer forensics) and dynamic phase (Network forensics).

The findings also confirmed one of the postulates of computer forensics, ie the need to preserve the integrity of the bits that make up the data where it is intended to draw information to be used during the trial. The need to apply all the Computer forensic principles in processing data for trial use have as sole purpose preserving the integrity of the sequences of bits in order to guarantee that the information derived can be reliable for all the parties.

2.2 Proof and Evidence

As mentioned above, digital forensic studies the rules concerning the processing of digital data for trial use, and in particular their relevance for probative purposes in a legal procedure. One of the terminological questions to clarify concerns precisely the definition of terms such as *proof*, *means of proof*(147), *evidence* and *digital evidence* that, although frequently recurrent in the field of computer forensics, are erroneously used as synonyms.

From a legal point of view, and according to the best classification148, the term "*proof*" can have at least four different meanings: source of proof, means of proof, evidence and evidential result.

Source of proof means "*everything that is suitable to provide appreciable results for the judge's decision, such as a person, a document, a thing...*", and sometimes the source of proof is the source of the evidence.

Means of proof is the instrument with which a trial element is acquired in the process which is used for the decision, such as a testimony[ii].

Evidence is composed by the raw data obtained from the source of proof, when it has not yet been evaluated by the judge.

Evidential result is what the Judge obtains as a result of the assessment of the credibility of the source and of the reliability of the element obtained.

The *evidence* tout-court can therefore be defined as the logical process that derives the existence of the fact to be proved from the known fact. The fact is that the clue and the means of proof assume the term "evidence" tout-court only after the outcome of the search, the admission, the assumption and the evaluation of the means of proof and then, after the completion of the decisional syllogism made by the judge, that is the decision transfused in its provision, sentence, order or decree. Secondly, until 2015, the Italian code did not comprehend a (sub) category that could be defined as a means of digital evidence or digital proof or IT documentary proof. With the introduction of art. 234 bis entitled "Documents and computer data"[iii], in the Italian legal system it is possible to speak, even formally, of a means of documental digital evidence. The concept of "digital evidence" derives from the transposition of the concept that can be found in the Common Law system, which, by providing for the legal category of *evidence*, where it relates to a set of digital data, may also include the sub-category of *digital evidence*.

[ii] Art. 194 of Italian c.p.p.

[iii] Art. 2, comma 1 bis del D.L. 18 febbraio 2015 n. 7, converted with L. 17 April 2016, n. 43, “Integrazione delle misure di prevenzione e contrasto delle attività terroristiche”.

Again from a legal point of view, regarding the use that forensic operators intend to make of digital data taken from electronic devices and systems, it is undeniable that they are fully part of the scientific evidence(167). In truth, the scientific content related to digital data pertains to the modalities with which they must be treated in order to preserve the information assets useful for the proceeding. In this regard, it highlights the whole question of the method used in data processing and therefore the best methods, also called best practices, protocols, guidelines, standards that include the best data processing techniques for procedural purposes, which will be addressed in the following.

2.3 Italian Legislation

On the subject there is no homogeneity in the normative production; the discipline can be traced back to at least three areas of interest: crimes committed through the use of new technologies (criminal law of information technology), electronic documents and the security of information systems, the protection of personal data. In all cases it is a recent regulatory production, developed over the last twenty years. IT crimes are regulated in the penal code and in some subsequent special laws.

2.3.1 Law n. 547/93 on computer crimes

Computer crimes were introduced into Italian regulatory system for the first time by Law n. 547 of 1993[iv]. The first novel in the matter of

[iv] Law 23 December 1993 n. 547, "Modificazioni ed integrazioni alle norme del codice penale e del codice di procedura penale in tema di criminalità informatica"

computer crimes inserted a series of new cases to "update" the traditional areas of protection from attacks on IT assets or through IT tools. The following forecasts were of particular importance:

- the arbitrary exercise of his own reasons with violence on things was supplemented by the prediction of violence on the computer program and on the computer or telecommunications system (Article 392, paragraph 2, paragraph 1);
- on the subject of damage, the rules on the damaging of computer or electronic systems (Article 635 bis of the Criminal Code) and on the dissemination of programs aimed at damaging or interrupting an IT or telematic system (Article 615 quinquies of the Criminal Code) were introduced;
- the law that already punished the attack on public utility facilities (Article 420 of the Italian Civil Code) was supplemented by the provision relating to IT and telematic systems;
- the protection of the domicile was extended to the digital domicile, providing for crimes such as the unauthorized access to an IT or telematic system (Article 615 ter of the Criminal Code), the improper possession and dissemination of access codes to IT systems (Article 615 quater cp);
- computer fraud (article 640 ter c.p.), was outlined on the model of traditional fraud;
- the falsification of the IT document (Article 491 bis of the Civil Code) extended the cases of false documents;
- the crimes of interception, impediment or illegal interruption of digital or electronic communications (article 617 quater of the Italian Criminal Code), of the installation of equipment designed to intercept, prevent or interrupt computer or electronic communications (Article 617 quinquies of the Criminal Code), and falsification, alteration or suppression of the content of computer or electronic communications (article 617 sexies cp), in order to protect correspondence and electronic communications;

- the extension of the applicability of the rules set out in articles from 617 to 617 sexies c.p. remote transmissions of sounds, images or other data (623 bis 6 c.p.).

The only rule introduced in the code of criminal procedure was that concerning the interception of electronic flows. Therefore, Law 547/93 did not affect the evidence's regime, nor did it provide anything in terms of digital evidence. However, Law 547/93 on the one hand represented the first attempt to adapt the system to the new demands imposed by the digital revolution, on the other it is the first sign of the poor understanding of new digital phenomena from part of legislator, proven by the inclination to adapt the pre-existing regulatory frameworks to completely new and only apparently similar realities. This inclination was then confirmed in the subsequent law which modified the procedural rules.

2.3.2 Law n. 48/08 - Convention on Cybercrime ratification

The study of the issues related to the investigation of cybercrimes can not today leave aside the most important novelty in the field of forensic science such as the Ratification Law of the Budapest Convention on Cybercrime[v], the normative text of 2001 which represents the first international agreement on the subject crimes committed through IT and telematic systems, Internet and other computer networks.

The Convention, adopted by the Council of Europe, was prepared by a Committee composed of 22 experts coming from some of the 41 Council countries, of observers coming from important countries outside Europe (USA, Canada, Japan, South Africa) and of delegates of some of the major international organizations (Interpol, European Union, Unesco).

[v] Law 18 March 2008, n. 48. "Ratifica ed esecuzione della Convenzione del Consiglio d'Europa sulla criminalità informatica, fatta a Budapest il 23 novembre 2001, e norme di adeguamento dell'ordinamento interno".

It was opened for signature in Budapest on November 23, 2001 and immediately signed by 30 countries, including Italy.

The Cybercrime Convention aims to harmonize cyber-crimes, to establish a rapid and effective international cooperation regime, to create a supranational common legislation, to provide for more streamlined procedures in the context of seizures, searches and interceptions, to assign powers necessary for the prosecution of this type of crimes in national procedural law. Moreover it has the task of outlining common definitions of crime between countries, defining common powers of investigation and preparing means of international cooperation. In the first chapter, consisting of a single article, the following four definitions are preliminarily provided:

"computer system" means any equipment or group of interconnected or connected equipment, one or more of which, on the basis of a program, perform automatic data processing;

"computer data" means any presentation of facts, information or concepts in a form likely to be used in a computerized system, including a program capable of enabling a computerized system to perform a function;

"service provider" means any public or private entity that provides users of its services the ability to communicate through an IT system, as well as any other entity that processes or stores IT data on behalf of such communication service or for users of this service;

"data transmission" means any computerized information related to a communication through an information system that constitutes a part of the communication chain, indicating the origin of the communication, the destination, the route, the time, the date, the size , duration or type of service.

In the second chapter of the Convention the measures to be taken at national level on substantive criminal law are enucleated. In particular, Articles 2 to 10 regulate a series of criminal cases that must necessarily be present in all signatory States in order to guarantee homogeneity of the indictments. The other articles are aimed at regulating the liability of legal persons and applicable sanctions (Articles 11 to 13) and the definition of the procedural law (Articles 14 to 22), within which the rules concerning the modalities for timely ensure electronic data that can be altered or modified are particularly relevant. The third chapter lays down the provisions on international cooperation regarding mutual assistance with provisional measures - rapid storage and dissemination of computer data - as well as investigative powers.

On July 1, 2004 the condition was reached for the entry into force of the Convention, which was established when five ratifications were reached, of which at least three of the Council of Europe states.

Although in the name explicitly refers to cybercrime, the scope of the Budapest Cybercrime Convention is of utmost importance for all legal practitioners as it includes not only pure IT crimes, but also all "common" crimes committed through an IT system or which evidences are in digital format.

The Italian Law of 18 March 2008, n. 48, cd. "Ratification and execution of the Council of Europe Convention on Cybercrime, made in Budapest on November 23, 2001, and rules for the adaptation of the internal legal system", makes limited modifications to the penal code, since Italy has already provided for the introduction of IT crimes in national law in previous years. Some important rules regarding the acquisition and recovery of data on which to undertake a forensic investigation are introduced in the code of criminal procedure: in relation to IT or telematic systems, the new formulation of art. 244 comma 2 c.p.p. provides for the need to adopt technical measures aimed at ensuring the preservation of the original data and preventing alteration; in all cases where there is a well-founded reason to believe that data, information, computer programs or traces in any case relevant to the offence are in an IT or telematic system, the new formulation of art. 247, paragraph 1-bis of the Italian Civil Code, provides for the search to be carried out by adopting technical measures aimed at ensuring the preservation of the original data and preventing it from being altered.

With this law the legislator confirmed the propensity, already manifested with Law 547/93, to set up new standards in existing structures that should have responded to the needs of regulatory modernization. This time, the regulatory intervention concerned certain provisions of the criminal procedure code, with the aim of fight cybercrime and electronic crime. Beyond the intentions, the law 48/08 should be recognized the merit of having turned a light on the problem of processing digital data for procedural purposes and to have sensitized the forensic operators on the need to adapt the remaining rules to the demands imposed by technological progress in the IT field.

On the other hand, in addition to the adaptation of some provisions of substantive criminal law already introduced by Law 547/93, the Italian legislator has limited itself to modifying the procedural rules concerning jurisdiction, acts at the initiative of the judicial police and the means of searching the test, editing the code and the pre-existing rules with additions and incisions, but without any intervention on the means of proof with digital content in the direction of the implementation of the *digital evidence*.[28–32] So, one of the most serious shortcomings, the Italian legislator has failed to bring the definition of data or IT data into our system, denying them the rank of an independent legal asset worthy of protection.

Therefore, the legislation to be implemented in the national systems was limited to providing for the procedures to which the processing of digital data subject to investigation must be subjected, as well as the technical procedures for the preservation of intact and reliable data, to protect the representative assets and therefore information that may derive from it.

From these considerations it follows that the data of which part of the proceeding, judicial police, public prosecutor, or other party, intends to use, have probative value only if they constitute a faithful "*representation of facts, information or concepts in a form suitable for digital processing*", and not if they result in a partial, erroneous, misleading, or worse, misrepresentative or even altered reality.

The achievement of this objective is delimited by two jambs: the first is constituted by the procedural rules governing the investigation activity to be applied in light of the rights recognized by the international conventions and by the Italian Constitution [vi];

the other jamb is instead constituted by the technical standards based on scientific principles proper of the Informatics and organized to guarantee to the computer data the integrity of the representative capacity of the facts, information and concepts for probative purposes. The architrave is made up of the Computer Forensics methods which connect and stabilize the two areas.

The procedural and collaboration rules provided for by the Convention, while not imposing or indicating specific technical and scientific instruments that implement the precautions imposed in the processing of data for procedural use, imply their adoption. See, for example, the art. 19 of the Convention concerning

"(...) Search and seizure of stored computer data":

[vi] Key point of Budapest Convention preamble: "*(...) Mindful of the need to ensure a proper balance between the interests of law enforcement and respect for fundamental human rights as enshrined in the 1950 Council of Europe Convention for the Protection of Human Rights and Fundamental Freedoms, the 1966 United Nations International Covenant on Civil and Political Rights and other applicable international human rights treaties, which reaffirm the right of everyone to hold opinions without interference, as well as the right to freedom of expression, including the freedom to seek, receive, and impart information and ideas of all kinds, regardless of frontiers, and the rights concerning the respect for privacy; Mindful also of the right to the protection of personal data, as conferred, for example, by the 1981 Council of Europe Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data; (...)*" transfused in art. 15 of Convention: "*(...) Article 15 - Conditions and safeguards 1 Each Party shall ensure that the establishment, implementation and application of the powers and procedures provided for in this Section are subject to conditions and safeguards provided for under its domestic law, which shall provide for the adequate protection of human rights and liberties, including rights arising pursuant to obligations it has undertaken under the 1950 Council of Europe Convention for the Protection of Human Rights and Fundamental Freedoms, the 1966 United Nations International Covenant on Civil and Political Rights, and other applicable international human rights instruments, and which shall incorporate the principle of proportionality. 2 Such conditions and safeguards shall, as appropriate in view of the nature of the procedure or power concerned, inter alia, include judicial or other independent supervision, grounds justifying application, and limitation of the scope and the duration of such power or procedure. 3 To the extent that it is consistent with the public interest, in particular the sound administration of justice, each Party shall consider the impact of the powers and procedures in this section upon the rights, responsibilities and legitimate interests of third parties".*

1. *Each Party shall adopt such legislative and other measures as may be required to empower its competent authorities to search or similarly: a computer system or part of it and computer data stored therein; and b a computer-data storage medium in which computer data may be stored in its territory.*
2. *each Party shall adopt that legislation and other measures as a requirement for the use of the search system or similarly access to a specific computer system or part of it, pursuant to paragraph 1.a, is stored in another computer system or part of it in its territory, and it is possible to extend the search or similar access to the other system.*
3. *Each Party shall adopt the following: legislative and other measures as may be required to empower the competent authorities to seize or similarly*
    a) *seize or similarly secure a computer system or part of it or a computer-data storage medium;*
    b) *make and retain a copy of those computer data;*
    c) *maintain the integrity of the relevant stored computer data;*
    d) *render inaccessible or remove those computer data in the accessed computer system.*
4. *Each Party shall adopt such legislation and other measures as may be necessary to empower its competent authorities to order any person who has the knowledge of the computer system or measures to protect the computer data to be provided, as is reasonable, the necessary information, to enable the undertaking of the measures referred to in paragraphs 1 and 2.*
5. *The powers and procedures referred to in this article shall be subject to Articles 14 and 15. (...)*

Well, the purposes set by the 3rd paragraph, lett. a), b), c), d), in addition to being subject to regulatory provisions can be implemented with "other measures" which can not be other than technical measures.

Such measures, being necessarily informatics, will have to guarantee the aims set by the legislative norms, with the proper techniques of the computer science, and therefore with measures that base their technical value on the scientific principles of the information technology. On the contrary, if those purposes set by legislative measures were pursued with non-scientific instruments, the same objectives could not be achieved.

In this regard, the Italian legislator has not typified specific data processing techniques for procedural purposes, a choice that, in light of the continuous technological progress, appears to be acceptable. Therefore, Digital Forensics, taking into account the aims pursued by law, turns its attention to computer science to acquire the methods that make them on a scientific basis so as not to depress representative loyalty and therefore the evidential effectiveness of computer data.

3

State of the art of Computer Forensics practices

Contents

3.1 Best practices in Digital Forensics

In regulating the modus operandi of the operations, it can be observed how the attention of the legislator has focused, rightly, more on the result that must be obtained rather than on the method to be followed: the canonization within legal norms of technical procedures more that to represent a guarantee, would have led, in the long run, to contrary and distorting effects represented by the constant evolution of the discipline and by the peculiarities peculiar to each case (see Section 3.2.1).

The Law 48/08 has introduced some recurrent principles in all the rules of the criminal procedure code in order to guarantee the reliability of the data object of the processing for procedural purposes, but does not fix the technical specifications. On this point, a large field of experts has simplistically invoked the need for investigative best practices, believing that a uniform protocol can be the solution to problems. In reality, this approach neglects that forensic IT techniques can not be proceduralized because they are only subjected to scientific verification and validation. Secondly, authorities lack the power to set procedures in best practice; moreover, the application of the law can not be subordinated to the existence or application of extra-legal rules; finally, technological progress is so rapid that every best practice would undergo a process of rapid obsolescence, because it is difficult to hypothesize such proactive best practices as to anticipate the placing on the market of new devices or software (think, for example, of the need to acquire data relating to a crime committed with a mobile phone just placed on the market).

3.1.1 Best practices definition

Especially in the period prior to Law 48/08, operators engaged in investigations concerning IT data, have felt the need for shared practical methodologies, indicated as best practices, guidelines, criteria, etc., for the implementation of data processing techniques for forensic purposes.

In a broader sense, for best practice[i] we mean the "*set of activities (procedures, behaviours, habits, etc.) that, organized in a systematic way, can be taken as a reference and reproduced to favor the achievement of the best results in the business, engineering, health, education, government environment and so on*". The expression also refers to the process of development and application of operating standards used by complex organizations, but there are equivalent expressions. In the field of computer forensics, best practice has often referred to practices developed overseas by federal agencies or by associations operating in the sector considered a point of reference for the approach to the technical-IT aspects of Computer Forensic.[33–37]

3.1.2 Inadequacy of the procedural law

The debate that took place in the first half of the 2000s began with real problems such as the inadequacy of the procedural law that did not provide for references or technical indications to be adopted, the inadequacy of the investigative procedures, not sufficiently developed, the inadequate technical training of the employees and the scarcity of the means available to the operators. However, it can be assumed that it was above all the judicial police that needed guidelines to best direct their

[i] from Wikipedia: "A best practice is a method or technique that has been generally accepted as superior to any alternatives because it produces results that are superior to those achieved by other means or because it has become a standard way of doing things, e.g., a standard way of complying with legal or ethical requirements."

activities to allow the results of the investigation to "resist" during the trial. The debate on best practices in Computer forensics has therefore focused on various elements:

- about the *object* to be defined, the need to outline uniform technical procedures for the standard processing of digital data for procedural purposes was highlighted;
- about their *purpose*, was identified in the need to provide theoretical and practical support to the technical investigation and investigation of the Judicial Police, the public prosecutor and the related consultants;
- with regard to the *figures* to be involved, some forensic operators were identified, and therefore essentially judicial police officers, magistrates, lawyers, technicians;
- on the recognition to *best practices* of a privileged procedural effectiveness, and therefore the presumption of the validity of the procedures followed and the binding nature of all the parts of the process.

However, at the best examination, this approach revealed serious limitations, which can be summarized as follows:

- regarding the object of the best practices of Computer Forensics, the lack of legal principles regarding the processing of digital data for the purposes of proof emerged clearly, moreover established by the first jurisprudence attested on positions still behind;
- the objectives pursued, in the absence of principles and reference to the possibility of mutability in the field of defensive investigations, concretized the risk of the instrumentalization of the procedures by the persons involved in the investigation and investigation activities;
- as far as the editors of the best practices were concerned, no valid selection criteria were highlighted in the context of a valid academic and scientific course, thus remaining the expression of a voluntary and extemporary activity; moreover, the omitted indication of the auxiliaries of the proceeding as subjects to be

involved (chancellors, bailiffs, custodians), constituted an index of the fundamental error proving a partial, and therefore unscientific, approach to the question;

- finally, with regard to the procedural effectiveness of best practices, it was unaware that their privileged effectiveness would have conditioned the judge (and judgment) by simultaneously violating the constitutional and procedural principles of subjection of the judge solely to the law (Article 101, c. 2 of the Constitution), of the formation of the judge's free conviction (article 192), of the adversarial and of the formation of the trial during the trial (article 111 of the Constitution).

3.2 Computer Forensics ISO standards

Until October 2012 the methodologies were defined in some best practices of the sector, aimed at outlining the paradigms of technical action in the forensic field, through a basic methodology aimed at: a) acquiring the test without altering or damaging the device original; b) authentication of the find and image (bit stream image) acquired; c) to guarantee the repeatability of the assessment; d) an analysis without modification of the original data; e) maximum impartiality in technical action.

On a practical level, however, the implementation and development of shared procedures clash with two orders of limits, attributable, on the one hand, to the "technological variable" represented by both the characteristics of the media in which the data are contained and the habitat technology in which the device is inserted and operates (consider, for example, the acquisition of data contained within a hard disk in a trust computing environment, that is equipped with mechanisms for encrypting content); on the other hand, he observes the "subjective variable", constituted by the subject operating and by the objectives connected to the action: for the Police Forces the aim will be to acquire useful elements for the investigations while preserving their authenticity, for the Judiciary it will be such findings to facts of a criminal nature, for the Technical Advisor of the defense or the defender in the field of defensive investigations will be to check that the processes followed allow an appropriate exercise of the right of defense.[38]

For example, the rules modified by Law 48/08 provide that the paradigmatic procedure to be carried out when the data are processed for judicial purposes, consists in the production of a copy of the data that must be carried out on adequate supports, through a procedure that ensures conformity of the copy to the original and its immutability, adopting technical measures aimed at ensuring the preservation of the original data and preventing its alteration; furthermore, the procedure may provide for the affixing to the copy of the electronic or computer seal. The tools that implement this obligation are the bit-by-bit image, on durable secondary memories, verified through the hash, digitally signed and marked temporally.

Moreover, some researchers working at the Department of Computer forensics in Bologna have developed the empirical procedures that have implemented the principles of the Convention (use of optical media, digital signature, calculation of hash by appropriate algorithms, etc.) that give technical form the precepts just examined, which were considered scientifically founded, and which recently met the endorsement of the ISO 2012:27037 standard.

Among the various standards promoted by the ISO, documents relating to Computer Forensics have recently been submitted, which are candidates for effective and internationally recognized technical standards of reference:

- ISO/IEC 27037:2012, issued in definitive version on October 15, 2012 with regard to guidelines for the identification, collection, acquisition and storage of digital tests;
- ISO/IEC 27041, with regard to guidelines on the guarantee of suitability and adequacy of the methods of investigation;
- ISO/IEC 27042, related to guidelines for the analysis and interpretation of digital evidence;

- ISO/IEC 27043, concerning principles and processes for the investigation of IT incidents.

3.2.1 ISO/IEC 27037:2012 standard

The ISO/IEC 27037:2012 standard is a document that in its definition refers to other ISO/IEC standards[ii] and contains some guidelines that can certainly be considered as the operational reference in the field of computer forensics for the identification, collection, acquisition and preservation of the digital evidences, necessary in any investigation that needs to maintain the integrity of digital evidence. The standard aims to provide guidance to those responsible for identifying, collecting, capturing and storing potential digital evidence:

- the Digital Evidence First Responders (DEFR), authorized, prepared and qualified to intervene first on the scene of an incident by collecting and acquiring the digital tests with responsibility for their management;

- the Digital Evidence Specialists (DES), a subject that performs the duties of an DEFR and has specialist knowledge, skills and abilities in managing a wide variety of technical issues;

- incident response specialists;

- managers of computer forensics laboratories.

[ii] ISO/TR 15801 - Document management - Information stored electronically - Recommendations for trustworthiness and reliability.
ISOI/IEC 17020 - Conformity assessment - Requirements for the operation of various types of bodies performing inspection.
ISO/IEC 17025:2005 - General requirements for the competence of testing and calibration laboratories.
ISO/IEC 27000 - Information technology - Security techniques - Information security management systems - Overview and vocabulary.

The document provides that the responsible parties manage the potential digital evidences with methodologies that are adequate on a global scale, with the aim of facilitating the investigation of digital devices and tests in a systematic and impartial manner, while preserving their integrity and authenticity. The standard also aims to provide information to decision-makers who need to determine the reliability of digital evidence. It is applicable to organizations that need to protect, analyse and present the potential digital evidences, ie the data that can be obtained from different types of digital devices, network devices, databases and anything else provided already in digital format (the standard doesn't provide for analog/digital conversion).

Because of the fragility of digital evidence, it is necessary to put in place an appropriate methodology to ensure the integrity and authenticity of potential digital evidence: the standard does not address the methodology of legal processes, disciplinary procedures and other actions related to management potential digital evidence that is unrelated to the purpose of identification, collection, acquisition and storage.

Applying the standard requires compliance with national laws, rules and regulations, will not replace the specific legal requirements of a jurisdiction while serving as a practical guideline for any DEFR or DES in investigations involving potential digital evidence. It does not extend to the analysis of digital evidence and does not replace specific jurisdictional requirements that relate to instances such as eligibility, persuasive value, relevance and other limitations subject to judicial control of the use of potential digital evidence in the courtrooms. The standard can help in simplifying the exchange between jurisdictions of potential digital evidence. In order to maintain the integrity of digital evidence, operators are required to adapt and correct the procedures described in compliance with the legal requirements of the evidences required by the specific jurisdiction. The ISO/IEC 27037:2012 standard integrates the ISO/IEC 27001[iii] and ISO/IEC 27002[iv] standards, and in particular, the control requirements regarding the acquisition of potential

[iii] ISO/IEC 27001:2013 - ISO/IEC 27001 - Information security management

[iv] ISO/IEC 27002:2013 - Information technology - Security techniques - Code of practice for information security controls.

digital evidences, offering an additional application address, as well as being applied in independent contexts two standards mentioned.

3.3 Digital Forensics phases

Taking up the definition of forensics, that is "the set of techniques and tools used to identify, acquire, analyse, evaluate and present the evidence found on a computer or other device", we can identify five different phases in it. So, in the literature, forensics has always been defined in five (sometimes four) phases:[39] identification; acquisition (with collection and conservation); analysis; evaluation and presentation (in the four-stage versions, the latter two were grouped). However, taking into consideration the various definitions of computer forensics, the state of technology, various guidelines defined in the best practices and, lastly, how correctly defined and illustrated within the ISO/IEC standards on the topic, the process of computer science forensic would more correctly be divided into seven phases according to the following scheme: identification; collection; acquisition; conservation and transport; analysis; evaluation; presentation.

Identification The identification phase consists in searching evidences useful to the case. The digital evidence has a physical form and a logical form: the physical form is given by the technique of impression of data on the support (for example, magnetization of matter in the case of magnetic supports or the representation in pit and land of optical media); the logical form is the virtual representation of the data in bits that can take the value of 0 or 1. Identification is the process of research, recognition and documentation of potential tests in digital format, ie bit storage devices that may be relevant to the investigation, identifying where possible also data that can be found outside or in virtual spaces such as cloud systems. However, this is not just a simple search for devices, but the right priorities must be defined taking into account the risk of data volatility, in order to minimize the damage to potential digital evidence and obtain the most intact and genuine data.[40]

This is probably one of the most difficult phase of the whole investigation, because today everyone is surrounded by dozens of supports that can store (or conceal) information. The investigator doesn't have to underestimate any element [GF09], but rather he has to consider any secondary functions of some devices, for example:

- multimedia players (mp3/mp4) can be used as generic mass storage devices;
- USB drives (mouse, hub, small desktop gadgets ...) can contain small flash disk memories;
- some printers may have a network interface that can be used as a file repository.

Collection The collection phase consists in obtaining "materially" the data to be analysed. The forms through which it can be implemented are the duplication or the seizure.

The most common findings are computers, which very often are not handled correctly during the seizure phase, not considering the fact that they have easily alterable storage media. If a computer is found turned on, it is advisable to make some considerations before proceeding further. First of all it is essential to understand what data to expect compared to the considered crime. In some cases it may be necessary to make a copy of the RAM memory, which loses its contents at shutdown, to ascertain which programs were running at the precise moment of the seizure. Then it is necessary to evaluate the best shutdown mode, between standard shutdown (shutdown) and the disconnection from the power supply (unplugging). The former is usually inadvisable because it causes numerous modifications to the system files, but the latter could be fatal for old or delicate hardware, with the risk to make the machine unusable.

Once the digital devices have been identified, the computer scientist (or rather the DEFR according to the indications of the ISO/IEC 27037:2012 standard) must decide whether to proceed immediately with the acquisition or whether to proceed with the support collection operations that will be followed by the acquisition operations. Collection is the phase of digital data processing in which devices that may contain potential digital evidence are removed from their original location to be transported to a laboratory or, more generally, to another controlled environment for the acquisition and the subsequent analysis. Each finding must be labelled with the case number, a description, the date and time of collection and the name of the person who detected it. Devices containing potential digital evidences may be in the on state or in the off state. Depending on the state and purpose of the investigation, as well as the legal limits, different methodologies and tools may be required.

Usually the form preferred by law enforcement is the seizure, whose critical point is only the care of the physical support and the correct maintenance of the chain of custody. However, it has some advantages:

- simplicity and tranquillity: the collection of the physical medium does not require the particular technical knowledge necessary for the acquisition, although the removal of digital storage media still requires some skills; moreover, the postponement of the acquisition operation helps to loosen the tension in critical moments of a seizure activity, avoiding errors;

- rapidity: the collection simply requires the annotation of the identification details of the support in a report, in addition to the workload required for transport;

- tangibility: the tangibility of the support conveys greater peace of mind to the operators and to the suspect who may not have adequate technical and legal knowledge to evaluate the current technical activity;

- preservation of further non-digital tests: in addition to the digital data, an IT evidence could be used to detect other types of tests such as fingerprints.

The alternative procedure, is the duplication of the medium through bit-to-bit copies, which can be necessary under different circumstances where physical collection is not possible:

- IT systems that can not be turned off: these are systems that deliver critical services in 24/7 mode; for example, systems for controlling the exchange of railway tracks;
- IT systems that provide services to third parties: these are systems that typically located in datacenters and provide resources, both computational and storage, to various users, allowing them to reduce costs by centralizing the investment of hardware and software, as well as the costs for the system activity; for example, systems of hosting service providers that host websites;
- virtual systems: these are systems that simulate a real machine whose physical consistency is that of the system on which the activity is performed.

In any case, the collection process must not be limited to the device that contains the digital data but must be extended to the material that concerns it, such as, for example, post-it with passwords, notebooks with notes and power supplies. Furthermore, the collection process must be documented in detail to justify the choice of one method instead of another. The IT media must be carefully packaged away from heat sources so as not to risk corrupting the media and causing accidental data loss.

The following scheme in Figure 3.1, taken from the ISO/IEC 27037:2012 standard, highlights the evaluation process by the DEFR regarding the possibility of making a collection or acquisition. Based on the choice made, the same standard provides detailed procedures to follow.

If you were to tend to collect, in the case of devices turned off the scheme in Figure 3.2 is to follow.

Instead, when you come into contact with a turned on device, the scheme, reported in Figure 3.3, is more complex because it requires the evaluation of some elements that would be lost permanently after switching off the system.

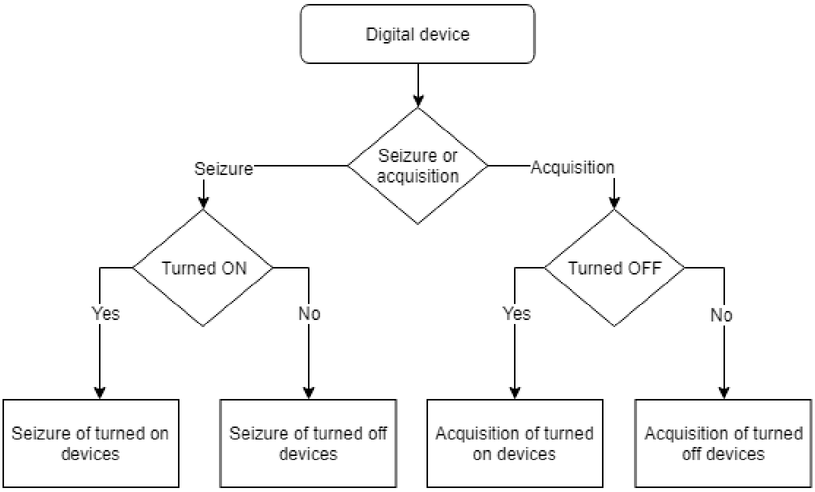

Digital device
Seizure or acquisition
Seizure
Acquisition
Turned ON
Turned OFF
Yes
No
Yes
No
Seizure of turned on devices
Seizure of turned off devices
Acquisition of turned on devices
Acquisition of turned off devices

Figure 3.1: Decision-making policy about the opportunity to collect or acquire potential digital evidence taken from the ISO/IEC 27037:2012 standard

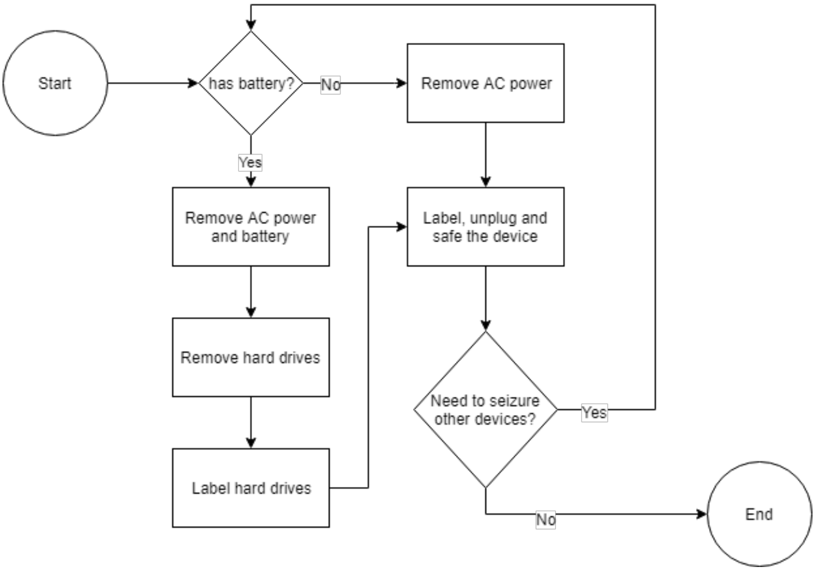

Start
has battery?
No
Remove AC power
Yes
Remove AC power and battery
Label, unplug and safe the device
Remove hard drives
Need to seizure other devices?
Yes
Label hard drives
No
End

Figure 3.2: Guidelines for the collection of turned off digital devices, taken from the ISO/IEC 27037:2012 standard

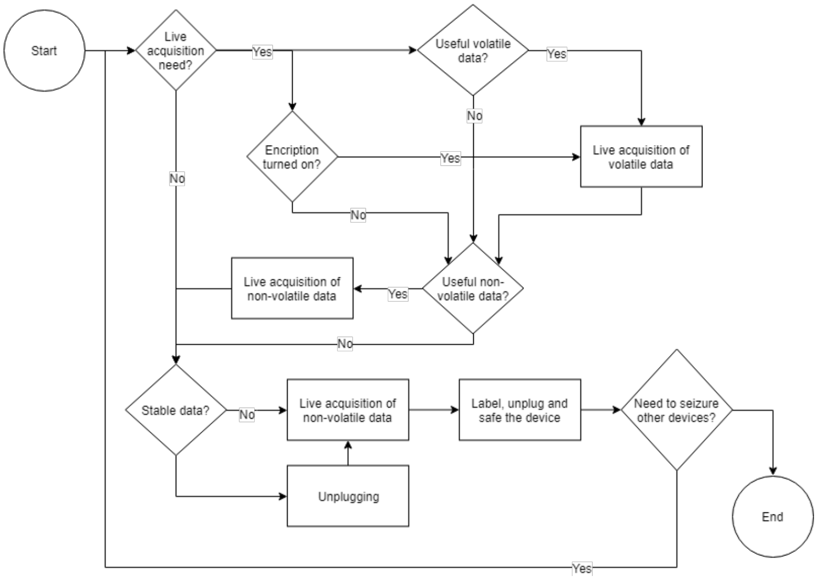

Start
Live acquisition need?
Yes
No
Useful volatile data?
Yes
No
Encription turned on?
Yes
No
Live acquisition of volatile data
Live acquisition of non-volatile data
Yes
Useful non-volatile data?
No
Stable data?
No
Live acquisition of non-volatile data
Label, unplug and safe the device
Need to seizure other devices?
Unplugging
End
Yes

Figure 3.3: Guidelines for the collection of turned on digital devices, taken from the ISO/IEC 27037:2012 standard

Note that after switching off the system (with unplugging or with normal shutdown), the procedure is based on that of the previous diagram for the devices switched off.

Acquisition Acquisition is the process of producing a copy of digital evidence called a forensic copy (or bit-stream image, or a bitmap image), which is a complete copy of the media including unallocated space and slack space[v]. The operating methods and tools to be used depend on the situation, cost and timing, but in any case must be documented in detail so that they are, where possible, reproducible and verifiable by a technical consultant of one of the other parties. The acquisition process must be as least invasive as possible, ie it must involve the alteration of the least possible number of bits, possibly aiming at the inalterability of the source support, in order to produce a sequence of bits representing

[v] Slack space is a set of digital data generated by the way the data is organized into a storage medium. Regardless of the file size, media are structured in fixed-size blocks (sectors). Any unused space within the block will contain data that existed until the time of deletion and will continue to contain it until the space is overwritten as a result of wiping or new file allocation

the original sequence. The final product of an acquisition can therefore be a clone, ie a device that contains the exact and identical sequence of the source device, or an image file (or a series of fragmented files) that represent the exact and identical sequence of the device source; in the second case it is possible to apply compression algorithms as in the case of the Expert Witness format[vi]. The identity between source and destination can be easily proved by hash algorithms, mathematical functions that allow to synthesize the representation of millions of bits in hexadecimal strings of a few tens of bits: in fact, the application of a same hash function to two bit sequences always produces the same result (digest), if and only if the two input sequences are identical. It should be noted that in some circumstances - such as the live acquisition[vii] of a system - this verification is not possible, therefore, it is necessary to resort to an in-depth documentation of the acquisition process, also by means of video and photographic recording.

If the machine is still on at the time of the seizure, it is necessary to carry out a first immediate analysis on site because the shutdown would result in the loss of volatile data in RAM memory or in the swap area and the inaccessibility of data protected by encryption. Live analysis must therefore verify network connections, open ports, running process and their behaviour, current users, files, kernel modules and devices in use.[41]

It can then be switched off and disconnected from the network to conduct *post-mortem* analysis, ie with the machine turned off, which consists of a survey of resident data (data files and programs), system file and logs analysis.

The investigations should not be carried out on the original hard disk but on a copy of the bit stream: the physical copy compared to the logical one allows to preserve data apparently not present on the disk that are instead highlighted with a forensic analysis.

[vi] EWF is a proprietary data format used for forensic copies devised by EnCase but now supported by most forensic software.

[vii] The term *live forensics* indicates a method of acquiring information from an IT system while it is operating, in order to capture those data, transiting or stored in it, which would not be acquired id the device is turned off.

Bit-stream image acquisition procedure The procedure for the acquisition and copying of data constitutes a crucial event during the investigation of IT data. At present, the best method for acquiring data from a memory is the production of a c.d. bit-stream image, or bit-by-bit image, which, due to the procedural value, is also called "forensic copy", ie a complete copy of the memory on which the bits are stored, including unallocated spaces, those apparently unallocated and slack spaces, and in ways that do not involve alteration of data (for example, creation dates, last access, last modification). In fact, cases may occur in which not only the data that "appear" are stored on the memory, ie those that are immediately usable, but also others that are not visible because they are deleted, or stored in particular ways, such as encryption or steganography or both simultaneously. The deletion of data from a memory through the standard procedures provided by the operating systems, unless it is carried out with particular techniques of *wiping*, that go to overwrite the data, it concerns only the index that allows you to rebuild the documents so that the user can no longer access them. However, the data c.d. deleted, as long as they are not overwritten, they are still stored in the memory and therefore potentially recoverable and analysable.

The same goes for the data recorded in the *slack space*. To understand the phenomenon we must start from the organization of the memory which is structured in blocks (sectors) of fixed size. The bits are stored on these blocks, regardless of the file size. When the files are smaller than the blocks, more or less large area of memory will remain unused. But if in that area had already been stored data of files subsequently deleted, in the area inside the blocks remained unused by the archiving of subsequent files will continue to insist data of previously archived files and will continue to be stored there until the space is not overwritten by following allocation of new files or wiping operations.

Therefore, the data acquisition and copying process must be carried out in such a way as not to alter any bit of the original memory and must be integral, that is (tend to) realize a sequence of bits to be stored on a destination memory that perfectly represents the original bit sequence stored on the source memory. To this aim, between the memory on which the original data are stored and the destination memory of the bit-by-bit image, hardware or software devices or combinations of the two are interposed, which prevent write access to the memory, and therefore the modification of the original data.

The final result of such a data acquisition procedure must be a perfect clone of the original memory, and therefore a memory on which the same identical sequence of the source memory is stored, or an image file (or a series of fragmented files) that reproduces the exact same sequence of bits stored on the original memory; in this second case, data compression algorithms can be applied. The verification of conformity between the original data and the obtained bit-by-bit copy is performed using hash functions: if the hash string consisting of bits of the bit-by-bit copy coincides with the hash string consisting of the original bits, then they are compliant.

Conservation After the acquisition operations, it is necessary to guarantee the correct preservation of the finds. With regard to information systems, a double regard must be given: it must be considered the modifiability of the storage media to which the exhibit belongs and subsequently it must be identified the technology in use on it. On the basis of all these factors it could be established the best method of conservation for the considered media.

The ISO/IEC 27037:2012 standard prescribes conservation requirements such as maintenance of the chain of custody, the use of suitable packaging that depends on the characteristics of the item to be treated and the control of the environment in which the find is stored (environmental threats, humidity, temperature); as regards transport, the standard requires that the moving parts be secured and that everything is properly packed.

Chain of custody In the conservation phase the chain of custody is fundamental, that is the detailed list of what has been done with the acquired copies, which must be maintained in order to reconstruct the history of the investigation: who took charge of the supports, where and when, how they were transported and where they are stored, who has had access to them and what they have done. The presence of a weak link in this chain could nullify the whole work. In some legal systems the chain of custody is included in the code of criminal procedure.

Analysis The analysis activity consists in recovering those data that may be useful for the purposes of a forensic investigation, which are therefore likely to be hidden, voluntarily or not, from the view of a common user.

*ISO/IEC 27042:2015 — Information technology — Security techniques — Guidelines for the analysis and interpretation of digital evidenc*e was published in 2015. This standard offers guidance on the process of analysing and interpreting digital evidence, which is of course just a part of the forensics process. It lays out a generic framework encapsulating good practices in this area.

Aside from the standard evidential controls (maintaining the chain of custody, scrupulous documentation etc.), the standard emphasizes the integrity of the analytical and interpretational processes such that different investigators working on the same digital evidence ought to come up with essentially the same results - or at least any differences should be traceable to choices they made along the way. Given the volume, variety and complexity of digital evidence these days, that's quite a challenge, hence the drive for standardization, good practices, common terminology and sound, rational approaches.

The standard touches on issues such as the selection and use of forensic tools, plus proficiency and competency of the investigators.

This phase will be discussed more in detail in the Section 4.1.

Valuation A further significant aspect concerns the determination of the circumstances in which a crime is committed and the modalities of the same; even if the victim can be known, reconstructing details is essential to shed light on what happened.

It should be clarified that the reason for which an evaluation phase of the exhibit is necessary lies in the fact that the bit is 0 or 1, hence absolutely poor information. Since the information can be altered, polluted, counterfeited, it is necessary to ascertain whether these events occurred, if they were potentially verifiable and whoever could have done these actions. Another aspect concerns the acquisitions operations, for which it is necessary to assess whether they have been carried out with rigour and in a correct manner in compliance with the current legislation. Once these requisites have been assessed, judgements must therefore be formulated regarding the reliability of the informatics finding, in the sense of its integrity and verifying possible alterations, and its authenticity, ascertaining the author or the authors. So, during the evaluation phase a meaning is given to the data emerged during the analysis phase and the legitimacy of the operations carried out to acquire them is ascertained.[6] A valid judgement is issued regarding the reliability, integrity and authenticity of the find, carefully evaluating every possible event of alteration (who could, how and when).

Presentation The last phase of the forensic examination consists in the presentation of all the evidence found by the consultant and his conclusions in a detailed report that will be considered during the trial. Within the technical report all the documentation acquired or produced during the analysis must be included. This means is essential to bring in the "knowledge" of the technical activities carried out during the investigation phase. The methodology used to analyse the data, the tools used, the discoveries made must be indicated in an exhaustive way, providing an explanation of what was done, why, by whom and how long each operation was performed.

The purpose of the presentation is to transmit to all the parts of the process the facts ascertained according to scientific techniques and methodologies, which will illustrate the steps taken. The report will be more effectively illustrated where there will also be audiovisual instruments (eg. slides, recordings, etc.), simulations and practical demonstrations to support deposition.

3.4 Hardware and software tools for acquisition

Computer Forensics laboratory is a scientific laboratory that must keep in mind a series of prescriptions defined within the internationally accepted standards, first of all, ISO/IEC 17025 and ISO/IEC 27037. The main objective of the laboratory must be to achieve the highest level of guarantees, legal and technical, on the goodness of the results, recalling the heavy reflection that such activities commonly have on people's lives.

The computer forensics laboratory deals with the identification, acquisition and analysis of digital evidences for legal purposes. Given the rapid evolution of the discipline, the laboratory should be in contact with other laboratories in order to set up an investigative network, also in order to compare the results of the technical investigations on the basis of the various software and hardware tools adopted. All of this, in order to determine with ever greater precision the investigative protocols applicable in most cases, avoiding to leave the imagination, the improvisation, the inexperience and the personal initiative of the technicians, which last one, however, is strongly used in presence of new situations to be investigated.[42]

As said, forensic activities require a method aimed at ensuring that the data is treated appropriately and that preserves its characteristics and genuineness throughout the duration of the investigative activity. For the first technical activities there are open source and free software tools that allow to approach the digital forensics without particular investments;[43] however, often open-source tools are not optimized for specific tasks, nor are they constantly updated, so it is certainly important - if not necessary - to use more professional equipment to perform operations more quickly and effectively.

3.4.1 Digital storage devices

The tools and software to be used for the acquisition of digital media are naturally variable depending on the type of support: just think that only for the acquisition of the two most classic and widespread media (hard disk and smartphone) both the hardware equipment and the software are completely different. In general, the essential requirement in each acquisition station is the write blocker, a solution that can be implemented in two ways:

1. the tool prevents writing attempts by informing the operating system of the impossibility to complete the operation;

2. the tool memorizes writing attempts throughout the session, making the operating system believe that the writing operations have actually been completed correctly.

As for hard disk media, solid state disks, optical media, floppy (in various sizes), USB sticks and tapes, apart from the appropriate cables and suitable readers, an instrument certainly reliable and free is the *dd* command of Linux operating systems; in fact, Linux allows to access devices in read-only mode, thus not requiring any additional hardware to prevent write accesses. There are numerous Linux distributions that collect a set of software specific to the computer forensics (not just for acquisition, but also for analysis) that are called forensic distributions[viii]. Although not recommended, because it is not verifiable, Microsoft Windows operating system offers the possibility of acquiring IT supports without the need to use as intermediary a hardware device that blocks the writing by introducing an appropriate registry entry. In general, to

[viii] Most famous distributions are:
DEFT (`http://www.deftlinux.net`),
CAINE (`http://www.caine-live.net`),
Kali Linux (`http://www.kali.org`),
Swift Linux (`http://www.swiftlinux.org`),
SIFT (`http://digital-forensics.sans.org/community/downloads`),
ForLEx (`http://www.forlex.it`),
NetSecL (`http://netsecl.com`),
Matriux (`http://www.matriux.com`),
BackBox Linux (`http://www.backbox.org`).

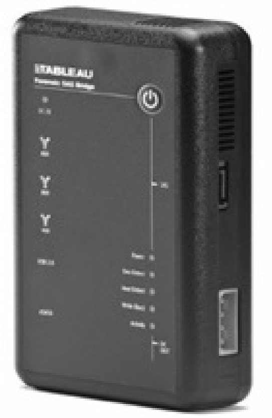


Figure 3.4: Example of write blocker made by Tableau

avoid altering - even accidentally - an IT find, it is good practice to use a write blocker.[43] For the actual acquisition are then available several software that also provide compression (one of the best known is EnCase but there are others as AccessData FTK).

The write blocker (see Figure 3.4 and 3.5) is a hardware tool that allows you to read data from the media preventing it from being modified. The use of this tool is certainly recommended when operating acquisitions with Windows operating systems that are extremely invasive and can lead to alteration of data with just the connection of the digital data storage device.[44] As an alternative to the use of personal computers (and possibly writeblockers) it is possible to use hardware devices built specifically for the acquisition of computer media. These devices, known as hardware cloners, allow to acquire in clone or image mode - sometimes even compressed - and to calculate the hash string.

In generating a forensic copy, it is essential to calculate an imprint which uniquely distinguishes the digital trace of the forensic analysis in order to comply with the aforementioned data integrity requirements. This guarantee seal is created with the hash computing and is a sure reference to the original trace. An additional technique to be applied to secure acquisitions over time is the adoption of time-stamp that allows the operations carried out to be placed over time.

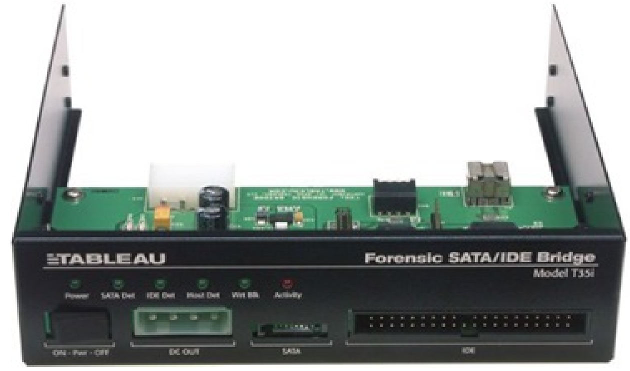


Figure 3.5: Example of write internal blocker made by Tableau

3.4.2 Mobile devices

As regards mobile devices, ie all devices that can be easily moved and contained in a pocket such as portable media players, satellite navigators, mobile phones, smartphones and tablets, the evolution of technology has led to the creation of tools that have computational capacity comparable to that of computers: a smartphone today offers better performance than most of the computer marketed only a decade ago.

Above all, smart-phones and tablets are extremely common devices, as they have become an integral part of everyday life and contain information that are often extremely sensitive. From a technical point of view, these devices are similar to small computers that run their own operating system, more or less specialized and feature-rich.

Modern devices use complex data structures to store data, such as heterogeneous media files, documents, GPS positions and SQLite databases that contain most of the information related to the applications installed on devices. This has led digital forensics investigators to focus more and more their activities on modern smart devices, like smart-phones, and to Cloud services connected to them. Therefore, it has been necessary to translate the investigative activities on these new devices, thus making necessary the adoption of new acquisition and analysis tools. However, mobile devices constantly improve the security and protection of the data they contain, which makes the work of the Mobile Forensics Expert increasingly difficult and complex.

In general mobile devices keep data:

- in the internal memory;
- in the removable external memory;
- on the SIM card.

For the acquisition and analysis of removable external memory it can be used the common computer forensics techniques related to acquisition and analysis of digital data storage devices, while for the SIM card there are now consolidated methods of acquisition and analysis with tools such as SIMcon[ix] or Paraben SIM Card Seizure[x]. The activity is more complex when it is necessary to acquire and analyse the hardware of the mobile device (without SIM and external memory), which however provides the greatest amount of results useful for the investigation[xi].

The analysis can be conducted in two ways:

1. invasive: physical extraction of the memory chip[45] or access as "root" to the system, exploiting hacking techniques that allow complete control of the device;

2. non-invasive: data acquisition and reconstruction by connecting the mobile device to an IT system (see Figure 3.6) which, through a specialized software, provides for the recovery of relevant data (possibly deleted) such as SMS, call list, address book, end so on; in reality, not knowing exactly how these software interfaces with the device, it would be appropriate to consider also this technique invasive, with the consequence that in general the acquisition of mobile devices should be carried out with the guarantees provided

[ix] `http://www.simcon.no/`

[x] `https://www.paraben.com/sim-card-seizure.html`

[xi] Most famous products are:
Cellebrite UFED (`http://www.cellebrite.com`),
Oxygen Forensics Suite (`http://www.oxygen-forensic.com`),
Paraben Device Seizure (`http://www.paraben.com/device-seizure.html`),
MobilEdit (`http://www.mobiledit.com`).

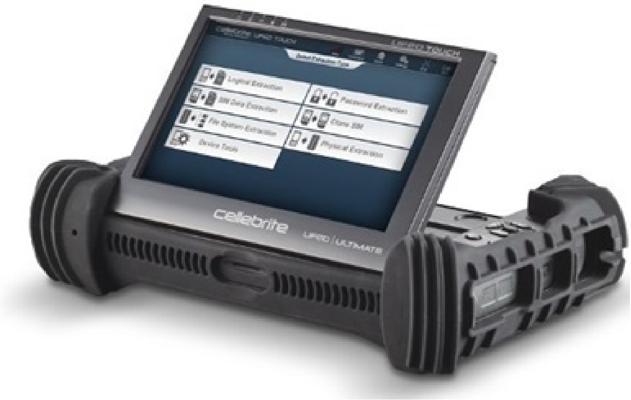


Figure 3.6: Example of a hardware device for the acquisition of media such as mobile phones, smartphones, tablets and SIM cards (Cellebrite UFED Touch), see Section 3.4.2

.

for by art. 360 Italian c.p.p. on the subject of non-repeatable technical assessments.

Depending to the actual considered device and its operating system version, there are different techniques that can be adopted in order to acquire a data stored in it, which vary in complexity, effectiveness and supported devices.

Physical Acquisition mode The physical acquisition mode allows bit-to-bit copy of all partitions of the entire memory, including unallocated space. This is certainly the most advanced, effective and valid technique to adopt in judicial domain. Usually it takes place with the device in "Download" or "Fastboot" mode for Android systems and in DFU mode for iOS systems (prior to the iPhone 4S model), for this reason it is also possible to bypass the device unlock code, if it should be present. The produced output is a memory dump that can be analysed with classic Mobile Forensics analysis tools.

File System Acquisition Mode The file system acquisition mode extracts all the files stored in the device memory, including system data, integrated application data and unallocated space within files. However, it allows the recovery of most of the deleted files and often, from the point of view of extracted data, the acquisition is equivalent to an acquisition performed in physical mode.

IOS devices can be acquired in file system mode through the use of the iTunes backup function, which is able to extract all the contents of the device (both present and deleted) as files .ipa (iPhone Application), multimedia content, chat, e-mail and all other information from applications.

On Android devices this mode of acquisition is carried out through the operation of Android backup, which backup all installed applications including installation files apk and shared memory, whether internal or external micro SD card. Starting from Android 5.0 Lollipop to adopt this mode of acquisition root privileges are required, which, of course, is a limitation, as not always it is possible to grant them. As an alternative, to circumvent authorization problems and any password protection or unlock code, this activity could be performed through a custom recovery, a system boot mode that replaces the default recovery mode and which contains many more features, including a full backup of all memory partitions (called NANDroid Backup). The NANDroid backup creates a copy of the individual system partitions and saves them in a shared memory or SD card folder. It will then be possible to analyse this data with one of the programs that will be described in Section 3.4.2.1. The use of a custom recovery alters the device and it is not always possible to perform this activity in the judicial field, however during the backup a log file and all the MD5 hashes are generated to guarantee data integrity.

Logical Acquisition Mode The logical acquisition mode extracts only the files present and shared by the operating system from the device. It does not perform any carving operation on the memory, but can be useful to demonstrate and certify the presence of some contents received, sent or stored by the analyzed device, during a legal proceeding. For example, this mode can be useful (in the event that the other, more complete, were not supported) for crimes like stalking or child pornography. For this mode the same considerations of the file system mode are valid: it is necessary to have access to the device or to know the unlock code, in order to authorize the PC to perform the data reading.

3.4.2.1 Acquisition and Analysis tools

Forensic acquisition and analysis techniques are constantly evolving, but the tools adopted for the extraction of data from mobile devices are always a step behind the security protection adopted by their operating systems. In fact it has became increasingly complex for engineers who attempt to access and correctly interpret the data stored on memory devices through reverse engineering process. The main difficulty in the creation and development of efficient and reliable forensic analysis tools lies in industrial secrets of manufacturers, the presence of proprietary file systems, new/proprietary operating systems, the lack of knowledge of the investigators and so on. There are several tools, both hardware and software, that allow forensic acquisition of mobile devices, those shown below are the most known and used.

UFED Cellebrite Cellebrite is the leading company in the Mobile Forensics sector. Its most used tools in the forensic field are UFED Touch, UFED Touch2 and UFED For PC for the acquisition phase and UFED Physical Analyzer and UFED Reader for the analysis phase. These tools support over 20,000 devices and therefore they offer the most complete solution currently on the market. UFED Touch is a small computer, running Windows Compact Edition, on which the Cellebrite acquisition software is executed. On the left side, it hosts USB and Ethernet ports, to which you can connect the device to be acquired, be it a smartphone or tablet, but also an SD card reader. On the right side, instead, there is the destination USB port, to which a mass memory can be connected. The UFED Touch also has a SIM card reader port on the front of the device, a wireless card for software updates and a Bluetooth antenna for acquisitions that require it. There are extensions to the kit, such as UFED Camera used to make photographic acquisitions and brute force unlock of the device, UFED Chinex used to acquire Chinese clone phones and UFED Memory Card reader, for the acquisition of SD cards and memory cards, also equipped with the functionality write block (Write-Block).

Oxygen Forensics Oxygen Forensics was initially designed for communication between mobile device and PC, then it has been specialized in reading and acquisition of smartphones and tablets. The suite provides a complete kit for acquisition and analysis similar to that proposed by Cellebrite. It contains the Getac F110 tablet PC, the Oxygen Forensics Extractor software, the USB license dongle, the software to be installed on a second PC for the analysis and the cable set for the connection of the devices. As the Cellebrite suite, it allows to acquire a great variety of models, even if it is limited in some aspects, as well as not allowing the acquisition of satellite navigators, drones and other more specific devices.

Magnet Axiom Thanks to its module Magnet Acquire is able to perform the acquisition of mobile devices running iOS and Android operating system. Magnet does not provide any hardware tools for the acquisition, unlike its rivals, so it supports a limited number of devices. As a suite, it focuses on Computer Forensics, then on the acquisition and analysis of traditional memory devices such as hard disks, pendrive, memory cards, and so on.

Open Source Alternatives Using open source alternatives is a double-edged sword, because, on the one hand you can save thousands of euros in commercial licenses of software or hardware devices, but on the other hand both the acquisition and the analysis phases become more complicated and cumbersome. Some open source techniques take advantage of features and functionality provided directly by the manufacturer or integrated into the Operating System, such as Android Backup mode, iTunes backup, and also DD copy on external SD card and NANDroid Backup. For Android devices Linux Santoku distribution and the Autopsy software are the most valuable tools, while for iOS it is possible to use the Zdziarski techniques and the iPhone Backup Analyzer software.

3.4.3 Networks

IT data can reside within IT systems and can be transferred across the network from one system to another: in this case, the data acquisition mode is the interception of the electronic traffic generated and received by the IT systems. The telematic interception can be divided into two categories:

1. single: monitoring all the traffic of a specific user regardless of the protocol used and the content of the communications;

2. parametric: traffic monitoring that meets certain requirements, such as the presence of a certain sequence of bits in the flow; usually parametric intercepts take place on very large geographical areas.

From an architectural point of view, an interception system consists of three parts:[46]

- capture tool: hardware and software (probe) capable of interfacing with the network to be monitored and producing a copy of the selected traffic with predefined filtering rules;

- computer: apparatus (core) responsible for the reconstruction and processing of communication sessions from which all the information requested is extracted;

- visualization and analysis tool: apparatus (viewer) that allows the use of the information collected and the reproduction of the contents at the application level of the protocol stack.

The interception is therefore effectively carried out by the probe which sets itself the goal of remaining transparent to the intercepted system, not altering the data and listening in order to passively catch (in technical jargon "sniffing") everything that is in transit.

The probe memorizes all the communications concerning the computer system to be monitored, placing itself in a topological area closer to the system in order to minimize the amount of non-relevant data and interfacing with the various communication channels used (local LAN or wireless network, twisted pair telephone, radio links ...): for example, if the computer system being analysed is connected to a LAN network, the probe should be connected to the network switch to which the system is connected, possibly using the *span port*[xii], or positioning the probe between the system and the switch.

WireShark[xiii] is one of the main software (moreover open-source) for monitoring and analysing network traffic;[47] in the field of computer forensics it is also useful in forensic acquisitions of content available on the network for which you also need to freeze the entire network traffic generated in the session of downloading data from the network.

[xii] The span port is a switch port that replicates all traffic exchanged with a prefixed system

[xiii] `http://www.wireshark.org`

4

A semantic-based methodology for forensic analysis

Contents

4.1 The Analysis of digital evidences

The phases described in Section 3.3 are quite mechanical and repetitive, requiring only a working methodology. The analysis phase, on the other hand, requires in-depth knowledge of computer architecture and operating systems, but also networks, communication protocols, systems administration and also a certain talent on the part of the examiner who must find relevant material for investigation purposes. The analysis phase must be performed on a copy of the find; it must be reproducible and every single operation performed must always produce the same result.

The analysis activity consists in recovering those data that may be useful for the purposes of a forensic investigation, which are therefore likely to be hidden, voluntarily or not, from the view of a common user: when a file is deleted it is only hidden to the user as it continues to reside on the disk. The delete operation does not destroy the entire file but modifies a bit that discriminates if the file has to be displayed or not by the user: with this behaviour it is possible to have the operation speed and life time of the storage medium much higher. The traces of the deleted file are lost only when that sector is rewritten to store another file and for this reason when it is necessary to analyse a disk it is necessary to make a bit-stream copy of the media, which preserves that data would appear to be non-existent.

Other techniques commonly used to hide data consist in modifying the file extension to "trick" the operating system and prevent it from opening the file with the default application or hide data written by encrypting them in other documents with the technique of *steganography*[i].

[i] The term steganography is composed of the Greek words steganòs (hidden) and gràfein (writing) and indicates a technique dating back to ancient Greece which aims to hide the communication between two interlocutors

The investigator must pay close attention also to the space not visible to the common user as it is in those areas that often reside the most useful data for forensic purposes. Some of them are:

- the content or update date of the files;
- data relating to the events of switching on and off the system;
- e-mail: e-mail is one of the most important sources because it keeps a very high number of information (not only the text, but also the date, the sender or the recipient ...);
- peer-to-peer files: they are the files shared by file-sharing applications (fundamental to go back to downloading pirated copies of software or music or sharing pedo-pornographic material);
- temporary internet files: browsers used for browsing save the files downloaded from the various sites to a temporary folder and then actually display them on video; it is also possible to find traces of the chronology of the last visited sites;
- conversations on Instant Messaging systems;
- temporary application files: some applications during execution make use of support files to keep track of any backups (for example a word processor periodically saves the changes that the user makes on the document) that will then be deleted at the exit of the application;
- installation files: during the installation process, several temporary files are copied or generated to determine which software was installed on the machine and on which date;
- print file: the print jobs are queued and the information saved by the operating system in a file which will then be deleted when the process is completed;
- partial files: copying files from one mass storage device to another may sometimes fail due to user interruption or insufficient space on the destination drive during a file generation operation; in this case, the copied data will still be present on the destination device

up to the point where space was available, but will be treated as a partially overwritten deleted file;

- elements that demonstrate a willingness to delete or hide documents.

During the investigation phase it may be necessary to analyse possible damages (or attempts) against a computer connected to the network: in this case the investigation aims to trace the intrusions in the computer system (this operation consists in checking the presence of any backdoor or by analysing log files) to find and analyse any malicious code present in the machine.

Several events, not only those mentioned above but also others deducible from physical or forensic elements, can be correlated to reconstruct the timeline of an individual's actions. This important activity of organization and correlation of the retrieved facts takes the name of "information management".

4.2 Techniques for semantic analysis

This section discusses research related to the realization of solutions for structuring textual information. In detail, first, a description of most relevant existing techniques for structuring narrative text is given. Finally, the motivations for the proposed work are clearly stated and its research contribution is diffusely discussed.

4.2.1 Information structuring

Generally speaking, a textual document is the product of a communicative act resulting from a process of collaboration between an author and a reader; the former uses language signs to codify meanings, the latter decodes these signs and interprets their meaning by exploiting the knowledge of:

- infra-textual context, consisting in relationships at a morphological, syntactic and semantic level;
- extra-textual context and, more in general, the encyclopedic knowledge involving the domain of interest.

This implies that the comprehension of a particular concept within a text requires information about the properties characterizing it as well as the ability to identify the set of entities the concept refers to. As a result, both the subjectivity of domain knowledge and the interpretation given by the author with respect to final readers make the structuring of text a thorny task to be performed.

Up to now, there has been extensive research on structuring narrative text,[48] encompassing a wide range of interdisciplinary methodologies which combine computer science, logic, mathematics, linguistics and others. In detail, existing solutions rely on different techniques to analyse texts and automatically extract relevant information.[49,50]

Many of these techniques utilize linguistic-based approaches, embracing Natural Language Processing and Computational Linguistics,[51] to gain an understanding of the text, others employ statistical or pattern-matching based methods[52] for analysing specialist or sectorial narrative texts, leading to the development of specific disciplines, like Corpora Linguistics, Textual, and Lexical Statistics,[53] or for mining information from texts and supporting document categorization.[54]

Linguistic-based techniques involve low-level activities, such as tokenization, which segments sentences and identifies minimal units of text, named tokens (e.g. words, word particles, abbreviations, acronyms, alphanumeric expressions punctuations, etc.) and normalization, consisting in handling variations of the same lexical expression in order to obtain a unique representation, by harmonizing spelling and capitalization. Moreover, Part-of-Speech (hereafter, POS) tagging identifies the part of speech of each word within a narrative text and categorized accordingly as a content word (e.g. nouns, verbs, adjectives and adverbs) or functional word (e.g. articles, prepositions and conjunctions).[51] Finally, lemmatization identifies the lemma (i.e. a dictionary form) of all the inflected forms of individual text tokens (i.e. diagnose, diagnosing, diagnoses, and diagnosed are all forms of diagnose). This leads to the identification of proper nouns as well as noun and verb phrases representing entities, concepts, events, and their relationships contained within a narrative text.

Standard statistical techniques use mathematical models to determine whether a word or phrase is a term that characterizes the target domain. To achieve this goal, they measure unit-hood and term-hood as the "degree of strength or stability of syntagmatic combinations and collocations" and "degree that a linguistic unit is related to domain-specific concepts", respectively.[55] Unit-hood is only relevant to complex terms (i.e. multi-word terms), while term-hood deals with both simple terms (i.e. single-word terms) and complex terms. On the one hand, most of the existing techniques for measuring unit-hood employs conventional measures such as mutual information[56] and log-likelihood,[57] and simply relies on the occurrence and co-occurrence frequencies from local corpora as source of evidence. Mutual information measures the co-occurrence frequencies of the constituents of complex terms to assess their dependency, whereas, log-likelihood attempts to quantify how much more likely the occurrence of one pair of words is than the other.[58] On the other hand, most of the existing approaches for evaluating term-hood makes use of distributional behaviour of terms in documents and domains, and some heuristics related to the dependencies between term candidates or constituents of complex term candidates.[58] Common measures for weighting terms employ frequencies of occurrences of terms in the corpus. They identify sets of terms or keywords that are collectively used to represent the content of documents, by assigning a weight for each term, which measures the importance of a term in a document. There are various implementations, but the most common one is the classical Term Frequency–Inverse Document Frequency (hereafter, TF–IDF) and its variants.[59] Statistical and pattern-matching techniques have been also used to group documents into clusters or to map individual documents or parts of them to pre-defined topic categories.[60] Algorithm types used in these methodologies include Bayesian Probability, Neural Networks, Support Vector Machines, and K-Nearest Neighbors.

Moreover, ontology learning methods, based on natural language processing, formal concept analysis and clustering, have been also developed to address the problem of automatically building conceptual structures out of large text corpora in an unsupervised process.[61] In addition, a plethora of ontology learning frameworks has been developed in the last decade, such as OntoLearn,[62] OntoLT,[63] Terminae[64] as well as Text-ToOnto[65] and its successor Text2Onto,[66] and integrated with standard ontology engineering tools. All these frameworks implement various and different ontology learning methods.

However, the integration of some of these different techniques sufficiently general to be used in many domains as well as their customization and instantiation to face the specific application requirements pertaining the forensic domain still constitute open issues. In particular, the main necessity is to design a reconfigurable solution able to handle the heterogeneity of existing forensic images and provide semi-automatic procedures for defining the peculiar lexicon that better represents the specific domain of interest.

4.2.2 Semantic approach advantages

Acquisition and analysis are critical phases in digital investigations: the heterogeneity of support that can store digital evidences, in addition to the technological evolution and the wide range of investigation scenarios, doesn't allow the identification of a unique procedure for digital evidences acquisition. Forensic tools play a major role in this phase, since they

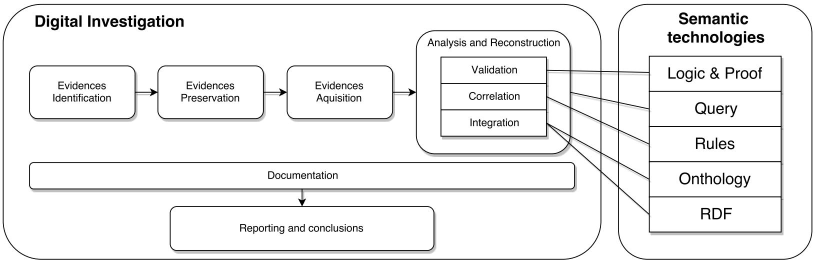


Figure 4.1: Digital Investigation phases and related semantic technologies.

adapt acquisition processes to different digital devices (e.g. hard disks, USB flash drives, mobile devices, IDS Firewall logs, memory, etc.)[67–69] The specialized and evidence-oriented design of forensic tools produces acquired data in different format and representations, that get more difficult the analysis process and so calls for advanced interoperability techniques for evidences correlation.

The development of automatic processes that assist detectives in data acquisition and analysis phases simplifies their work, especially when dealing with large amount of data. Currently, the most of forensic tools work on plain text data which does not allow advanced analysis processes. A semantic approach, eventually based on ontologies, use a unique and reliable representation of domain concepts, enabling data structuring on one hand, and standardised representation of data on the other hand.

Unlike structured data formats (i.e. relational databases), ontologies analysis tools easily make inference of new information, checking knowledge consistency, etc. This because an ontology explicitly represents relationships between entities.

The methodology proposed in this work improves, through semantic technologies, all the main phases of a digital investigation, with respect of evidence discovery, integrity and correlation. It can provide a framework able to describe, in a more expressive and formalized way, the representation of a given scenario. Figure 4.1 shows relationships between adopted semantic technologies and phases of digital investigation.

Main potential advantages of such an approach regard:

- Information Integration: the RDF data model simplify integration of data coming from multiple sources, due to its schema independence and standardized representation of knowledge in the subject-predicate-object form.

- Inference: the RDF/OWL combination can infer class membership and typing information from ontological definitions; a reasoner can them in order to infer dynamically class membership of the instances.

- Extensibility and Flexibility: RDF/OWL provide compatibility with data model of forensic tools input and output. OWL provides flexibility by defining custom ontologies according to the scope and integrating multiple existing ontologies through ontology mapping processes.

- Search: queries can be enhanced taking advantage of the reasoning engine and the semantic mark-up used along traditional keywords during document indexing.

4.3 Methodology for forensics evidences correlation

This section describes the proposed methodology for digital evidence integration and correlation. The methodology workflow is presented in Figure 4.2.

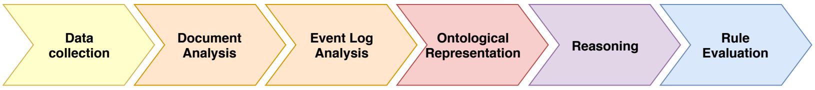


Figure 4.2: Phases of the proposed methodology.

4.3.1 Data Collection

The first "Data Collection" phase involves all the acquisition operations and aims at generating inputs for next phases. During this phase the examiner must respect the chain of custody principles and has to be compliant with acquisition best practices, depending on analysed media (hard disks, mobile devices, etc.). During this phase preprocessing and data reduction may be applied too, using techniques such as KFF (Known File Filter), in order to reduce the amount of data managed by next phases.

The complexity (in terms of time and space-memory) of this phase strictly depends on the kind of device to be examined, on its specification and on the amount of data stored by it.

4.3.2 Document analysis

Then the data are processed by a *Document Analysis* phase that extracts the relevant concepts and the relationships among them. The approach adopted for structuring documents is essentially aimed at properly locating and characterizing resources in a text by recreating the domain model to which that text pertains. In detail, terms convey the fundamental concepts of a specific knowledge domain: they have their realization within texts and their relationships constitute the semantic frame of both the documents and the domain itself. For this reason, the detection of a series of relevant and peculiar terms in a text allows determining the set of concepts that can be used to define features characterizing a resource.

In order to extract relevant terms from a text, the proposed approach hybridizes linguistic and statistical techniques. In particular, linguistic

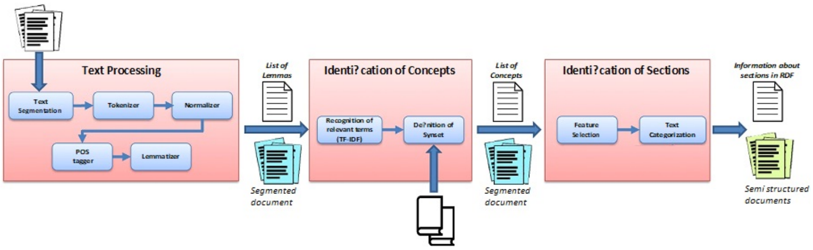


Figure 4.3: The approach proposed for structuring evidences from unstructured documents.

filters are applied to words in order to extract a set of candidate terms, whereas a statistical method is used to calculate word occurrences within a text and, consequently, assign a value measuring the “strength” or “weight” of a candidate term. Indeed, not all words are equally useful to describe documents: some words are semantically more relevant than others, and among these words, there are lexical items weighting more than others do. The whole approach is outlined in Figure 4.3 and diffusely described in the next subsections.

Text processing The first stage of the proposed approach aims at segmenting, extracting and filtering text from unstructured documents in order to make it partitioned into coherent blocks and enriched with metadata specifying morphosyntactic information and citation form for each text element and, thus, suitable for automatic processing. It has been arranged in the form of a sequence of five basic techniques, namely Text Segmentation, Tokenization, Normalization, POS Tagging, Lemmatization, opportunely adapted to this specialist textual universe.

In detail, Text Segmentation consists in performing a complete global segmentation of an unstructured document into distinct homogeneous regions by using features like punctuation marks or white-spaces. Specifically, documents here considered are single columned documents not including graphics and photographs, and are composed of one or more blocks, each of which belongs to a coherent section of the document. One block corresponds to a set of text lines with the same typeface, a consistent line spacing and ending with known punctuations such as “.”, “.. .”, “!”, “?”. The algorithm here adopted iteratively examines each line of the document in turn, from top to bottom. Lines are merged into complete blocks according to heuristic rules using their contents, typography information, or both. Details of the heuristic rules are not discussed here for the sake of brevity. As an example, in the following, a heuristic rule defined for segmenting a block is described in terms of its conditions which have to be simultaneously verified:

Condition 1: The first text line in the block corresponds to either a new line or a normal text line.

Condition 2: The last text line in the block is neither a new line nor a centered line.

Condition 3: The last text line in the block is ended by a known punctuation.

Condition 4: All text lines except for the first and last in the block are normal text lines.

Successively, Tokenization has been applied to each block, once it is segmented, and it has been realized by means of special tools, defined tokenizers, including glossaries with well-known expressions to be regarded as forensic domain tokens, and mini-grammars containing heuristic rules regulating token combinations. The synergic combination of glossaries and mini-grammars has been motivated by the need of a high level of accuracy, even in presence of texts with acronyms or abbreviations that can increase the error rate.[70] Tokenization has been further partitioned into the following sequence of phases: (i) grapheme analysis, to define the set of alphabetical signs used within a block of a segmented document; (ii) disambiguation of punctuation marks, aimed at realizing the token separation; (iii) separation of continuous strings, to recognize strings not separated by blank spaces; and (iv) identification of separated strings, to be considered as complex tokens and, therefore, single units of analysis. Normalization has been automatically performed by first comparing a block of a segmented document to external lexical lists in order to recognize and standardize particular expressions (like well-known abbreviations and acronyms, toponyms, as well as grammatical phrases and specific noun phrases) and, successively, setting proper parameters in order to uniform the different forms (e.g. reduction of capital letters into small letters according to some pre-arranged conditions, such as a capital letter used after a punctuation mark identifies the beginning of a sentence).

POS Tagging has been realized by using Key-Word In Context (hereafter, KWIC) Analysis, i.e. a systematic study of the local context where the various occurrences of a lexical item appear,[60] in order to provide a procedure for word-category disambiguation. In detail, occurrences of each concept are computed in the text and co-text (i.e. the textual parts before and after it). The analysis of the co-text allows detecting the role of the words in the phrase in order to disambiguate their grammar category. As a result, ambiguous forms are first associated to the set of possible POS tags, and then disambiguated by adopting the KWIC analysis: the set of rules defining the possible combinations of sequences of tags, proper of the language, enables the detection of the correct word category. After being categorized, words are enriched with additional morphological specifications, such as inflectional information.

Finally, Lemmatization has been implemented by introducing a partitioning scheme establishing an equivalence class on the list of tagged terms in order to reduce all the inflected forms to the respective lemma coinciding with the singular male/female form for nouns, the singular male form for adjectives and the infinitive form for verbs.

Identification of concepts The second stage is aimed at identifying relevant concepts for each block of a segmented document and organizing them in synsets, i.e. lists of terms that are considered semantically equivalent for the purposes of information retrieval.[71] In more detail, preliminarily, the vocabulary of relevant terms from a block is extracted. It is worth remembering that some words are semantically more significant to describe resources, and among these words, some lexical items weight more than others do. In the proposed approach, the semantic relevance is assessed by *TF–IDF index*,[53] computed over the corpus vocabulary based on term frequency and term distribution within the corpus. This information enables the selection of relevant terms, by filtering all terms having an associated index value under an empirically established threshold. The set of selected terms constitutes the peculiar lexicon used to define features in classification tasks.

On the extracted peculiar lexicon, lexicometric analyses are then applied in order to evaluate the rate of coverage of the extracted terms over the vocabulary of the input block.[70] In the case that the coverage rate results inadequate, the whole process is reiterated by enlarging the empirically established thresholds in order to extract terms with lower associated indices. Once relevant terms are detected, this stage proceeds to clusterize them in synsets, in order to associate the proper concept to the list of terms denoting it. In this way, it is possible to refer a concept independently of the particular term used to denote it. Each concept, then, is referred by a list of terms representing it, codified in a synset. The clustering has been performed by integrating two external resources: the forensic ontology given by "*FORE*" and a thesaurus of forensics terms. The adoption of specialized external resources has a double purpose, i.e. endogenous, since inside the documental base, the same concepts can be referred by different terms, and exogenous, since in a natural language query, a given concept can be denoted by using terms that are different from those occurring in the documental base.

Identification of sections The last stage is in charge of performing a text categorization aimed at assigning labels to all the blocks of the input document depending on the presence/absence of concepts in them.[72] This categorization has been performed by using features extracted from each block as inputs to a combination of supervised linear classifiers, namely Naïve Bayes, Decision Tree and K-Nearest Neighbor.[73] For each block of the document, the feature space is represented by the set of all the concepts included in the synset and appearing in the block itself, with the number of their occurrences in it as values. In other words, the bag-of-words representation is used, since each block is represented with a vector of the concept counts that appear in it. In order to make the use of aforementioned classifiers possible, and, contextually, improve generalization accuracy and avoid "overfitting", a feature selection method based on term frequency has been preventively applied to reduce the high dimensionality of the feature spaces and select the most representative features. In particular, this method makes use of the TF–IDF index, calculated for each concept included in the synset, as evaluation metric to measure the ability of each concept to differentiate each section.

Each single classifier has been trained in order to learn the most predictive values for the features belonging to the set of the ones preliminarily selected and that can characterize a specific section. All the considered classifiers require only a small amount of labelled trained data as input and can be successively evaluated, with respect to their effectiveness, in a testing phase when previously unseen instances of data are considered. They are easy to construct and update since they require only subject knowledge and not programming or rule-writing skills.

In detail, the Naïve Bayes Classifier is constructed by using the training data to estimate the probability of each section among the set of possible ones given some feature values calculated for a new block. This probability is calculated by using the Bayes theorem, with the simplifying assumption of conditional independence, since the feature space contains more elements. In other words, the conditional probability of a concept given a section is assumed to be independent of the conditional probabilities of other concepts given that section. This classifier finally labels a block as the section with the highest probability.

The Decision Tree is constructed by using the C.4.5 algorithm among the possible learning ones. In particular, this algorithm computes a tree, where each internal node denotes a test on a feature, each branch represents an outcome of the test, and leaf nodes represent final categories, i.e. the sections. Gain ratio is used as splitting criteria, i.e. to select the set of features which best partition the different blocks into distinct sections, and, in addition, pruning is enabled to identify and remove branches reflecting noise or outliers in the training data.

Finally, K-Nearest Neighbors is used to classify a section based on the majority category amongst its K-Nearest Neighbors. It is based on learning by analogy, i.e. by comparing a given block with training data that are similar to it. In particular, closeness is defined in terms of a distance metric, i.e. the Euclidean distance. This algorithm computes the distances between a new block and all the previous ones already classified, sorts these distances in increasing order and selects the k blocks with the smallest Euclidean distance values. Finally, it assigns the new block to the largest section out of k selected ones. The results produced by these different algorithms are then combined by means of a voting strategy. It is a very efficient strategy, since previous knowledge about the results that are being decided as well as a large set of data to be analysed are not required. In more detail, every vote has a fixed weight and a fixed probability of occurrence, being independent of the other types of voting. At the end of this process, the assigned output category is the one that gets the majority of votes. Its advantage is that it is almost impossible that more classifiers can produce the same text categorization result, so almost all errors are reduced. This text categorization can be successively refined by domain experts, who can move or reallocate one or more blocks classified as belonging to a section.

The final output of this stage is a semi-structured document subdivided into one or more sections. The information about the different sections produced is coded in RDF for being processed by the framework as described in the next chapter.

4.3.3 Event log Analysis

Event is a change in the system state, independently from its source or nature. Some events reflect normal system activity (e.g., a successful user login), other can occur upon the occurrence of an abnormal behaviour or system faults (e.g., a disk failure). When a system component encounters an event, it could emit an event message that describes the event. So, the system component that emits event messages for event logging, is called Log client. Event logging is the procedure of storing event messages in a form that can be analysed. This data can be structured in many ways for analysis, but the plain text format (the event log, a text file that is modified by appending event messages) is the most common implementation, thanks to its minimal dependencies on other system processes, and its capability to log all phases of computer operation, including start-up and shut-down, where system processes might be unavailable.

To efficiently manage logging information generated by multiple clients, many organizations adopt Security Information and Event management (SIEM) software products and services to combine security information management (SIM) and security event management (SEM). They provide real-time analysis of security alerts generated by applications and network hardware. Main features offered by a SIEM system are the follows:

- Data collection: Logs are the main source of data analysed by an SIEM. Each security system, software, database, present in the system sends the data contained in the log files to the main server on which the SIEM resides. Sending data can be managed by software agent or by allowing SIEM to directly access the device. The choice on which method to use is related to the devices we use.

- Parsing and normalization: Each device manages and stores data in its own way, SIEM provides to standardize the collected data, cataloguing them by type of device and data type, facilitating their interpretation.

- Correlation: The correlation between different events is one of the main functions, allowing to integrate and analyse events from different sources. Although SIEM has a set of already predefined correlation rules, it provides the possibility to create customized rules in order to meet the needs of administrators.

- Reporting: Long-term data archiving combined with the ability to take advantage of customized queries for data extraction enable reporting. Reports can be used for audits, compliance or forensic analysis.

- Dashboard: Dashboards provide an overview of the environment in real time. Using these tools it is possible to provide a representation of the data in the form of diagrams or other models, allowing analysts to quickly identify abnormal activities.

- Notifications: Notifications and warnings are generated when certain events occur, informing users of a possible threat. Reports can be made via dashboard or using third-party services such as e-mail or text messages.

The methodology make use of Information Extraction and semantic techniques and tools by combining the relevant information extracted from textual logs into a semantic representation that enrich the correlation capabilities.

Considering the main features offered by a SIEM system, the methodology introduces additional functionalities that can be embedded in a such system. In particular, the methodology foresee an additional log text analysis phase, to be execute after the Collection or the Normalization, in order to annotate event messages with additional metadata. This processing task aims to automatically extracting structured information from unstructured and/or semi-structured messages and concerns processing human language texts by means of Natural Language Processing (NLP). In fact, a typical Information Extraction (IE) pipeline can recognize names of people, organizations, locations, dates, references, etc. or identify domain-specific terms, automatically extending text metadata to improve search quality.

However, a traditional IE process produces results based on a flat structure, but according to the Semantic Web principles, information have to be represented in a hierarchical structure. So, the idea is, from one hand, to attach the results of NLP process (semantic metadata) to the log file chunks, and, on the other hand, to point them to concepts in an ontology. Information can be exported as a text file annotated with links to the ontology, useful for further ontology population or for indexing to provide semantic search.

4.3.4 Ontological Representation

The goal of "Ontological Representation" phase is to transform the acquired data, by software or even hardware parsing tools, into a set of triplet constituting the RDF data model. Ontologies used in this step can be created ad-hoc or even fetched from shared repositories. Outputs can be stored using different formats, such as RDF/XML or RDF/OWL.

4.3.5 Reasoning

Ontological representation of data is processed during "Reasoning" phase, through an OWL-based reasoner that infers additional axioms on the basis of instances' relations. The reasoner can infer different types of new axioms, enriching the asserted instances with information regarding the definition of their class, properties or subclasses. Moreover, thanks to subclass hierarchy, property relations or property restrictions, reasoner can dynamically classify and correlate instances with higher precision compared to asserted data.

4.3.6 Rule Evaluation

SWRL Rules can assert additional axioms that cannot be inferred through OWL. SWRL Rules are evaluated, during "Rule Evaluation" phase, by a rule engine in order to insert newly inferred axioms into the ontology. This operation is realized with the support of external Rule Engines that translate inferred axioms into RDF data model to permit ontology integration.

4.3.7 Query

By using SPARQL language it is possible to query endpoint hosting for the sets of triplets constituting asserted and inferred axioms. This implements the "Query" phase.

A complete discussion about time and space complexities of the whole methodology is complex since they depend, during each step, on many, different, elements. During the first stage (collection), as previously described, we collect all digital evidences from the devices we are analysing. Hence, the elements that mainly influence time and space complexities are the number of elements to collect, i.e. the dimensions of file systems and of logs to analyse. In addition, the whole time for collection depends on access time too.

The complexity of all the other phases of the methodology depends on the ontology definition (in terms of how many classes, properties and relationship are defined in it): the more relationships are in the ontology, the more is the complexity of each step, because relationships need cross analysis. However, the number of generated triples depends also on the actual values of collected evidences, because each evidence can (or not) contain properties or relationship expected by the ontology definition.

5

A system architecture for digital evidences analysis

Contents

5.1 The proposed framework

5.1.1 Document model

In order to manage the different kinds of multimedia data, their relations and the particular structure imposed by e-Gov applications, the adopted document model uses three different representation layers, as described in the following.

Data management layer describes the semantic content of each single multimedia objects composing the document (such as a text fragment or an image), providing functionalities for managing each single media; as an example, information extraction and indexing over text and images are performed in this layer.

Integration layer describes the relations among the heterogeneous multimedia components of the same document or belonging to different ones, providing functionality for their integration and composition. For example, the property of a text fragment of referring to a given image belongs to this layer.

Presentation layer regulates the way by which the information has to be shown to final users. It provides different representations of the same informative content, according to the formats, the final user's access rights, user preferences and needs and the available technology and user devices.

This approach allows the management of heterogeneous contents, by separating the presentation logic from the content management one. In order to give a concrete example, it permits to give an immutable legal validity to the content of a document even if the format of representation

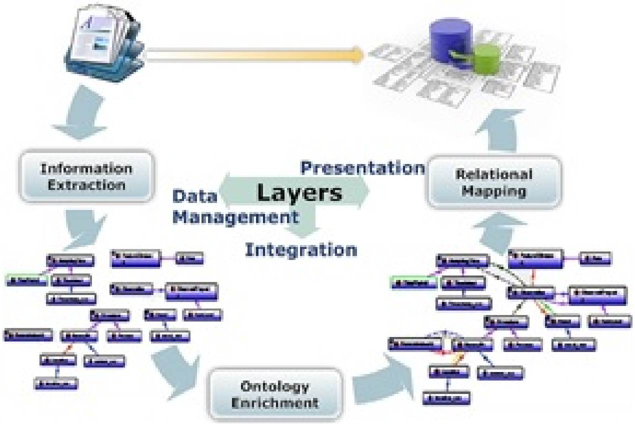


Figure 5.1: General schema of documents processing.

changes, evolving with technology. According to the different description layers of a document, information is semi-automatically extracted and tagged with respect to the concepts contained in the available domain ontologies: associations among concepts and their instances are picked out. A general schema of documents processing is depicted in Figure 5.1.

More in detail, the tagging process leverages different types of ontologies. A Domain Ontology is exploited to formalize the concepts of interest in the reference domain and relationships among them. Some domain ontologies[74] can be further divided into a Structural Ontology that describes how information is organized within the document and models the associations between the internal sections of the document and the set of concepts that can be found in it, and a Lexical Ontology that contains the terms of the general language and can be used to refer wide-ranging concepts presented in the documents, not enclosed in the domain of reference.

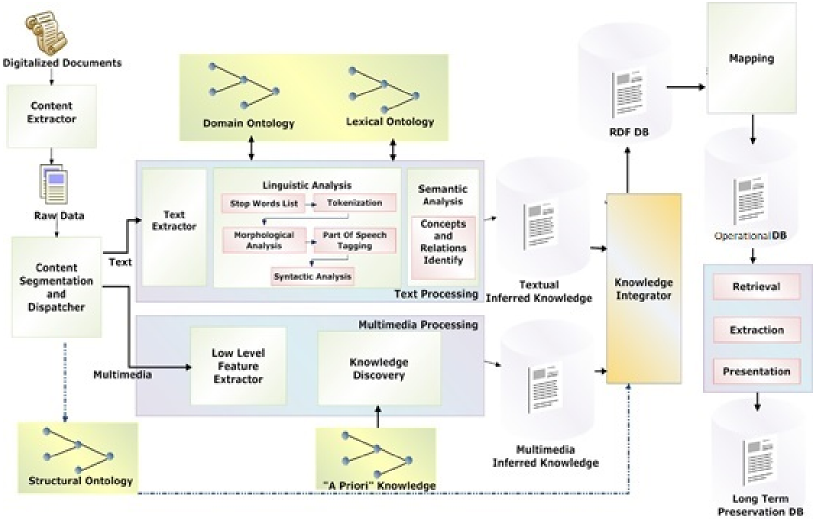


Figure 5.2: Document processing in details.

5.1.2 Processes overview

A general framework is proposed and the related instance for the management of the forensic investigation life cycle. As already stated, forensic images contain text that can be supplied with multimedia information as pictures, video streaming and audio information. The framework is composed of several processes: text processing, multimedia data processing, and the integration, retrieval, preservation and presentation tasks (see Figure 5.2), as described in the following.

The Text Processing process aims at extracting relevant information from text, associating specific concepts to the related key terms and defining relationships among them. The text is processed in according to the following pipeline:[74]

1. *Structural analysis*: performs the text segmentation and the related classification in order to identify the different sections constituting the structure of the document.
2. *Linguistic analysis*: performs a morpho-syntactic analysis of the text (i.e., text tokenization and normalization, Part-of-Speech

Tagging, lemmatization and complex terms analysis) combined with statistic analysis, thus enabling the extraction of relevant terms. These terms and the information about them, refined with the help of domain experts, will constitute a lexicon that is exploited for the building of the set of concepts used for the domain formalization via ontologies.

3. *Semantic analysis*: by using the information of the early analysis, it detects properties and associations among terms, defining the concepts and relationships, allowing ontology building and final documents annotation.

The Multimedia Data Processing process has the aim of classifying the other kinds of multimedia objects, associating concepts from the domain ontology. It is composed of two components implementing innovative methods that have been presented in:[75, 76]

1. *Analyzer*: identifies relevant media parts and produces a low-level description that permits to create some indices to help the tagging and retrieval tasks.
2. *Classifier*: uses the indexing information to automatically deduce which concepts, from the domain ontology, are being associated to media elements.

Final information are stored as RDF assertions into a Knowledge Base and all the knowledge associated to a documents is in turn managed by proper ontology repositories. Different processes (i.e., Knowledge Integrator, Retrieval, Extractor, Presentation) are finally devoted to realize the other discussed tasks. It is to note that the multimedia knowledge is then managed by a Multimedia DataBase Management System (MMDBMS). It supports different multimedia data types (e.g. images, text, graphic objects, audio, video, composite multimedia, etc.) and, in analogy with a traditional DBMS, facilities for the indexing, storage, retrieval, and control of the multimedia data, providing a suitable environment for using and managing multimedia database information[i].

[i] A MMDBMS meets certain requirements that are usually divided into the following broad categories: multimedia data modeling, huge capacity storage management,

5.1.3 Semantic processing

The document semantic processing supported by the designed system needs a preliminary domain formalization stage that has the aim of codifying, with proper data structures (ontologies), the information of interest pertaining to the domain which the documents belong to. On the top of such structures, the different described tasks for semantic processing of documents can be activates.

Information extraction and ontology population Once associations between document segments and ontology fragments have been resolved, the methodology foresees the population of concepts and relationships in the ontology by adding the detected instances. Relevant information are then extracted, document segments are annotated and results are presented in RDF triples containing the properties identified in the segments. Concepts and relations are extracted by exploiting an inference mechanism performed by a Rule-Based System. A generic rule is formed by a combination of token and syntactical patterns, which codifies the expert domain knowledge. In order to derive instances of relevant concepts or relationships, rules exploit Named Entity Recognition (NER), eventually using subsumption on a TBox-Module for deriving more specific concepts. The detected instances can be shown by using tools like KIM[77] that highlights the associations among detected instances and the concept defined in the domain ontology.

Information retrieval Once relevant information related to the domain of interest has been codified for document corpus, it is possible to execute a semantic-based search which is able to retrieve information by content and not only by keywords. The system combines ORDBMS technologies, NLP techniques, proper domain and structural ontologies and inference

information retrieval capabilities, media integration, composition and presentation, multimedia query support, multimedia interface and interactivity, multimedia indexing, high performances and distributed management rules in order to retrieve significant concepts related to each document and to provide semantic querying facilities for users. When a user submits a query, the system identifies the concepts associated to the terms used in the query. These concepts are represented by means of ontologies as synsets, which are the set of linguistic elements linked by a synonymy relationship, i.e. terms that can be used in the same statement without modifying its whole meaning. Furthermore, same terms can be used with different acceptations (the meaning in which a word or expression is understood). In this case, different synsets are related to different meanings. If these ambiguities are present, the system will provide features to discriminate the synset of interest in the search. Once users have selected the desired synset (all synsets are chosen if no selection is specified) a query expansion[78] mechanism is used in order to perform queries on corpus where all lemmas in the selected synsets become lemmatized keywords for a text-based search. The collection of all the documents retrieved from these searches constitutes the results of the semantic-based query. A ranking algorithm is used to score results depending on a similarity measure, based on Tf-Idf index evaluation.

5.2 System Architecture

In order to validate the proposed methodology, a system prototype was implemented. In this section is presented an implementation of a system architecture of the proposed methodology, with a brief discussion of the tools and techniques used. Thanks to its flexibility, methodology can be implemented in different ways, so system architecture can be updated among the evolution of semantic technologies.

The overall system consists of an ontology and five modules: Evidences Manager, Semantic Parser, Inference Engine, SWRL Rule Engine and a Query and Visualization module. An overview of the architecture is presented in Figure 5.3. Its main components are further discussed below.

Evidence Manager The evidence manager loads binary content of digital evidences, identifying the type of given source and verifying its integrity through hash values. This module provides tools to extract knowledge from a forensic image, like user files, browser history, Windows registry, etc. The extraction process uses forensics tools like Hachoir (for binary file manipulation) and Plaso (for timeline creation). The knowledge extracted consists of a set of attributes, including temporal information (date, time and timezone), a description of the information source (source and source type), a description of the event. Many of these fields are structured and ready to be used as instances attributes; instead, some of them, like description of the event, are non-structured fields and they require further processing based on regular expression or NLP techniques. The output of this module is a file containing all the footprints retrieved from the disk image.

Semantic Parser Semantic parser module generates an OWL representation of knowledge extracted from digital evidence in the previous step. This module instantiate the ontology, combined from public domain ones, if available, or custom ones. Ad-hoc ontologies can be created to integrate all the referenced domain ontologies or to define new classes, additional restrictions, new object properties, etc. Ontology integration also promotes reuse of entities defined in other ontologies and increases system flexibility, since domain ontologies may not be easily modifiable.

Ontology population is made by creating an instance for each footprint item of acquired data, and by linking them each other according to formal object properties defined in the ontology schema.

Inference Engine Inference engine performs automated reasoning, according to the OWL specifications, coming from domain ontologies or investigators knowledge. Reasoners can specify the granularity of inferences to be made, such as hierarchy relationship (an individual that is a member of a subclass is also a member of the parent class), or the generation of inverse object properties. Such kind of inferences can improve the performance of query execution. The goal of this step is to enrich the knowledge base with new inferred facts, increasing the examiner's knowledge on evidence. For example it is possible to infer a file type from its extension or to deduce the author of an action from user information in its active account.

SWRL Rule Engine SWRL rules play a major role in the automated integration and correlation parts of the methodology. SWRL Rule Engine adopts SWRL rules in order to correlate different individuals or to establish relationships among individuals belonging to different ontologies but representing similar concepts. Saving SWRL rules in a separate text file promotes the decoupling and enables rules reuse. This module can be called before the reasoning engine make it able to process the axioms that can only be inferred by the SWRL rules, or you may execute the reasoning step twice, one before the rule engine and the other after it.

SPARQL Queries The final module is responsible for accepting SPARQL queries from the user and evaluating them against a SPARQL query engine. The set of RDF triples that have been either asserted during the semantic parsing of the source data or inferred by the reasoning or the rule engine are loaded in-memory and SPARQL queries can be evaluated against it. Once more, the queries can be saved in separate files promoting reuse and decoupling.

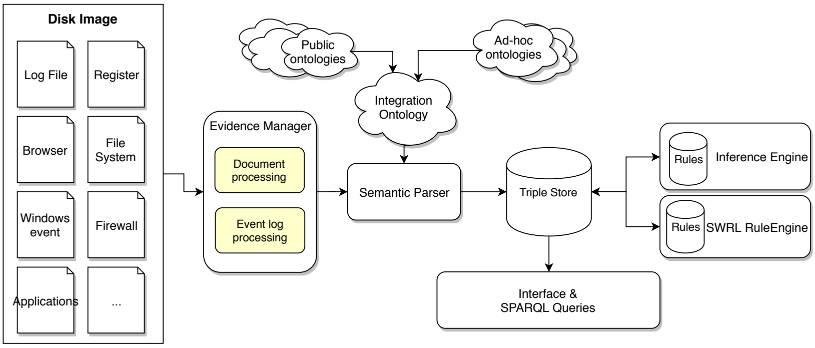


Figure 5.3: System architecture for proposed methodology.

5.3 System Implementation

A prototype version of a Forensic Evidences Management System has been implemented, according to the following features:

1. it exploits a unified data model that takes into account content-based and document-based characteristics;
2. it uses ontological support for managing the semantics of data;
3. it has a multi-layer architecture with different kinds or user interfaces;
4. it provides advanced functionalities for document indexing and semantic retrieval.

Figure 5.4 shows at glance the component architecture of the system. The Digital Documents (DD) are managed by a dedicated component, named Digital Document Repository (DDR).

Its objectives are, from one hand, to allow for interoperability among the different data formats by providing import/export procedures and, from the other one, to manage security in the data access. Moreover, documents can be organized in specific folders to easy management and

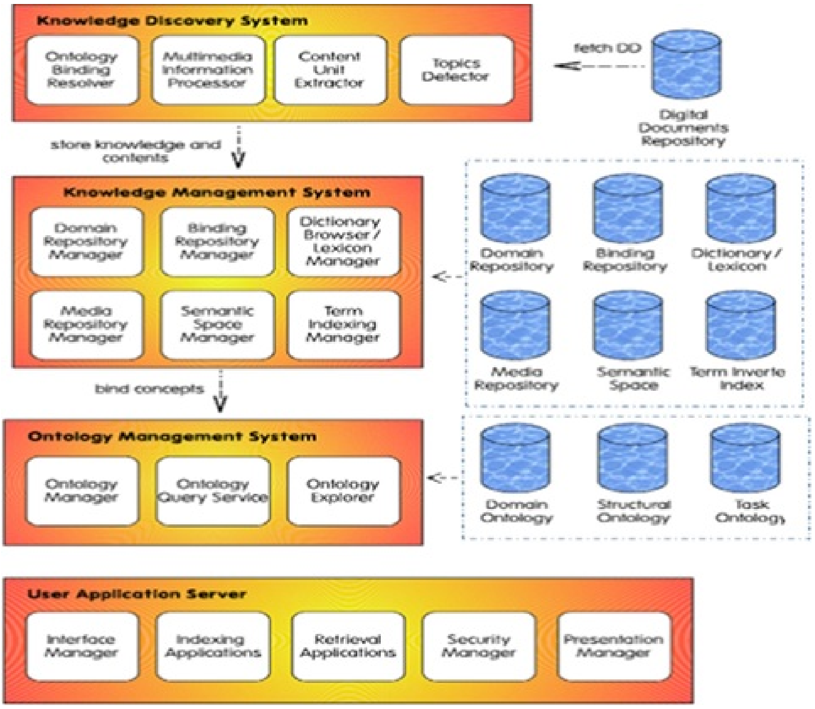


Figure 5.4: Architecture's components

retrieval. According to the introduced data model, it is possible to associate a digital document to a set of semantic concepts – retrievable by semi-automatic information extraction procedures and related to single content units of a document – and a set of keywords – defined as particular properties of the whole document. In the indexing stage, digital documents are picked up from DDR by a particular module called Knowledge Discovery System (KDS). The KDS analyses digital documents with the goal of obtaining useful knowledge from raw data. In particular, a Content Unit Extractor has the task of extracting (by a human-assisted process) content units from a document (and of generating an instance that can be stored in the system knowledge base), while the Multimedia Information Processor sub-module infers knowledge in terms of semantic concepts from the different kinds of multimedia data[79] (e.g. text, audio, video, image). Furthermore, a Topics Detector sub-module operates on the not-structured view of a document and aims at detecting using NLP techniques the most relevant topics for the whole document. Eventually, the Ontology Binding Resolver sub-module has the objective of creating – for each discovered concept/topic – a binding association with a node of the domain ontology. The extracted knowledge is then stored in the Semantic Knowledge Base (SKB) managed by a Knowledge Management System (KMS). The KMS performs indexing operations on the managed information, providing features for the browsing and the retrieval of the documents. The components of the SKB (and the related KMS managing modules) are described in the following.

Dictionary (for each supported language) It contains all the terms of a given language with the related possible meanings and some linguistic relationship (e.g. WordNet). Each dictionary is managed by an apposite management module, called Dictionary Browser.

Lexicon It contains all the terms known by the system: dictionary terms and named entities (names of people and organizations). The lexicon is managed by a proper module, called Lexicon Manager.

Term Inverted Index It is the data structure used for indexing terms inside documents. For each term known by the system (and contained in the lexicon) a posting list, that contains identifiers of documents and contents referring to terms with the related frequency, is created. The inverted index is managed by a Term Indexing Manager.

Semantic Space It allows the storage of atomic pieces of knowledge belonging to document content units, which are called document segments. It is an abstraction of a shared virtual memory space (with read/write methods) by which applications can exchange multimedia data. This space is called "semantic" because each element is associated to a particular structural ontology that allows for relating segments of the same content unit to content units of different documents. The Semantic Space Manger provides functionalities for reading, writing, removing and searching tuples in the space.

Domain Repository contains the description of application domain concepts and it is managed by a Domain repository Manager.

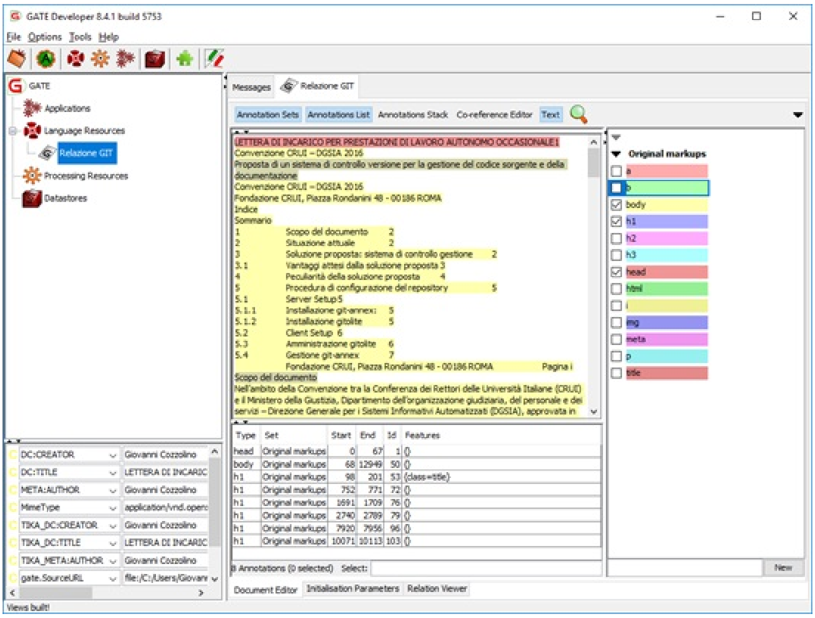
Binding Repository contains the associations between document and domain repository concepts and it is managed by a Binding Repository Manager.

Media Repository is an Object Relational DBMS able to manage different kinds of multimedia contents. It is managed by a particular module, called Media Repository Manager able to support classical multimedia query for the different kinds of multimedia data – e.g. query by example/feature for images, query by content/keywords for images and text, and so on.

The semantics associated to the data contained in the knowledge base is then managed by the *Ontology Management System* (OMS), that contains the ontology models used by the system. In particular, three kinds of ontologies were exploited (managed by an *Ontology Manager*): (i) a set of *domain ontologies* that relate the semantic concepts in a given domain, (ii) a set of *task ontologies* that determine the role/meaning of a content unit in a document and (iii) a set of structural ontologies that code the relationships between contents and segments. The *Ontology Explorer* allows browsing of the concepts in the ontologies, while the *Ontology Query Service* is a component devoted to execute queries on the ontologies. From the user point of view, the features provided by the system are the indexing of documents and the semantic retrieval of information. The application interfaces are realized both as web services and desktop programs (and managed by an *Interface Manager*). Finally, two different modules provide *security* and *presentation* management.

5.3.1 Log analysis with GATE

Log file documents offer a number of challenges, including a highly specialised vocabulary as well as common English words used with a domain specific sense.



LETTERA DI INCARICO PER PRESTAZIONI DI LAVORO AUTONOMO OCCASIONALE 1
Convenzione CRUI – DGSIA 2016
Proposta di un sistema di controllo versione per la gestione del codice sorgente e della documentazione
Convenzione CRUI – DGSIA 2016
Fondazione CRUI, Piazza Rondanini 48 - 00186 ROMA
Indice
Sommario
1 Scopo del documento 2
2 Situazione attuale 2
3 Soluzione proposta: sistema di controllo gestione 2
3.1 Vantaggi attesi dalla soluzione proposta 3
4 Peculiarità della soluzione proposta 4
5 Procedura di configurazione del repository 5
5.1 Server Setup 5
5.1.1 Installazione git-annex: 5
5.1.2 Installazione gitolite 5
5.2 Client Setup 6
5.3 Amministrazione gitolite 6
5.4 Gestione git-annex 7
Fondazione CRUI, Piazza Rondanini 48 - 00186 ROMA Pagina i
Scopo del documento
Nell'ambito della Convenzione tra la Conferenza dei Rettori delle Università Italiane (CRUI) e il Ministero della Giustizia, Dipartimento dell'organizzazione giudiziaria, del personale e dei servizi – Direzione Generale per i Sistemi Informativi Automatizzati (DGSIA), approvata in

DC:CREATOR	Giovanni Cozzolino
DC:TITLE	LETTERA DI INCARIC
META:AUTHOR	Giovanni Cozzolino
MimeType	application/vnd.openc
TIKA_DC:CREATOR	Giovanni Cozzolino
TIKA_DC:TITLE	LETTERA DI INCARIC
TIKA_META:AUTHOR	Giovanni Cozzolino
gate.SourceURL	file:/C:/Users/Giovann

Type	Set	Start	End	Id	Features
head	Original markups	0	67	1	{}
body	Original markups	68	12949	50	{}
h1	Original markups	98	201	53	{class=title}
h1	Original markups	752	771	72	{}
h1	Original markups	1691	1709	76	{}
h1	Original markups	2740	2789	79	{}
h1	Original markups	7920	7956	96	{}
h1	Original markups	10071	10113	103	{}

8 Annotations (0 selected) Select:

- □ a
- □ b
- ☑ body
- ☑ h1
- □ h2
- □ h3
- ☑ head
- □ html
- □ i
- □ img
- □ meta
- □ p
- □ title

Views built!

Figure 5.5: GATE annotation editor, showing the annotation sets of a test fil

In order to implement log analysis, the General Architecture for Text Engineering (GATE) was adopted. GATE is a framework for the development and deployment of language processing technology in large scale. It provides three types of resources: Language Resources (LRs) which collectively refer to data; Processing Resources (PRs) which are used to refer to algorithms; and Visualisation Resources (VRs) which represent visualisation and editing components. GATE can be used to process documents in different formats including plain text. When a document is loaded or opened in GATE, a document structure analyser is called upon which is in charge of creating a GATE document, a LR which will contain the text of the original document and one or more sets of annotations, one of which will contain the document mark-ups (for example HTML). Annotations are generally updated by PRs during text analysis - but they can also be created during annotation editing in the GATE GUI, as shown in Figure 5.5. Each annotation belongs to an annotation set and has a type, a pair of offsets (the span of text one wants to annotate), and a set of features and values that are used to encode the information.

GATE comes with a default information extraction system called ANNIE. The ANNIE system identifies generic concepts such as person names, locations, organisation, dates, etc. Therefore it is necessary to develop new rules, or adapt the existing ones, to target entities of specific domain applications that are not covered by ANNIE.

Moreover, the GATE architecture provides GATE Mímir, a multi-paradigm information management index and repository which can be used to index and search over text, annotations, semantic schemas (ontologies), and semantic meta-data (instance data). It allows queries that arbitrarily mix full-text, structural, linguistic and semantic queries and that can scale to gigabytes of text. Mímir provides a framework for implementing indexing and search functionality across all these data types.

Many GATE components can be adapted with little effort to process log files. The ANNIE Base Gazetteer can be initialized against a domain ontology in order to annotate many different domain specific concepts. Also many plugins can be used "as is" to extract information from event messages. The rest of this section, however, documents the resources included with GATE which are focused purely on processing log event messages.

Two main GATE features are relevant to face the problem of indexing and searching annotation data:

- a finite state transduction language called JAPE (Java Annotation Patterns Engine) that defines a rich regular expression language for matching within annotation graphs;

- GATE Developer includes ANNIC (ANNotations In Context), a visualisation tool.

The two features come together to a degree in that ANNIC allows queries using a JAPE-like language.

JAPE allows to recognise regular expressions in annotations on text documents. A JAPE grammar consists of a set of phases, each of which consists of a set of pattern/action rules. The phases run sequentially and constitute a cascade of finite state transducers over annotations. The left-hand-side (LHS) of the rules consist of an annotation pattern description, that is a pattern to be matched to the annotated text document. The right-hand-side (RHS) consists of annotation manipulation statements, that is what is to be done to the matched text. Annotations matched on the LHS of a rule may be referred to on the RHS by means of labels that are attached to pattern elements.

Considering the following JAPE rule entitled ‘EventName’:

Listing 5.1: EventName JAPE rule

```
Phase: EventRecognition
Input: Lookup
Options: control = appelt debug = true

Rule: EventName
(
{Lookup.majorType == event}
(
{Lookup.majorType == event}
)?
)
:event
-->
:event.EventName = {rule = "EventName"}
```

it will match text annotated with a ‘Lookup’ annotation with a ‘majorType’ feature of ‘event’, followed optionally by further text annotated as a ‘Lookup’ with ‘majorType’ of ‘event’. Once this rule has matched a sequence of text, the entire sequence is allocated a label by the rule, and in this case, the label is ‘event’. On the RHS, we refer to this span of text using the label given in the LHS; ‘event’. We say that this text is to be given an annotation of type ‘EventName’ and a ‘rule’ feature set to ‘EventName’.

ANNIC supports complex queries, such as a query that searches for event annotations followed by past tense verbs followed by device name.

All matching text ranges then appear in the lower half of the tool, with a graphical representation of the individual annotations concerned in the middle part.

For large datasets can be exploited the HPC Cloud computing paradigm, e.g. the Hadoop framework provides reliable data storage by Hadoop Distributed File System and MapReduce programming model which is a parallel processing system for large datasets. Hadoop distributed file system breaks-up input data and sends chunks of the original data to several machines in hadoop cluster to hold blocks of data. This approach helps to process log data in parallel using multiple machines and computes result efficiently. This approach reduces the response time as well as the load on to the end system.

6

Experimental results

Contents

Here, a set of experiments aimed at evaluating the framework's performances are presented. This section aims at demonstrating the capabilities of the approach in the context of a digital investigation examination. For this purpose, I perform an analysis of effectiveness and efficiency in a simulated cases. I start by showing how the data related to user behaviours were collected. Then, I discuss in detail the experimental setup in terms of execution time scalability, detection accuracy, and effectiveness.

6.1 Preparing the dataset

For the experimental campaign, it was set up a virtual machine, in which were stored more than 30,000 files, properly anonymized, coming from different sources and regarding different topics. Within this dataset, a set of 6,000 document belonging to specific topics and a set of queries with relevance judgements have been chosen as training set. The relevance judgements include relevant paragraphs marked within the whole document.

Further, to guarantee the repeatability of the experiments, it was implemented a batch script in order to perform a set of predefined known actions on the machine that simulate the user's behaviour. In this way the events generated by the script are always the same and the log analysis performances can be evaluated.

6.2 Experiments and results

To set up the experimentation, a disk image of the virtual machine containing the dataset is prepared in order to be fed to the Evidence Manager Module, that collects all the evidences retrieved from the disk image. Then, before being used as input of the Semantic Parser Module, all the evidences were pruned of all known files, such as operating system files with known hash. This has been done in order to consider only potentially relevant user files and to reduce the number of files to be processed by the system.

The objective is to evaluate the system correctness during the Document Analysis and Event log Analysis in automatically discovering relevant concepts within a forensic and in particular:

- User Personal Data
- System Events (Logon, Shutdown, Application run, etc.).
- Timeline of Significant Events (e.g., Browser chronology, instant messaging activities, File creation/editing/deletion, etc.).

6.2.1 Document Analysis Evaluation

Relevant concepts discovery procedures exploit a domain ontology built from scratch from the dataset, with the help of domain experts. The comparison between the proposed system semantic indexing output and the ground truth relies on the well-known recall and precision evaluated on the test set (24,000 documents). *Recall* measures the ratio between the relevant documents retrieved by the system and the total relevant ones (in the ground truth) with respect to a set given concept, while *Precision* measures the ratio between the relevant retrieved documents and the retrieved ones by the system.

The proposed method achieves an average recall value of 94.4% in finding a single concept for recall with respect to an average precision value of 83.7%, a 79.5% recall rate with a 86.8% precision rate in finding ten different concepts and, finally, an average recall of 74.4% with a precision of 97.3% in finding 20 concepts.

6.2.2 Log analysis evaluation

In this experiment, we show that we can automatically discover many abnormal behaviours in a logging system application. To measure ad-hoc information retrieval effectiveness in the standard way, we used a test collection consisting of three things:

1. A log file events collection
2. A test suite of information needs, expressible as queries
3. A set of relevance judgements, standardly a binary assessment.

For each event in the test collection is given a binary classification as either relevant or non-relevant. This decision is referred to as the *gold standard* or ground truth judgement of relevance. For this purpose, we manually labelled each distinct message, not only marking them as normal or abnormal, but also reporting the type of anomaly. We classified the events into four main categories (Critical, Error, Warning, Information) and for each category we further annotated each event accordingly to the event description. Table 6.1 shows the number of events of each category for the considered dataset.

Table 6.2 shows the manual labels. Each label group different kind of event, which can also belong to different event category. In Table 6.2 are reported the number of events for all the categories and the manual

Category	**N. of events**
Critical	42
Error	7344
Warning	1193
Information	31102
Total	**39682**

Table 6.1: Number of events for each category of considered dataset.

labels. In the last column is also reported the number of detected events through the presented methodology.

Figure 6.1 shows that the adopted methodology can detect a large fraction of relevant events in the test dataset.

In Figure 6.2 we can clearer observe that the detection rate is greater than 85% for those categories containing a considerable number of events. The methodology does also report some false positives, which are inevitable in any unsupervised technique, but the rate is never greater than 3% of each considered category.

Manual annotation	**Critical**	**Error**	**Warning**	**Information**	**Total**	**Detected**
Application Popup				16	16	15
bowser		8			8	7
BROWSER				13	13	12
BTHUSB			288	144	432	371
Disk		10	28		38	37
EventLog		24		1314	1338	1150
i8042prt			2		2	1
MEIx64			2	134	136	127
Microsoft-Windows-Dhcp-Client				380	380	330
Microsoft-Windows-DHCPv6-Client				250	250	245
Microsoft-Windows-Directory-Services-SAM				72	72	70
Microsoft-Windows-DistributedCOM		6813			6813	6336
Microsoft-Windows-DNS-Client			39		39	35
Microsoft-Windows-DriverFrameworks-UserMode				85	85	76
Microsoft-Windows-Eventlog		2			2	1
Microsoft-Windows-FilterManager				4740	4740	4171
Microsoft-Windows-GroupPolicy				23	23	21
Microsoft-Windows-HttpEvent			400	4	404	379
Microsoft-Windows-Kernel-Boot				1122	1122	987
Microsoft-Windows-Kernel-General				5960	5960	5900
Microsoft-Windows-Kernel-PnP			356		356	302
Microsoft-Windows-Kernel-Power	42			975	1017	884
Microsoft-Windows-Kernel-Processor-Power		16		4752	4768	4338
Microsoft-Windows-Ntfs				864	864	838
Microsoft-Windows-Power-Troubleshooter				100	100	96
Microsoft-Windows-Resource-Exhaustion-Detector			56		56	54
Microsoft-Windows-Setup				8	8	7
Microsoft-Windows-SetupPlatform				6	6	5
Microsoft-Windows-TerminalServices-RemoteConnectionManager				2	2	1
Microsoft-Windows-Time-Service			18	262	280	260
Microsoft-Windows-UserModePowerService				565	565	542
Microsoft-Windows-UserPnp				241	241	204
Microsoft-Windows-WindowsUpdateClient		55		2839	2894	2807
Microsoft-Windows-Wininit				132	132	125
Microsoft-Windows-Winlogon				248	248	210
Microsoft-Windows-WLAN-AutoConfig			4	130	134	123
Microsoft-Windows-WPDClassInstaller				100	100	86
NETwNb64				609	609	609
Schannel		3			3	2
Service Control Manager		39		4485	4524	4524
User32				434	434	381
VBoxNetLwf		286			286	280
Virtual Disk Service				2	2	1
volmgr		88			88	74
Volsnap				91	91	81
Total	**42**	**7344**	**1193**	**31102**	**39681**	**37105**

Table 6.2: Number of events for manually labelled events.

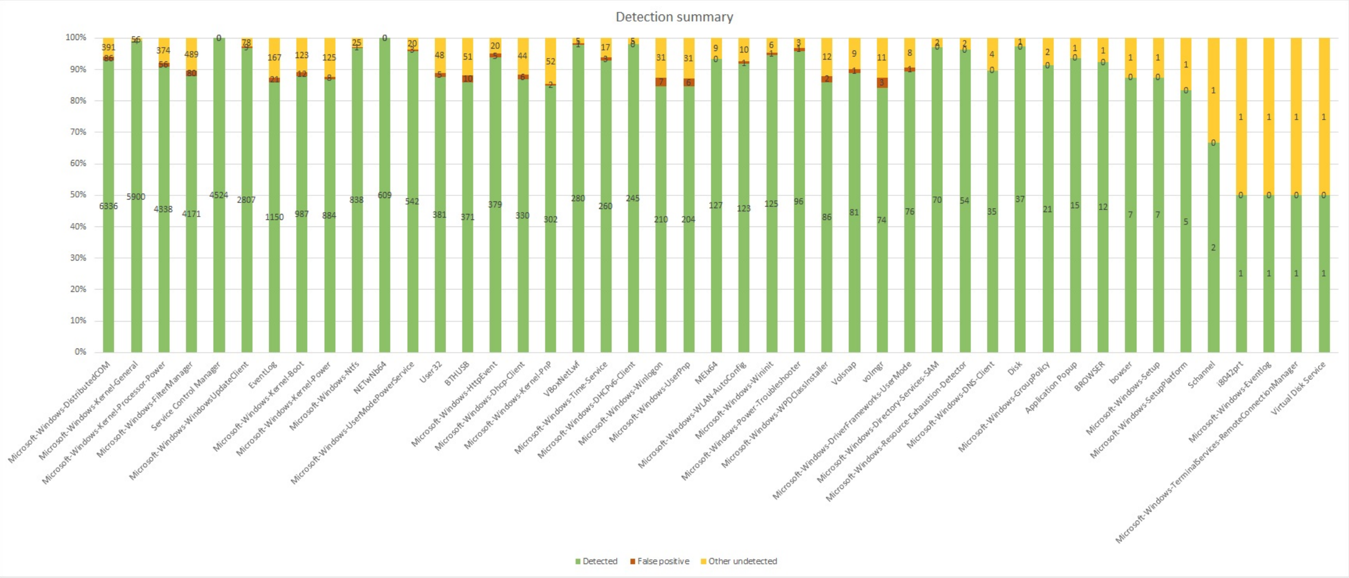


Figure 6.1: Numbers of detected, false positives and undetected events.

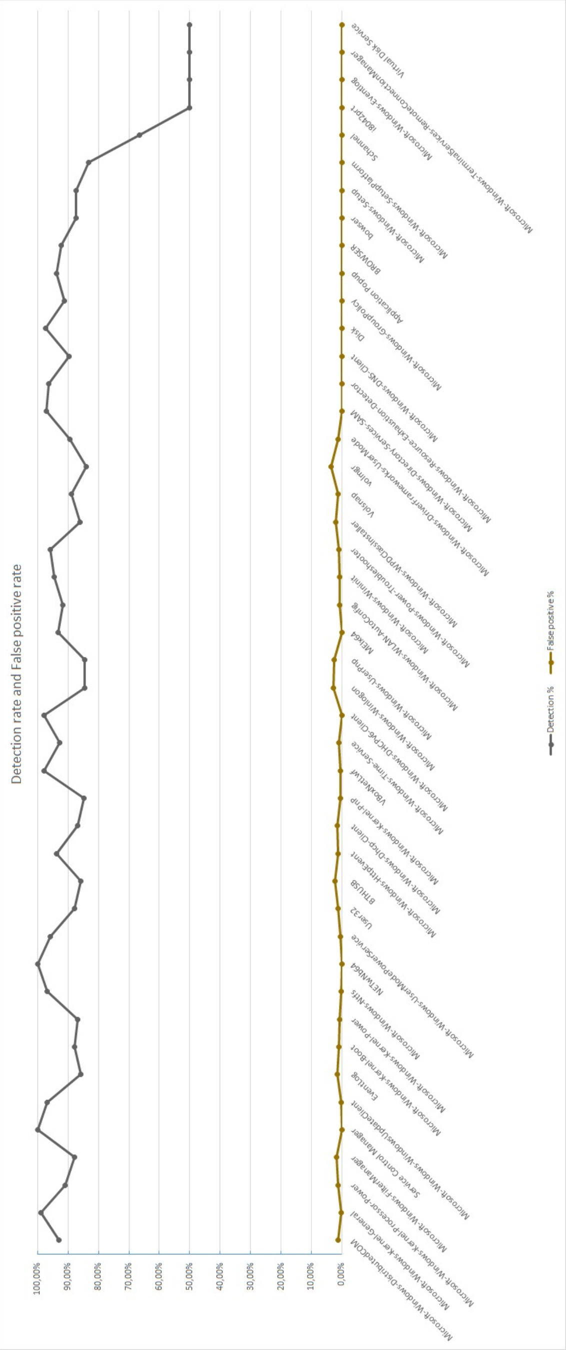


Figure 6.2: Detection rates and False positive rates for considered dataset.

6.3 Evaluating effectiveness and execution time

After successfully completing the data collection phase, the ontological representation phase and the enhancement (reasoning and rule evaluation) phase, the investigator starts the analysis by searching all significant correlations between events. To search for all traces of potentially suspicious executables is possible to execute a SPARQL query to retrieve all executables. Among the results of the query can be identified entries with a suspicious name and then can be figured out who interacts with this malware and in which circumstances.

Table 6.3 provides information on the volumes with the number of entries extracted by the Evidence Manager, the number of triples generated during the enhancement phase and finally the number of significant correlations found. Table 6.4 gives the execution times (in seconds) for the different phases of the process.

From Figure 6.3 we can see that execution time of each phase is linearly dependent by the number of entries processed in it. It can be assessed that the analysis phase is the most time-consuming step that requires pre-filtering of non-critical information to reduce the amount of data to analyse. Moreover, we can see that the number of triples does not increase linearly with the number of the acquired evidences. It depends on the type of each evidence which can carry more or less information

Table 6.3: Volumes of processed data through the reconstruction and analysis process

Criterion	Dataset No1	Dataset No2	Dataset No3
Extracted entries	6798	4263	15983
Generated triples	8786	10475	32498
Deduced triples	18	21	31
Correlations	1231	535	14982

Table 6.4: Volumes of processed data through the reconstruction and analysis process

Steps	Dataset No1	Dataset No2	Dataset No3
Collection	1.87	1.43	3.67
Representation	24.37	18.3	167.45
Reasoning	0.29	0.247	0.314
Analysis	776.3	241.3	12314.7

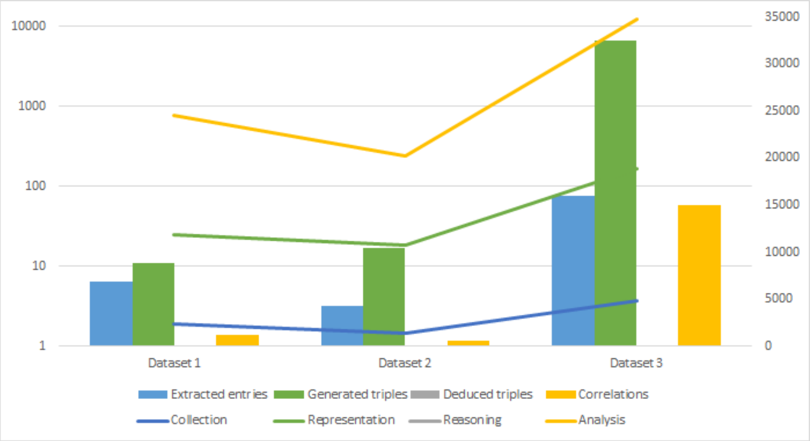


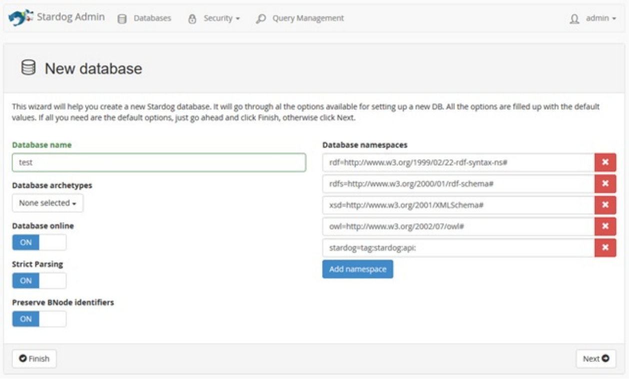
Figure 6.3: Analysis of execution time of each phase of methodology, related to entries processed.

and therefore require a larger or smaller number of triples to be modelled in the ontology.

Also, the variance of the results stresses the dependence of the methodology on the complexity of the input documents and logs.

6.3.1 How to reproduce the results

STEP 1 Add a database to Stardog (Figure 6.4)



New database

This wizard will help you create a new Stardog database. It will go through al the options available for setting up a new DB. All the options are filled up with the default values. If all you need are the default options, just go ahead and click Finish, otherwise click Next.

Database name: test

Database archetypes: None selected

Database online: ON

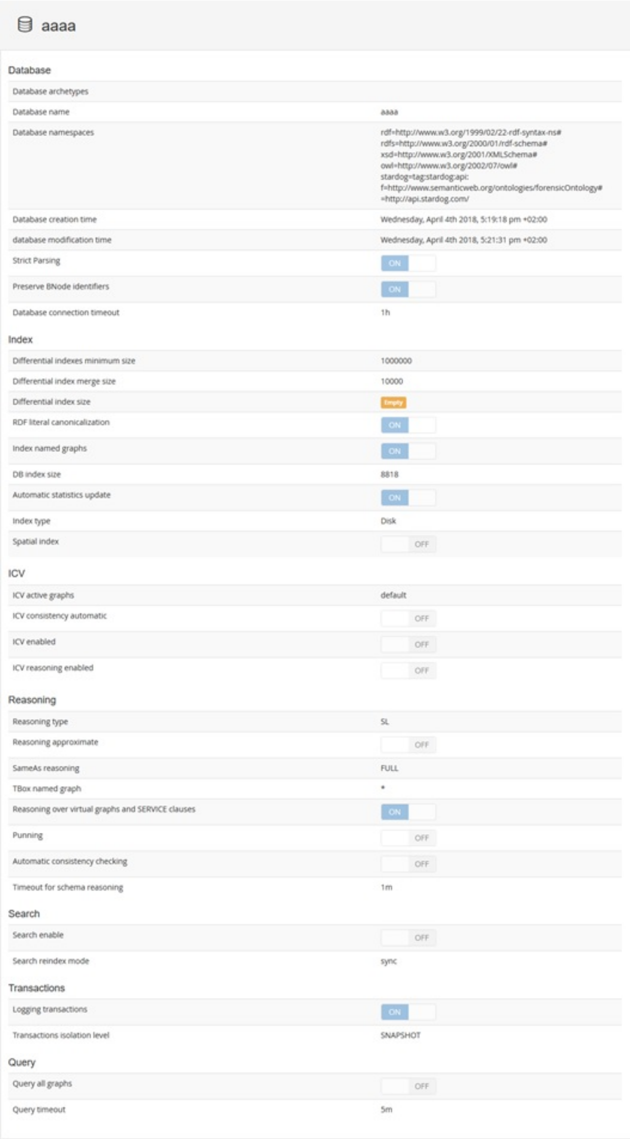
Strict Parsing: ON

Preserve BNode identifiers: ON

Database namespaces:
- rdf=http://www.w3.org/1999/02/22-rdf-syntax-ns#
- rdfs=http://www.w3.org/2000/01/rdf-schema#
- xsd=http://www.w3.org/2001/XMLSchema#
- owl=http://www.w3.org/2002/07/owl#
- stardog=tag:stardog:api:

Figure 6.4

Database full configuration (Figure 6.5)



aaaa

Database

Database archetypes	
Database name	aaaa

Database namespaces	rdf=http://www.w3.org/1999/02/22-rdf-syntax-ns#
	rdfs=http://www.w3.org/2000/01/rdf-schema#
	xsd=http://www.w3.org/2001/XMLSchema#
	owl=http://www.w3.org/2002/07/owl#
	stardog=tag:stardog:api:
	f=http://www.semanticweb.org/ontologies/forensicOntology#
	=http://api.stardog.com/

Database creation time	Wednesday, April 4th 2018, 5:19:18 pm +02:00
database modification time	Wednesday, April 4th 2018, 5:21:31 pm +02:00
Strict Parsing	ON
Preserve BNode identifiers	ON
Database connection timeout	1h

Index

Differential indexes minimum size	1000000
Differential index merge size	10000
Differential index size	Empty
RDF literal canonicalization	ON
Index named graphs	ON
DB index size	8818
Automatic statistics update	ON
Index type	Disk
Spatial index	OFF

ICV

ICV active graphs	default
ICV consistency automatic	OFF
ICV enabled	OFF
ICV reasoning enabled	OFF

Reasoning

Reasoning type	SL
Reasoning approximate	OFF
SameAs reasoning	FULL
TBox named graph	*
Reasoning over virtual graphs and SERVICE clauses	ON
Punning	OFF
Automatic consistency checking	OFF
Timeout for schema reasoning	1m

Search

Search enable	OFF
Search reindex mode	sync

Transactions

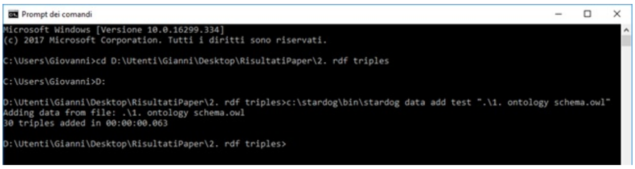
Logging transactions	ON
Transactions isolation level	SNAPSHOT

Query

Query all graphs	OFF
Query timeout	5m

Figure 6.5

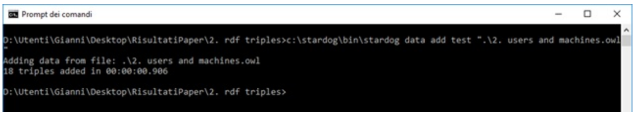
STEP 2 Add ontology schema to the database (we don't count these triples for the generated triple value) (Figure 6.6)



```
Microsoft Windows [Versione 10.0.16299.334]
(c) 2017 Microsoft Corporation. Tutti i diritti sono riservati.

C:\Users\Giovanni>cd D:\Utenti\Gianni\Desktop\RisultatiPaper\2. rdf triples

C:\Users\Giovanni>D:

D:\Utenti\Gianni\Desktop\RisultatiPaper\2. rdf triples>c:\stardog\bin\stardog data add test ".\1. ontology schema.owl"
Adding data from file: .\1. ontology schema.owl
30 triples added in 00:00:00.063

D:\Utenti\Gianni\Desktop\RisultatiPaper\2. rdf triples>
```

Figure 6.6

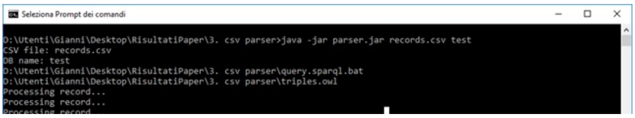
STEP 3 Add user and machine configuration to the database (we count these triples for the generated triple value)(Figure 6.7)



```
D:\Utenti\Gianni\Desktop\RisultatiPaper\2. rdf triples>c:\stardog\bin\stardog data add test ".\2. users and machines.owl
"
Adding data from file: .\2. users and machines.owl
18 triples added in 00:00:00.906

D:\Utenti\Gianni\Desktop\RisultatiPaper\2. rdf triples>
```

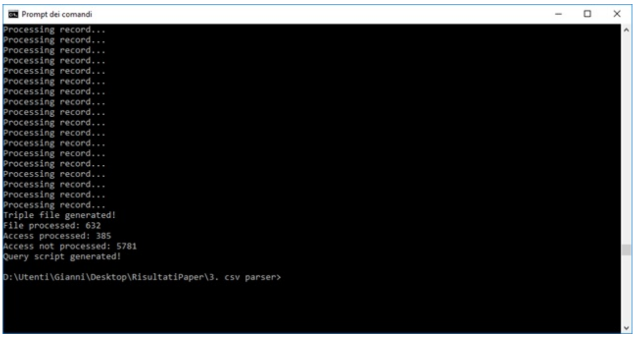
Figure 6.7

STEP 4 Run the CSV parser to generate RDF representation and the script to run correlation queries (Figure 6.8 and 6.9)



```
D:\Utenti\Gianni\Desktop\RisultatiPaper\3. csv parser>java -jar parser.jar records.csv test
CSV file: records.csv
DB name: test
D:\Utenti\Gianni\Desktop\RisultatiPaper\3. csv parser\query.sparql.bat
D:\Utenti\Gianni\Desktop\RisultatiPaper\3. csv parser\triples.owl
Processing record...
Processing record...
Processing record
```

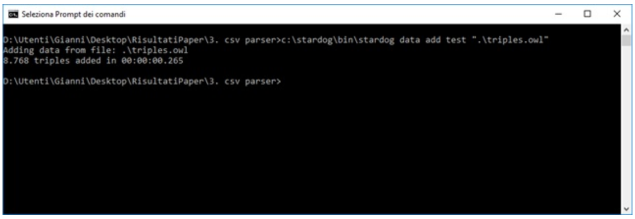
Figure 6.8



```
Processing record...
Processing record...
Processing record...
Processing record...
Processing record...
Processing record...
Processing record...
Processing record...
Processing record...
Processing record...
Processing record...
Processing record...
Processing record...
Processing record...
Processing record...
Processing record...
Processing record...
Processing record...
Triple file generated!
File processed: 632
Access processed: 385
Access not processed: 5781
Query script generated!

D:\Utenti\Gianni\Desktop\RisultatiPaper\3. csv parser>
```

Figure 6.9

STEP 5 Add generated RDF triples to the database (we count these triples for the generated triple value)(Figure 6.10)

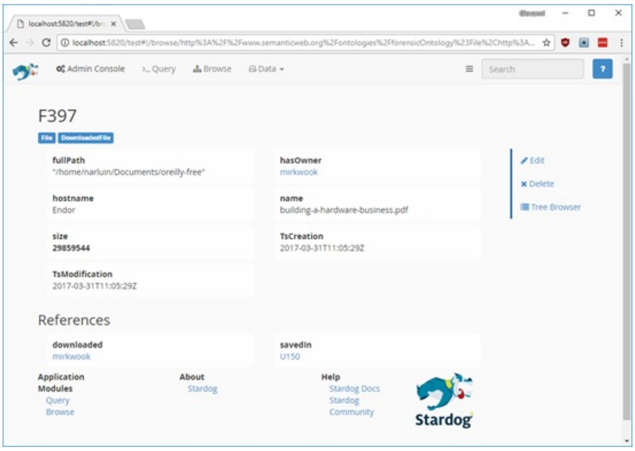


```
D:\Utenti\Gianni\Desktop\RisultatiPaper\3. csv parser>c:\stardog\bin\stardog data add test ".\triples.owl"
Adding data from file: .\triples.owl
8.768 triples added in 00:00:00.265

D:\Utenti\Gianni\Desktop\RisultatiPaper\3. csv parser>
```

Figure 6.10

You can notice that instances of File, Access and URL have no correlation.

STEP 6 Run the bat file for the correlations. It adds them into the knowledge base... See the *results.txt* for correlation triples number (Figure 6.11)



F397

File DownloadedFile

fullPath
"/home/narluin/Documents/oreilly-free"

hasOwner
mirkwook

hostname
Endor

name
building-a-hardware-business.pdf

size
29859544

TsCreation
2017-03-31T11:05:29Z

TsModification
2017-03-31T11:05:29Z

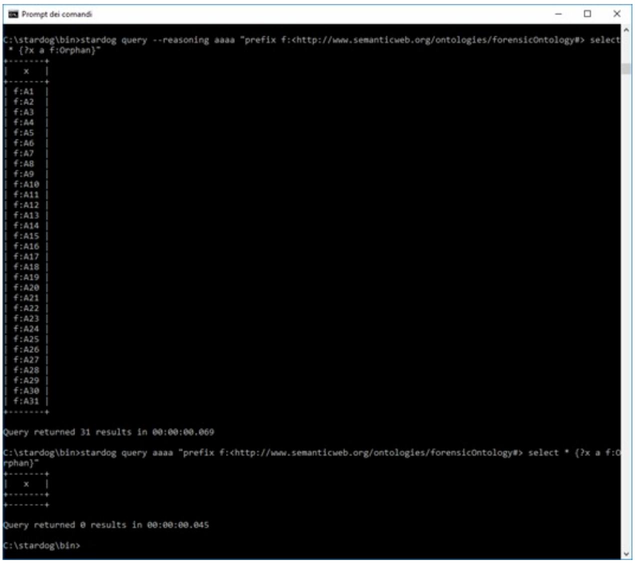
References

downloaded
mirkwook

savedIn
U150

Figure 6.11

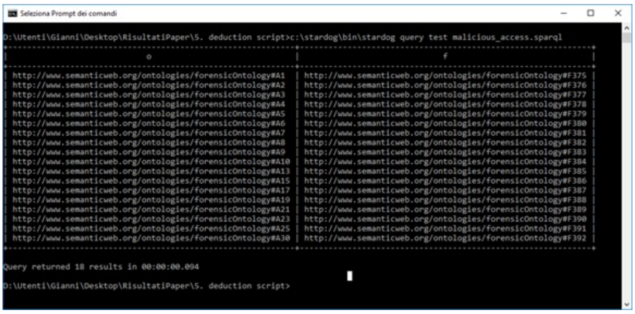
STEP 7 Add the SWRL rule to the database and query it with reasoning on (Figure 6.12)



```
C:\stardog\bin>stardog query --reasoning aaaa "prefix f:<http://www.semanticweb.org/ontologies/forensicOntology#> select
 * {?x a f:Orphan}"
+-------+
|   x   |
+-------+
| f:A1  |
| f:A2  |
| f:A3  |
| f:A4  |
| f:A5  |
| f:A6  |
| f:A7  |
| f:A8  |
| f:A9  |
| f:A10 |
| f:A11 |
| f:A12 |
| f:A13 |
| f:A14 |
| f:A15 |
| f:A16 |
| f:A17 |
| f:A18 |
| f:A19 |
| f:A20 |
| f:A21 |
| f:A22 |
| f:A23 |
| f:A24 |
| f:A25 |
| f:A26 |
| f:A27 |
| f:A28 |
| f:A29 |
| f:A30 |
| f:A31 |
+-------+

Query returned 31 results in 00:00:00.069

C:\stardog\bin>stardog query aaaa "prefix f:<http://www.semanticweb.org/ontologies/forensicOntology#> select * {?x a f:O
rphan}"
+-------+
|   x   |
+-------+
+-------+

Query returned 0 results in 00:00:00.045

C:\stardog\bin>
```

Figure 6.12

STEP 8 Run the query for deductions. We don't add the results to the knowledge base because the results are not certain (Figure 6.13)



```
D:\Utenti\Gianni\Desktop\RisultatiPaper\5. deduction script>c:\stardog\bin\stardog query test malicious_access.sparql
+------------------------------------------------------------+--------------------------------------------------------------+
|                             o                              |                              f                               |
+------------------------------------------------------------+--------------------------------------------------------------+
| http://www.semanticweb.org/ontologies/forensicOntology#A1  | http://www.semanticweb.org/ontologies/forensicOntology#F375 |
| http://www.semanticweb.org/ontologies/forensicOntology#A2  | http://www.semanticweb.org/ontologies/forensicOntology#F376 |
| http://www.semanticweb.org/ontologies/forensicOntology#A3  | http://www.semanticweb.org/ontologies/forensicOntology#F377 |
| http://www.semanticweb.org/ontologies/forensicOntology#A4  | http://www.semanticweb.org/ontologies/forensicOntology#F378 |
| http://www.semanticweb.org/ontologies/forensicOntology#A5  | http://www.semanticweb.org/ontologies/forensicOntology#F379 |
| http://www.semanticweb.org/ontologies/forensicOntology#A6  | http://www.semanticweb.org/ontologies/forensicOntology#F380 |
| http://www.semanticweb.org/ontologies/forensicOntology#A7  | http://www.semanticweb.org/ontologies/forensicOntology#F381 |
| http://www.semanticweb.org/ontologies/forensicOntology#A8  | http://www.semanticweb.org/ontologies/forensicOntology#F382 |
| http://www.semanticweb.org/ontologies/forensicOntology#A9  | http://www.semanticweb.org/ontologies/forensicOntology#F383 |
| http://www.semanticweb.org/ontologies/forensicOntology#A10 | http://www.semanticweb.org/ontologies/forensicOntology#F384 |
| http://www.semanticweb.org/ontologies/forensicOntology#A13 | http://www.semanticweb.org/ontologies/forensicOntology#F385 |
| http://www.semanticweb.org/ontologies/forensicOntology#A15 | http://www.semanticweb.org/ontologies/forensicOntology#F386 |
| http://www.semanticweb.org/ontologies/forensicOntology#A17 | http://www.semanticweb.org/ontologies/forensicOntology#F387 |
| http://www.semanticweb.org/ontologies/forensicOntology#A19 | http://www.semanticweb.org/ontologies/forensicOntology#F388 |
| http://www.semanticweb.org/ontologies/forensicOntology#A21 | http://www.semanticweb.org/ontologies/forensicOntology#F389 |
| http://www.semanticweb.org/ontologies/forensicOntology#A23 | http://www.semanticweb.org/ontologies/forensicOntology#F390 |
| http://www.semanticweb.org/ontologies/forensicOntology#A25 | http://www.semanticweb.org/ontologies/forensicOntology#F391 |
| http://www.semanticweb.org/ontologies/forensicOntology#A30 | http://www.semanticweb.org/ontologies/forensicOntology#F392 |
+------------------------------------------------------------+--------------------------------------------------------------+

Query returned 18 results in 00:00:00.094

D:\Utenti\Gianni\Desktop\RisultatiPaper\5. deduction script>
```

Figure 6.13

6.4 Conclusions

In this work, a reusable methodology based on semantic representation, integration and correlation of digital evidence is proposed and an architecture that implements it is presented. The approach is based on ontologies defining domain of digital incidents and on a set of tools for extracting information from disk image, populating the ontology, inferring and analysing new facts. The use of an ontology allows for the representation of knowledge with a unified model and for simplifying the building of analysis processes.

Despite the aforementioned advantages, the approach has still some limitations that will be studied in future works. Performances can be improved, especially for what the execution time of the analysis phase concerns. Regarding the extraction phase, additional sources have to be integrated to reach a deeper analysis of an incident.

In future work, it is planned to improve extraction phase by adding more information sources related to additional devices, such as mobile devices (Android or iOS) or IoT devices. From the forensic perspective, also IoT devices can provide important evidences that could help in the investigation process. In the main while IoT causes some challenges for forensics examiner including but not limited to the location of data and heterogeneous nature of IoT devices such as differences in operating systems and communication standards.[80,81] The proposed methodology can be feasible also to IoT devices in the measure that information extracted by them can be represented through RDF format.

For the analysis, can be integrated new tools for event correlation based on pattern matching algorithm to detect illegal actions by identifying specific event sequences. Concerning the interface and query layer, it will be proposed some enhancement regarding a graph visualization tool that easily show instances correlations.

In conclusion, through the proposed approach a digital forensics examiner is able to easily extract the knowledge related to a digital incident.

This knowledge can be enriched using a semantic representation that alleviate heterogeneity problem of forensic tools outputs. The reasoning process allows to deduce new knowledge from the existing one, improving analytical skills by allowing easy and fast ways to integrate and correlate forensic evidences. The use of a SPARQL interface allow to understand the interactions between facts.

Bibliography

[1] Weiqiang Liu, Saket Srivastava, Liang Lu, Máire O'Neill, and Earl E Swartzlander. Are qca cryptographic circuits resistant to power analysis attack? *IEEE Transactions on Nanotechnology*, 11(6):1239–1251, 2012.

[2] Pieraugusto Pozzi, Roberto Masotti, and Marco Bozzetti. *Crimine virtuale, minaccia reale: ICT security: politiche e strumenti di prevenzione*, volume 7. Franco Angeli, 2004.

[3] Ian Walden. *Computer crimes and digital investigations*. Oxford University Press Oxford, 2007.

[4] A Attanasio, F Cajani, G Costabile, and W Vannini. Lo stato dell'arte della computer forensics in italia. *IISFA Memberbook*, pages 123–151, 2012.

[5] Albert Marcella Jr and Doug Menendez. *Cyber forensics: a field manual for collecting, examining, and preserving evidence of computer crimes*. Auerbach Publications, 2007.

[6] Cesare Maioli. Dar voce alle prove: elementi di informatica forense. 2004.

[7] Warren G Kruse II and Jay G Heiser. *Computer forensics: incident response essentials*. Pearson Education, 2001.

[8] G Fagioli and A Ghirardini. Digital forensics, 2013.

[9] Luca Lupária and Giovanni Ziccardi. *Investigazione penale e tecnologia informatica. L'accertamento del reato tra progresso scientifico e garanzie fondamentali*. Giuffrè Editore, 2007.

[10] Michael A Caloyannides. *Computer forensics and privacy*. Artech House Publishers, 2001.

[11] John R Vacca. *Computer forensics: computer crime scene investigation*. Charles River Media, Inc., 2002.

[12] Eoghan Casey. *Digital evidence and computer crime: Forensic science, computers, and the internet*. Academic press, 2011.

[13] H. Seo, Z. Liu, J. Choi, and H. Kim. Multi-precision squaring for public-key cryptography on embedded microprocessors. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8250 LNCS:227–243, 2013.

[14] Peter Mell, Tim Grance, et al. The nist definition of cloud computing. 2011.

[15] Terrence V Lillard. *Digital forensics for network, Internet, and cloud computing: a forensic evidence guide for moving targets and data*. Syngress Publishing, 2010.

[16] Rainer Poisel, Erich Malzer, and Simon Tjoa. Evidence and cloud computing: The virtual machine introspection approach. *JoWUA*, 4(1):135–152, 2013.

[17] Sebastiano Battiato, Giovanni Maria Farinella, and Giovanni Puglisi. Image/video forensics: Casi di studio. *IISFA memberbook*, pages 261–292, 2011.

[18] S Battiato, G Messina, and R Rizzo. Image forensics. contraffazione digitale e identificazione della camera di acquisizione: Status e prospettive. *IISFA Memberbook 2009 DIGITAL FORENSICS*, 2009.

[19] Oreste Dominioni. *La prova penale scientifica: gli strumenti scientifico-tecnici nuovi o controversi e di elevata specializzazione.* Giuffrè, 2005.

[20] Paolo Tonini and Carlotta Conti. *Il diritto delle prove penali.* Giuffrè Editore, 2012.

[21] Gianluca D'Aiuto and Luigi Levita. *I reati informatici. Disciplina sostanziale e questioni processuali.* Giuffrè Editore, 2012.

[22] Scientific Working Group on Digital Evidence (SWGDE) International Organization on Digital Evidence (IOCE). Digital evidence: Standards and principles, 2014.

[23] Brian Carrier and Eugene H Spafford. An event-based digital forensic investigation framework. In *Digital forensic research workshop*, pages 11–13, 2004.

[24] Bilal Khan, Khaled S Alghathbar, Syed Irfan Nabi, and Muhammad Khurram Khan. Effectiveness of information security awareness methods based on psychological theories. *African Journal of Business Management*, 5(26):10862, 2011.

[25] Bradley Lawrence Schatz. *Digital evidence : representation and assurance.* PhD thesis, Queensland University of Technology, 2007.

[26] Andrew S Tanenbaum. *Structured computer organization.* Pearson Education India, 2016.

[27] Nicholas Negroponte, Franco Filippazzi, and Giuliana Filippazzi. *Essere digitali.* Sperling & Kupfer Milano, 1995.

[28] Lorenzo Picotti. La ratifica della convenzione cybercrime del consiglio d'europa. profili di diritto penale sostanziale. In *Diritto penale e processo*, volume 6, pages 700–716, 2008.

[29] L LUPÀRIA. La ratifica della convenzione cybercrime del consiglio d'europa. i profili processuali. In *Dir. pen. proc*, volume 6, page 717, 2008.

[30] Francesco Cajani. La convenzione di budapest nell'insostenibile salto all'indietro del legislatore italiano, ovvero: quello che le norme non dicono. *Ciberspazio e diritto*, 11(1):185–210, 2010.

[31] P Perri. Computer forensics (indagini informatiche). 2011.

[32] Giovanni Neri. *Criminologia e diritto penale dell'economia*, volume 10. Edizioni Nuova Cultura, 2014.

[33] Scientific Working Group on Digital Evidence (SWGDE). Best practices for computer forensics, 2014.

[34] International Organization on Computer Evidence (IOCE). Best practices for computer forensics, 2014.

[35] Scientific Working Group on Imaging Technology (SWGIT). Best practices on imaging, 2015.

[36] National Institute of Standards and Technology (NIST). Forensic science — digital and multimedia evidence, 2013.

[37] AGIS Project founded by European COmmission. The admissibility of electronic evidence at court: Fighting against high tech crime, 2007.

[38] Onofrio Signorile. Computer forensics guidelines: un approccio metodico-procedurale per l'acquisizione e l'analisi delle digital evidence. *Ciberspazio e diritto*, 10(2):197–209, 2009.

[39] Brian Carrier, Eugene H Spafford, et al. Getting physical with the digital investigation process. *International Journal of digital evidence*, 2(2):1–20, 2003.

[40] John R Vacca. *Computer Forensics: Computer Crime Scene Investigation (Networking Series)(Networking Series)*. Charles River Media, Inc., 2005.

[41] G Convey. Collecting volatile and non-volatile data. *IISA Journal, August*, 2007.

[42] D. Cowen. *Computer Forensics InfoSec Pro Guide*. InfoSec Pro guide. McGraw-Hill Education, 2013.

[43] P. Carretta, A. Cilli, A. Iacoviello, A. Grillo, and F. Trocchi. *L'acquisizione del documento informatico. Indagini penali e amministrative*. Laurus Robuffo, 2012.

[44] M. Solomon, D. Barrett, and N. Broom. *Computer Forensics JumpStart*. Wiley, 2015.

[45] Svein Willassen. Forensic analysis of mobile phone internal memory. In Mark Pollitt and Sujeet Shenoi, editors, *Advances in Digital Forensics*, pages 191–204, Boston, MA, 2005. Springer US.

[46] S. Aterno. *Computer forensics e indagini digitali: manuale tecnico-giuridico e casi pratici*. Number v. 1 in Computer forensics e indagini digitali: manuale tecnico-giuridico e casi pratici. Experta, 2011.

[47] Bill Nelson, Amelia Phillips, and Christopher Steuart. *Guide to Computer Forensics and Investigations.* Course Technology Press, Boston, MA, United States, 4th edition, 2009.

[48] Song Mao, Azriel Rosenfeld, and Tapas Kanungo. Document structure analysis algorithms: a literature survey. In *Document Recognition and Retrieval X*, volume 5010, pages 197–208. International Society for Optics and Photonics, 2003.

[49] Fernando Gomez. A representation of complex events and processes for the acquisition of knowledge from texts. *Knowledge-Based Systems*, 10(4):237–251, 1998.

[50] Yu-Chieh Wu. Integrating statistical and lexical information for recognizing textual entailments in text. *Knowledge-Based Systems*, 40:27–35, 2013.

[51] Graeme D Kennedy. *An introduction to corpus linguistics.* London ; New York : Longman, 1998. Bibliography: p295-309. - Includes index.

[52] Anna Siewierska. *Language*, 81(3):737–740, 2005.

[53] F. Mancini M. Vedovelli M. Voghera. De Mauro, T. Lessico di frequenza dell'italiano parlato. *International Journal of Corpus Linguistics*, 2(2):302–308, 1997.

[54] Ian Golledge. Online self-organised map classifiers as text filters for spam email detection. 2009.

[55] Kyo Kageura and Bin Umino. Methods of automatic term recognition - a review, 1996.

[56] Kenneth Ward Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Comput. Linguist.*, 16(1):22–29, March 1990.

[57] Ted Dunning. Accurate methods for the statistics of surprise and coincidence. *Comput. Linguist.*, 19(1):61–74, March 1993.

[58] W.Y. Wong, Wei Liu, and Mohammed Bennamoun. *Determination of Unithood and Termhood for Term Recognition*, volume 2, pages 500–529. IGI Global, 2009.

[59] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513 – 523, 1988.

[60] Sergio Decherchi, Simone Tacconi, Judith Redi, Alessio Leoncini, Fabio Sangiacomo, and Rodolfo Zunino. Text clustering for digital forensics analysis. In *Computational Intelligence in Security for Information Systems*, pages 29–36. Springer, 2009.

[61] Philipp Cimiano. *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications.* Springer-Verlag, Berlin, Heidelberg, 2006.

[62] Paola Velardi, Roberto Navigli, Alessandro Cucchiarelli, and Francesca Neri. Evaluation of OntoLearn, a methodology for automatic population of domain ontologies. In Paul Buitelaar, Philipp Cimiano, and Bernardo Magnini, editors, *Ontology Learning from Text: Methods, Applications and Evaluation*. IOS Press, 2006.

[63] Paul Buitelaar, Daniel Olejnik, and Michael Sintek. A protégé plug-in for ontology extraction from text based on linguistic analysis. In Christoph J. Bussler, John Davies, Dieter Fensel, and Rudi Studer, editors, *The Semantic Web: Research and Applications*, pages 31–44, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg.

[64] Nathalie Aussenac-Gilles, Sylvie Despres, and Sylvie Szulman. The terminae method and platform for ontology engineering from texts. In *Proceedings of the 2008 Conference on Ontology Learning and Population: Bridging the Gap Between Text and Knowledge*, pages 199–223, Amsterdam, The Netherlands, The Netherlands, 2008. IOS Press.

[65] Alexander Maedche and Raphael Volz. The text-to-onto ontology extraction and maintenance system. In *International Conference on Data Mining (ICDM), San Jose, California, USA, November 29 - December 2, 2001, ICDM-Workshop on Integrating Data Mining and Knowledge Management*, 2001.

[66] Philipp Cimiano and Johanna Völker. Text2onto: A framework for ontology learning and data-driven change discovery. In *Proceedings of the 10th International Conference on Natural Language Processing and Information Systems*, NLDB'05, pages 227–238, Berlin, Heidelberg, 2005. Springer-Verlag.

[67] Ben Martini and Kim-Kwang Raymond Choo. Remote programmatic vcloud forensics: a six-step collection process and a proof of concept. In *2014 IEEE 13th International Conference on Trust, Security and Privacy in Computing and Communications*, pages 935–942. IEEE, 2014.

[68] Quang Do, Ben Martini, and Kim-Kwang Raymond Choo. A forensically sound adversary model for mobile devices. *PloS one*, 10(9):e0138449, 2015.

[69] Niken Dwi Wahyu Cahyani, Ben Martini, Kim-Kwang Raymond Choo, and AKBP Al-Azhar. Forensic data acquisition from cloud-of-things devices: windows smartphones as a case study. *Concurrency and Computation: Practice and Experience*, 2016.

[70] C. Butler. *Structure and Function: From clause to discourse and beyond.* From Clause to Discourse and Beyond: A Guide to Three Major Structural-functional Theories. J. Benjamins Publishing Company, 2003.

[71] F. Amato, V. Casola, N. Mazzocca, and S. Romano. A semantic-based document processing framework: A security perspective. In *2011 International Conference on Complex, Intelligent, and Software Intensive Systems*, pages 197–202, June 2011.

[72] F. Amato, V. Casola, N. Mazzocca, and S. Romano. A semantic approach for fine-grain access control of e-health documents. *Logic Journal of the IGPL*, 21(4):692–701, Aug 2013.

[73] Bruno Grilheres, Stephan Brunessaux, and Philippe Leray. Combining classifiers for harmful document filtering. In *Coupling Approaches, Coupling Media and Coupling Languages for Information Retrieval*, RIAO '04, pages 173–185, Paris, France, France, 2004. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE.

[74] Flora Amato, Antonino Mazzeo, Antonio Penta, and Antonio Picariello. Knowledge representation and management for e-government documents. In Antonino Mazzeo, Roberto Bellini, and Gianmario Motta, editors, *E-Government Ict Professionalism and Competences Service Science*, pages 31–40, Boston, MA, 2008. Springer US.

[75] Francesco Colace, Massimo De Santo, Luca Greco, Vincenzo Moscato, and Antonio Picariello. A collaborative user-centered framework for recommending items in online social networks. *Comput. Hum. Behav.*, 51(PB):694–704, October 2015.

[76] Vincenzo Moscato Francesco Piccialli Angelo Chianese, Fiammetta Marulli. Smartweet: A location-based smart application for exhibits and museums. *3D Digital Imaging and Modeling, International Conference on*, pages 408–415.

[77] Borislav Popov, Atanas Kiryakov, Angel Kirilov, Dimitar Manov, Damyan Ognyanoff, and Miroslav Goranov. Kim – semantic annotation platform. In Dieter Fensel, Katia Sycara, and John Mylopoulos, editors, *The Semantic Web - ISWC 2003*, pages 834–849, Berlin, Heidelberg, 2003. Springer Berlin Heidelberg.

[78] Dipasree Pal, Mandar Mitra, and Kalyankumar Datta. Improving query expansion using wordnet. *J. Assoc. Inf. Sci. Technol.*, 65(12):2469–2478, December 2014.

[79] Francesco Colace, Massimo De Santo, Luca Greco, Vincenzo Moscato, and Antonio Picariello. A collaborative user-centered framework for recommending items in online social networks. *Computers in Human Behavior*, 51:694 – 704, 2015. Computing for Human Learning, Behaviour and Collaboration in the Social and Mobile Networks Era.

[80] Ibrar Yaqoob, Ibrahim Abaker Targio Hashem, Arif Ahmed, SM Ahsan Kazmi, and Choong Seon Hong. Internet of things forensics: Recent advances, taxonomy, requirements, and open challenges. *Future Generation Computer Systems*, 92:265–275, 2019.

[81] Jacques Boucher and Nhien-An Le-Khac. Forensic framework to identify local vs synced artefacts. *Digital Investigation*, 24:S68–S75, 2018.