Journal of Universal Computer Science, vol. 29, no. 12 (2023), 1482-1509
submitted: 21/11/2022, accepted: 12/9/2023, appeared: 28/12/2023 CC BY-ND 4.0

OntoFoCE and ObE Forensics. Email-traceability Supporting Tools for Digital Forensics

Herminia Beatriz Parra de Gallo
(Universidad Católica de Salta, Salta, Argentina
https://orcid.org/0000-0002-3230-3108, bgallo@ucasal.edu.ar)

Marcela Vegetti
(Development and Design Institute, INGAR (CONICET-UTN), Santa Fe, Argentina
https://orcid.org/0000-0003-4016-1717, mvegetti@santafe-conicet.gov.ar)

Abstract: This paper shows the research conducted to respond to a continuous requirement of justice regarding the application of scientifically supported forensic tools. Considering ontological engineering as the appropriate framework to respond to this requirement, the article presents OntoFoCE (Spanish abbreviation for Ontology for Electronic Mail Forensics), a specific ontology for the forensic analysis of emails. The purpose of this ontology is to help the computer expert in the validation of an email presented as judicial evidence. OntoFoCE is the fundamental component of the ObE Forensics (Ontology-based Email Forensics) tool. Although there are numerous forensic tools to analyze emails, the originality of the one proposed here lies in the implementation of semantic technologies to represent the traceability of the email transmission process. From that point on, it is possible to provide answers to the items of digital evidence subject to the expert examination. These answers make it possible to support these evidence items in the forensic analysis of an email and to guarantee the gathering of scientifically and technically accepted results that are valid for justice. Thus, the research question that is tried to be answered is: Is it possible to apply ontological engineering as a scientific support to design and develop a forensic tool that allows automatic answers to the evidence items subject to the expert examination in the forensic analysis of emails?

Keywords: Ontology, Forensics, Email, Email Header
Categories: D.3.1, H.1.m, H.2.3, H4.3, I.2.1
DOI: 10.3897/jucs.97822

1 Introduction

Digital Forensics is defined as “the use of scientifically proven and derived methods towards the preservation, recollection, validation, identification, analysis, interpretation, documentation and presentation of digital evidence from digital sources; with the purpose of facilitating or promoting the reconstruction of events, which are considered criminal, or helping to anticipate unauthorized actions that may be detrimental to planned operations” [Palmer, 2001].

In the case of emails, the forensic analysis consists in obtaining from the email header all the necessary data to trace the route followed by the email from a sender’s account to a recipient’s account. If the traceability of the delivery can be established, its authenticity can be validated, and the admissibility of the email as digital evidence is supported.

Procedural Law requires that all expert evidence meet technical and scientific criteria that formally support the results obtained. The procedure that is carried out must follow valid protocols to review the digital evidence. Therefore, it is of interest to generate tools based on scientific criteria that help the expert in their task.

During the email transmission process, the header stores the identification data of each device through which the email traveled. In the forensic analysis of an email, the expert examines the header and identifies the active devices and servers through the IP address. As a result, all the devices participating in the transmission process are individualized, so the judges' requests about the evidence can be answered. These requirements are usually the same for the expert analysis of an email, and they are known as "evidence items subject to expert examination" (EISTEEs). The expert must respond to these items by producing a report for the judge with the answers found in their analysis. It is important to highlight how restrictive the EISTEEs are for the expert's activity, as the expert must respond to them without providing an incomplete or excessive answer.

If a single email is examined, the expert can analyze it manually without a forensic tool that processes the header and shows the relevant data. However, a manual analysis is very complicated if done on numerous emails because it requires much time and effort and does not guarantee correct or complete results. Although several tools can make this analysis easier, none of them automatically provides the answers that the expert must give to the judge's questions.

Having a tool in which the expert can enter the email headers to be examined, select the required EISTEEs, and automatically produce the report to deliver to the judge is very useful. This tool helps the computer expert keep the integrity, confidentiality, and availability of the evidence required in a judicial process.

The formalization of the information in an email's header is an essential requirement to automate the answers to the EISTEEs. In this sense, ontologies are an excellent tool for formalization.

In addition, linking EISTEEs with competency questions (the requirements that an ontology must meet) contributes to the automatic generation of answers to the EISTEEs. So, the research question this paper tries to solve is: Is it possible to apply ontological engineering as a scientific support to design and develop a forensic tool that allows the automatic response to the evidence items subject to expert examination in the forensic analysis of emails?

This article presents an ontology called OntoFoCE (Spanish abbreviation for Ontology for Electronic Mail Forensics) that offers a specific framework for the forensic analysis of emails. It formalizes the information contained in email headers and provides an answer to the research question. OntoFoCE is the fundamental component of the ObE Forensics (Ontology-based Email Forensics) tool, which provides automatic responses to the EISTEEs and helps the IT expert verify the authenticity of an email presented as judicial evidence.

Although there are many and varied forensic tools for email analysis, the originality of ObE Forensics lies in the use of semantic technologies to represent the traceability of the email transmission process to support the automatic generation of the responses to the EISTEEs usually required from experts. This characteristic guarantees obtaining scientifically and technically accepted results considered valid according to justice.

This work is organized as follows: Section 2 explains the problem of the forensic analysis of emails. Section 3 describes the state of the art of the topic. Section 4 defines

OntoFoCE from the ontology development phases, while section 5 details the processes followed for its validation. Section 6 considers the implementation of OntoFoCE in ObE Forensics, and section 7 includes the conclusions from this work.

In order to organize the article adequately, the documents with additional information on this research have been stored in the repository ZENODO DOI 10.5281/zenodo.7977291, which has open access for those who wish to explore these components in more detail. There, the reader will find the OntoFoCE conceptual model, the description of the different classes, the SPARQL queries, the corpus used for the ontology's vocabulary validation, the study of the email items subject to expert examination, a detail of the updated bibliographic review, and a description of the three different scenarios used to validate OntoFoCE from a bank of test instances. The doctoral thesis resulting from the research is also stored in this repository.

2 Problems of the Forensic Analysis of Emails

This section includes the essential aspects of the problem under analysis.

2.1 Email Validation

An email is a block of data that travels through a network from a sender to a recipient device using the appropriate software and hardware for its transmission. The RFC 822 [Crocker, 1982] states that emails "are viewed as having an envelope and contents. The envelope contains whatever information is needed to accomplish transmission and delivery. The contents compose the object to be delivered to the recipient." *So, in this paper, an email is a digital document that consists of two parts: a) a header (the envelope) that contains information about the transmission process, identifying the accounts involved and the different devices where the email was stored during the transmission; and b) a body (the content) that contains the message that is transmitted, plus the attached files that optionally make up the message*" [Parra, 2019].

In this regard, the NIST Technical Note 1945 report [Nightingale, 2017] shows the authentication mechanisms (SMTP, DKIM and DMARC) that act on the communication process to ensure its security through email validation. Thus, from the point of view of forensic analysis, an email's validation consists in checking the sequence of events that make up the entire transmission process and are included in the email's header.

The header contains all the data that make it possible to identify the route traveled between the points of origin and destination. The forensic analysis of an email is carried out by taking the header as input, which records the transmission process that took place.

Considering the route an email takes from its sender to its recipient, different processes executed during the transmission occur. Among these, those that include the storage of the email in different memory spaces (servers and transmission/reception device) are of interest to forensic analysis.

The forensic analysis of emails requires that the message under analysis is authentic. *An email is authentic when the sender's data (email account and IP address), its traceability (different devices involved in the transmission with their respective IP addresses), and the recipient's data (email account and IP address) can be identified.*

In this article, determining whether an email is authentic is referred to as *email validation.*

2.2 Traceability of the Email Transmission Process

Section 3.6.13 in ISO 9000:2015 [ISO 9001:2015, 2015] defines traceability as the "*ability to trace the history, application or location of an object*." In the case of a product or service, the traceability can be related to the origin of the materials or parts, the history of the process and the distribution and location of the product or service after delivery. The main advantage of traceability (or reverse logistics) is knowing for sure the origin and history of a product. In this work, that "product" is the email, and its traceability is critical in forensic analysis to determine if the mail is authentic.

The procedural guidelines for the forensic examination of emails establish that the received email must be accessed to identify in the email header the IP address of the device that sent the email and the one that received it. This identification should be performed by reversing the transmission route from the recipient's device to the sender's. The *transmission process traceability concept* technically supports this inverse path. Scientific support is also possible if the traceability is represented using an ontology.

During the transmission process, the email is stored on each device or server through which it circulates. These copies should be the same in terms of the message in the body and the attached files, but differ in the header, which is updated with the data of the server where it is hosted. Every time the email reaches a device or server, the process adds the IP address data and the date and time it arrived at the beginning of the header. Thus, the header of the received email will contain the identification data of all the devices on which the email was stored during the transmission process. For this reason, the header of the received email is always considered to perform the forensic analysis and establish its validity as digital evidence. So, the email header is the input that OntoFoCE takes to represent the traceability of the transmission process.

2.3 Evidence Items Subject to Expert Examination

To [Cafferata Nores & García, 2003] expert evidence is an evidentiary means through which there is an attempt to obtain an opinion based on special scientific, technical, or artistic knowledge, useful for the discovery or evaluation of a piece of evidence. Its regulation is defined in a general way in the procedural codes of legal use. The objective of the expert evidence is defined as the "evidence items subject to expert examination" (EISTEEs). Although there is no legal-doctrine definition of what these items are, it is possible to explain them according to their characteristics:

- Through them, the judge or the requesting lawyer defines the scope of the expert's activity.
- It is requested to determine or clarify something to duly offer the evidence, which contributes to expanding the judge's knowledge and criteria when pronouncing judgement on the case. - They constitute the initial questions or unknowns the expert must answer when carrying out their activity. - These elements are usually expressed as verbs: "*verify... confirm... report... explain...*" The expert performs these actions using the scientific knowledge of their area to respond to the judge's request, focusing on each discipline's rules and good practices.

- They are instructions that are only the expert's competency and define the space within which the expert must perform their task.

The expert is free to select the tools, techniques or methods used to analyze the evidence, and in this regard, only the use of scientific methods that guarantee a formal procedure of the forensic analysis is required.

By using a tool that, based on the analyzed emails, automatically provides answers to the EISTEEs and generates the corresponding reports, the analysis methods can be objective and not rely on the expert's subjective criteria. According to [Robledo, 2015], the contribution of scientific expert evidence is crucial when it comes to scientific tests whose analysis methods are standardized and have a margin of error known and scientifically accepted, as they do not depend on the expert's subjective criteria or experience.

To achieve this automatic response to the EISTEEs, they must be written in formal language. Likewise, it is necessary to formalize the details in the header of the emails because it is there where the information to answer these questions will be obtained. An ontology is proposed to meet these requirements. Its development has been guided by the competency questions derived from a set of EISTEEs. It is widely accepted that the requirements an ontology must match are defined in terms of competency questions [Gruninger & Fox, 1995]. Through their formalization, it is possible to analyze the ontology and determine whether it answers them. By defining EISTEEs as competency questions, they can be formalized and used to analyze an ontology that represents the information in the email header, which allows the retrieval of the answers.

The set of competency questions that represent the most requested EISTEEs, according to a survey taken by several forensic experts from Argentina, are introduced in section 4.2. The ontology that answers these questions is described in section 4.3.

It is essential to clarify that the proposal does not imply that a new ontology must be made for each expert analysis on emails since the claims subject to examination are very similar in most cases. The competency questions are generated based on the set of EISTEEs that are most required, and, as a result, the ontology is created.

If a claim subject to examination is not answered in any of the initial competency questions, it must be added, and the expert must check that the proposed ontology answers it.

3 Related Works

The publication [Parra B. et al., 2019] describes the bibliographic review conducted at the beginning of this research, focusing on relevant works from 2014 to 2018.

Exploratory research was conducted to identify the related works on the application of ontologies to Digital Forensics, particularly on the expert analysis of emails. A critical study was made, and the scope was limited to the following objectives: (i) To identify and study the most up-to-date research contributions on Ontologies and Digital Forensics; (ii) To find areas without information on the application of ontologies to Digital Forensics; and (iii) To relate works based on attributes of proximity (or distance) with the application of ontologies for the forensic analysis of emails.

A literature review method was defined based on two phases: A) The definition of the study framework and the scope of the review, and B) the search and selection processes, which were previously adapted with a pilot test.

First, the two main topics (ontologies and digital forensics) were combined. Then, a detailed analysis focused on the objects of the expert report (the email and its header) was made to establish the search criteria with keywords, following the restrictions and exclusion criteria from the publications used.

The second phase included three stages: the initial search, the pre-selection of the keyword count, and the final selection. The first stage involved an ETL (Extraction, Transformation and Load) process to standardize the metadata of the publications, following the format provided by each digital library. It also involved a manual load when the search could not be exported. The "pre-selection of the keyword count" consisted in the use of the *KeyWordFinder*[1] tool, which allows the upload of files in a portable format (PDF) so that the tool counts the words entered, making the review of relevant texts much faster.

The sources of information used were IEEE Xplore Digital Library, ScienceDirect, Scopus, Scholar Google, The Journal of Digital Forensics, Security and Law (JDFSL), and ACM Library. Only open-access publications have been considered.

Concerning email-centered investigations, those related to emails that used various technologies (data mining, computer security, natural language processing, network traffic analysis methods, forensic methods, and tools) were discussed, primarily for dealing with spam-related issues. Notably, we found only one work that uses ontologies applied to emails [Mehta, 2017] . This work describes an ontology that represents the semantic addressing of an email, allowing users to address emails to semantically specific groups and providing secure authentication to groups of email accounts. The work of [Msongaleli, 2018] proposes an algorithm for the forensic analysis of emails based on three work layers: the email header, the logs of the email servers and the analysis of local devices. Although it does not resort to an ontological model, this is one of the few studies that contemplates both the internal structure of the email header and the external components (sender's and recipient's servers and devices) in line with the OntoFoCE proposal.

From this bibliographic review from 2014 to 2018, the conclusion is that no research was found that addresses in an integrated way the concepts of traceability, header analysis and ontologies applied to emails to answer the EISTEEs automatically. Using the same methodology and sources of information, we looked for relevant texts published between 2019 and the first quarter of 2023. From this second bibliographic review, some relevant publications were selected. They are shown in the document "BIBLIOGRAPHIC SURVEY 2019 - FIRST QUARTER 2023.pdf" in the ZENODO repository DOI 10.5281/zenodo.7977291. The relevant findings are emphasized next.

Further research is being conducted to apply ontological engineering in the scientific formalization of Digital Forensics. Other works that provide scientific support to forensic processes were found, such as the study of [Ellison, 2020], in which an ontology on reactive techniques in digital forensics is defined, contributing to the development of a valuable knowledge base for those working in digital forensics applied to medicine. The research of [Reedy, 2023] compiles approximately 260 publications from the period between 2019 and 2022 that show the progress of Digital Forensics, with an emphasis on the contribution of ontological engineering to

[1] A desktop version available for Linux can be downloaded from this link: https://mega.nz/file/hfpw2KTT#Ip3eSF4jAkfVxlBkIQVOlhZr-gWL2xk0g3rS1Ye_b-4

methodologies, analysis of file systems, reconstruction of events, the taxonomy of data of an Android device and the development of a database of forensic cases. Moreover, Ransomware is studied by [Keshavarzi & Ghaffary, 2023] and [Gopinath et al., 2022], who developed two ontologies called Rantology and CiberOntology, respectively, for the analysis of digital extortion through the representation of knowledge. Also, [Sikos, 2021] offers a review of ontologies that refer to unstructured forensic data and ontologies that could be applied in automating the processing of digital evidence. The work of [Peppes et al., 2020] fosters the use of emerging technologies such as semantic analysis, data mining and Big Data to develop the necessary software to predict and combat crime. Similarly, the work of [Srimukh & Shridevi, 2020] describes an extended ontology for the criminal investigation process, and [Arshad et al., 2020] presents a model of knowledge of events related to an incident under investigation from the forensic analysis of online social networks. The CFRaaS proposal [Kebande et al., 2020] is innovative, not only because it uses cloud services–in constant development currently–but also because it addresses issues related to safeguarding the forensic space and to problems with the admissibility of digital evidence.

There are plenty of studies on the spam problem [Dada et al., 2019; Karim et al., 2020; Krause et al., 2019; Méndez et al., 2019; Saidani et al., 2020], but the ones which stand out are those that resort to the analysis of email attributes (header and body), among which we can mention: the work of [Hina et al., 2021] that proposes a neural network to analyze both the header and the body of the email in search of evidence of cyberattacks. Also, [Fang et al., 2020] perform a forensic analysis with four attributes of an email ("From", "To", "Date" and "Body") using social media mining and semantic patterns in emails. Finally, the work of [Soni, 2020] separates the components of an email (header and body) to analyze them with tools based on neural networks.

Several works dealing with the application of ontologies to represent objects or actions linked to the forensic analysis of emails were found. [Tchakounté et al., 2020] suggest studying phishing attacks by building a knowledge base based on an ontological formalization with semantics. Although this paper proposes an ontology in the forensic field, it does not provide enough details for the forensic analysis of the header, nor does it allow the automatic answer to the EISTEEs. [Dimitriadis et al., 2022] combine ontological reasoning with other cybersecurity frameworks to study attacks that use emails as an attack vector. This proposal coincides in some aspects with ours in the analyzed attributes of emails; however, its model does not automatically answer the EISTEEs, nor does it represent the traceability of the transmission process. [Shukla et al., 2020] describe the detection of spoofing cases through the forensic analysis of the email header captured from the live running processes from memory. Also, there is the work of [Apoorva & Sangeetha, 2020], who use SVM algorithms for authorship attribution in the forensic analysis of emails. These works are only relevant because they consider the two datasets (header and body). However, they do not make an ontology to represent the traceability of the transmission process or to answer the EISTEEs.

It is observed in these works that, although they also address the analysis of the header, like OntoFoCE, none of them focuses on answering the EISTEEs. Thus, the conclusion is that none of the works proposes an automatic response to the EISTEEs from the ontology's competency questions in a single integrated context, nor do these works represent the traceability of email communication. These two are the main characteristics of OntoFoCE and make it different from the other ontological proposals dealing with emails. Various proposals have become increasingly strict concerning spam control on emails and the use of different technologies to conduct a forensic analysis of those emails that are an attack vector for illegal access. Nonetheless, the appropriate technologies do not support the validation of the sending account's legitimacy with automated identification tools. When a single email account has to be analyzed, the identification of the sending account is quick and easily visualized by simply reading the mail header. However, this manual task is not advisable when it comes to analyzing a high number of emails. For the latter case, the representation of the traceability of the transmission process described in OntoFoCE is especially useful.

4 Ontology for the Forensic Analysis of Emails

This section introduces the ontology for the forensic analysis of emails called OntoFoCE, which aims to represent the email and its transmission process to verify the email's authenticity as digital evidence and, consequently, the non-repudiation condition of the proof. The proposal uses ontological engineering to guarantee obtaining scientifically and technically accepted results regarded as valid according to the rules that justice requires for expert results.

Next, the most relevant components of OntoFoCE are explained, such as the methodological aspects of its creation, the basic requirements that were considered, and the conceptualization, implementation and evaluation of the developed ontology.

4.1 Methodological Aspects

The term "ontology" was taken from philosophy and has been widely used in the past years in knowledge engineering, artificial intelligence (AI), computer science and emerging fields, like the Semantic Web. Therefore, there are many definitions of "ontology". Several are presented and compared in [Gómez-Pérez et al., 2007]. For this paper, we adopted the definition by [Studer et al., 1998]: "*An ontology is a formal, explicit specification of a shared conceptualization*." Ontologies can be represented with different knowledge modeling techniques and implemented in various languages. However, not all of them can represent the same knowledge with the same degree of formality and granularity.

Several methods for developing ontologies have been reported in the literature in the last two decades. The first contributions in the field by [Gruber, 1993; Gruninger & Fox, 1995; Uschold et al., 1996; Uschold & Gruninger, 1996] gave rise to many subsequent proposals. Gruber's work discussed some basic criteria for ontology design related to quality and development methodology. Grüninger and Fox introduced a development methodology based on competency questions. The use of competency questions as specifications of the ontologies' requirements has been widely adopted by the different development methodologies devised so far. Some other methodologies proposed have been developed later; for instance, KACTUS [Schreiber et al., 1995], SENSUS [Swartout et al., 1997], On-To-Knowledge [Fensel et al., 2000], TERMINAE [Szulman & Biébow, 2002], Methontology [Corcho et al., 2005], NeOn [Suárez-Figueroa, 2010]. Although many ontology development methodologies have been proposed in recent years, none has emerged as a clear reference yet [De Nicola, A. et al., 2009].

In this proposal, the Methontology methodology has been adopted due to several reasons. One of them is that Methontology introduced an ontology lifecycle based on evolving prototypes and specific techniques to address each activity in the methodology. So, it allows an iterative development like the ones we need for OntoFoCE. This methodology is used to develop ontologies in most of the works reported in section 3. Another reason is that Methontology has been used to develop several ontologies in the legal field, as reported by [Corcho et al., 2005]. This application helped us understand how to interpret the judicial context to improve the representativeness of the concepts involved.

The iterative processes in the development of OntoFoCE included the main activities involved in the ontology's creation in each iteration, with different degrees of importance and progress in each activity, depending on the creation time.

The proposal of [Suárez-Figueroa, 2010], detailed in section 4.2, was used in the requirements specification phase. The OntoFoCE conceptualization was developed using intermediate representations, expressed with Unified Modelling Language (UML) class diagrams in which the classes represent the concepts and the relationships between them are represented with UML associations. These diagrams are detailed more precisely in section 4.3. In the formalization and implementation stages, the Protégé tool was used, in its 5.5.0 version, in which OWL language was used for the definition of the OntoFoCE classes and properties, Semantic Web Rule Language (SWRL) for the formulation of the inference rules and SPARQL for the definition of the queries that formalize the competency questions, as described in section 4.4. At last, the ontology was validated using an integrated methodology detailed in section 5.

4.2 Specification of Requirements

In this stage, the scope of OntoFoCE was defined, with details of the application domain and the requirements set for the ontology. The domain is defined by the context in which the email forensic analysis is performed. It is particularly interesting to represent three elements: the object of study (email), the traceability of the transmission process and the EISTEEs that act as the ultimate goal of making the expert's report.

The specification of the requirements of an ontology to establish the intended uses, the users and the requirements defined by them is based on *competency questions* that must be answered by the ontology [Gruninger & Fox, 1995]. The user's basic requirements are translated into unknowns that must be answered with the knowledge base considered in the ontology (classes and their relationships). The origin of the competency questions is the identification of the purpose, scope, level of formality, intended uses and end users of the ontology. In the case of OntoFoCE, the purpose and scope of the questions are defined to identify the validity of an email as a digital document and its non-repudiation as evidence. For this, an ontological model based on tools and structures that guarantees the acquisition of scientific and technically valid results is proposed. The application of OntoFoCE through its implementation in ObE Forensics aims at fulfilling the requirements of the computer expert working with emails as digital evidence, and who must answer precisely the judge's questions about the claims subject to examination. As they are compulsory, the claims subject to the expert examination define OntoFoCE's requirements and allow them to be adequately expressed as competency questions.

The competency questions represent the initial requirements of the model, and if the ontology built verifies them, it is assumed that the ontological model fully represents the domain. In the case of the forensic analysis of emails, the initial requirements are defined in the claims subject to examination, on which the expert must act in order to resolve them; and OntoFoCE is modelled based on them.

Given the importance of the EISTEEs, any ontology supporting the expert in the forensic analysis should be capable of answering them. So, the choice of the competency questions, which constitute the requirements of OntoFoCE, derived from a set of EISTEEs used in email forensic analysis. To this end, a survey of a limited group of nearby expert users (Argentine Computer Experts working in the criminal and labor fields mainly) was carried out to identify the most common EISTEEs related to emails, obtaining 86 different EISTEEs. With these results, a second analysis was conducted considering the unification of EISTEEs where the same requirement was expressed but with different words. Thus, 46 EISTEEs were obtained, which gave rise to the 21 competency questions that can be answered with OntoFoCE. The two sets (EISTEEs and competency questions) were also combined in order to verify that each identified EISTEE has one or more associated competency questions that answer it.

The following are the competency questions that have been identified with the mentioned process:

- CQ01: Given an email, what is the date and time when the email was sent and the sender's IP address?
- CQ02: Given an email, what is the date and time when the email was sent and the recipient's IP address?
- CQ03: Given an email, what accounts was the email sent to?
- CQ04: Given an email, what is the user alias and email address of the Sender?
- CQ05: Given an email, what is the user alias and email address of the Recipient?
- CQ06: Given an email, what was the email client used by each user?
- CQ07: Given an email, what device was the email sent from?
- CQ08: Given an email, what device was the email received on?
- CQ09: Given an email, a sender S and a recipient R, what is the sequence of devices this email traveled through?
- CQ10: Given an account A, what emails did it send?
- CQ11: Given an account A, what emails did it receive?
- CQ12: Given an account A1, has an email been sent to account A2?
- CQ13: Given an account A1, has an email been received from account A2?
- CQ14: Given an IP address, what is its geographic location?
- CQ15: What emails traveled through the device that has a given IP?
- CQ16: What emails were sent from a particular account on a given date?
- CQ17: What emails were received by a particular account on a given date?
- CQ18: Given a keyword, does it appear in the subject of an email?
- CQ19: Given a keyword, does it appear in the body of an email?
- CQ20: Given a keyword, does it appear in an email attachment?
- CQ21: What emails were exchanged between accounts A1 and A2 in a given date range?

In the document “SURVEY OF EVIDENCE ITEMS SUBJECT TO EXPERT EXAMINATION.pdf” (in the repository ZENODO DOI 10.5281/zenodo.7977291), you can find more details about the survey conducted to get the competency questions.

4.3 Conceptualization

This section describes the most important aspects of OntoFoCE from three partial perspectives: the representation of the email itself, the transmission process, and the occurrences. During the conceptualization stage, all the information collected during the forensic analysis of emails was studied and organized into intermediate representation structures, such as UML diagrams and tables, obtaining, as a result, the definition of the basic concepts of the ontology: classes, relationships, sample instances and properties.

It is important to note that bidirectional UML associations are implemented in OWL as two object properties (one property and its inverse) [Atkinson, 2008]. To better understand the queries in the case study presentation, it was decided to represent two unidirectional paths in the UML diagrams instead of a single bidirectional path between two classes. These two unidirectional paths correspond to the implementation of the object properties of the OWL.

In addition, restrictions and rules were defined using Descriptive Logic (DL) and the SWRL to limit the interpretation of the UML diagrams. In particular, the OWL Manchester syntax was used for expressions in DL.

The complete OntoFoCE conceptual model, a tabular representation of concepts, relationships, their descriptions, and rules, can be found in “OntoFoCE CLASS DESCRIPTION.pdf” in ZENODO (DOI 10.5281/zenodo.7977291).

4.3.1 Email Representation

As explained in Section 2, the forensic analysis of an email is carried out on the email header, which includes all the data referring to the email transmission process. To properly represent the email as an object of the expert’s report, OntoFoCE includes a set of associated concepts that conveniently shape the entire email concept. These concepts, illustrated in Figure 1, are explained in definitions 1 through 8.

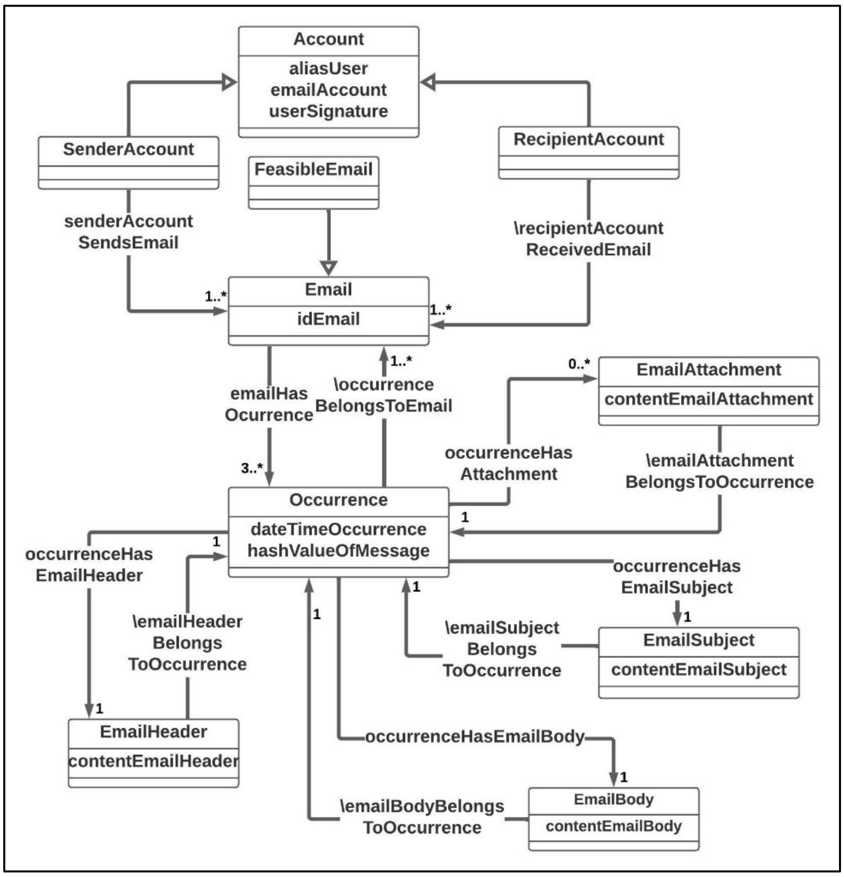


Figure 1: Email Concept Representation

Definition 1. An *EMAIL* is defined as "*a digital document that consists of two parts: a) a header that contains information about the transmission process that is taking place, with identification of the intervening accounts and the different devices on which the mail was stored during transmission; and b) a body that contains the message that is transmitted, plus the attached files that optionally make up the message*." This concept is represented in OntoFoCE with the *Email* class. In order to be represented conveniently in OntoFoCE, the email is separated into different concepts described later.

Each email is linked to two or more accounts and at least three occurrences. The concepts *OCCURRENCE* and *ACCOUNT* are described below.

Definition 2. An *OCCURRENCE* is defined as a "*Copy of the email that is stored on each device that participates in the transmission process*."

Definition 3. An *ACCOUNT* is an "*online service that provides a space for receiving, sending and storing email messages*." This concept is associated with a single user who acts as the sender/recipient of the email. The same account can be used to send and/or receive emails, so it is necessary to identify the account's role in each email. Thus, the subclasses *SenderAccount* and *RecipientAccount* are introduced in the ontology.

There is a unique axiom of the *SenderAccount* subclass, so for each email instance, there is a single account from which the email is sent.

In order to carry out the forensic analysis of an email, the email must meet certain requirements (it must have a header, and it must include the IP addresses of the email's sender and recipient). For this reason, the *FeasibleEmail* class is incorporated into OntoFoCE.

Definition 4. A *FEASIBLE_EMAIL* is defined as "*an email that meets the feasibility requirements for its forensic analysis*", because it is necessary for the email to meet certain requirements. For this reason, the *FeasibleEmail* class, which makes the *Email* class more specific, is incorporated in OntoFoCE. Instances of this class are inferred by the rule that establishes that "*An email is feasible to be analyzed when it has a Header, which contains the IP address of the Sender's Device and the IP address of the Recipient's Device.*".

Because each of the email parts has its own role in email forensics, it was necessary to partition the email representation using different concepts which represent its parts: *EMAIL_HEADER, EMAIL_SUBJECT, EMAIL_BODY* and *EMAIL_ATTACHMENT.* The proposed ontology incorporates them through classes with the same names.

Definition 5. An *EMAIL_HEADER* is defined as "*a flat text block that contains information relating to the email and the transmission process carried out*".

Definition 6. An *EMAIL_SUBJECT* is a concept that is defined as the "*Text that expresses the subject of the email*".

Definition 7. An *EMAIL_BODY* is defined as a "*message contained in the email*".

Definition 8. An *EMAIL_ATTACHMENT* is defined as a "*file associated with the email with complementary information to the content of the emai*l".

In email forensic analysis, keyword searches are often requested on the email's subject, body, and attachment, so it is important to identify these elements individually and represent them as classes.

Moreover, when an email circulates from one device to another during the transmission, its main components (header, subject, body, and attachment) are stored on the device or server through which it travels. These components are expected to be the same and not be modified during the entire transmission, except for the updated email header with the identification data of each device through which it circulates. For this reason, in OntoFoCE, the classes representing these components (*EmailHeader, EmailSubject, EmailBody,* and *EmailAttachment*) are linked to the *Occurrence* class.

4.3.2 Representation of the Transmission Process

It is important to consider that the email transmission process involves three moments (the sending stage, the internal transmission stage, and the receiving stage) which must be shaped *to establish the traceability* of the email sent. Figure 2 shows the partial view of two of them (the receiving and sending stages) in the email transmission process. The internal transmission process is explained in section 4.3.3 by describing the representation of the email occurrences.

The process of sending/receiving an email requires two primary components: the devices used for sending/receiving and the email managers, commonly known as Email Clients. The information that makes it possible to indicate and identify the acting users from the device used is relevant to forensic analysis. Figure 2 shows the concepts and the most representative relationships of the corresponding classes, which are detailed in items 8 to 10.

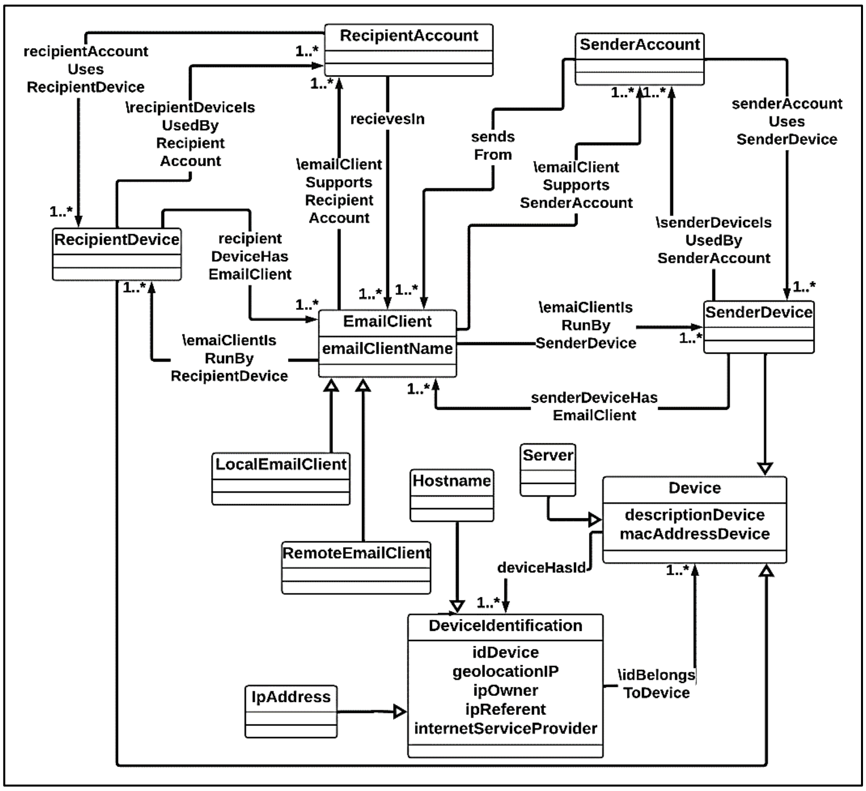


Figure 2: Representation of the Transmission Process

Definition 9. A *DEVICE* is defined as the "*Hardware component that stores an email.*" This concept encompasses any device (PC, mobile, notebook, email server, etc.) used during the email transmission process.

Three types of devices are identified according to the function they fulfill in the email transmission process: the sending device (used by the user to write and send the email); the servers (used by the email service during its transmission); and the receiving device (used by the user to receive and read the email). This division is expressed in OntoFoCE by the subclasses *SenderDevice*, Server and *RecipientDevice* which are more specific divisions of the Device class.

Definition 10. The *DEVICE_IDENTIFICATION* class is defined as a "*Unique identification of the hardware connected to the internet*" and comprises the set of data referring to the device's location in the context of the network used during the transmission of an email.

OntoFoCE makes the *DeviceIdentification* class more specific by dividing it into the *IpAddress* and *Hostname* subclasses, because if the email header contains a domain name instead of an IP address, it is likely that the domain has a dynamic IP address. It is important to highlight this distinction in the ontology since this data will be taken later as a reference for the geographical location of the users.

As the identification of a device can vary due to the random allocation of IP addresses that the service provider might make, the same device may have more than one identification, hence the need to represent them as a separate class.

Definition 11. An *EMAIL_CLIENT* is defined as a "*Computer application that manages an email account*." This concept refers to the software run by the user to access the email account. This class is divided into *LocalEmailClient* and *RemoteEmailClient* to show the possible places where a copy of the email under analysis can also be found (on the device itself if it is a local client or on the email server if it is a remote client).

4.3.3 Representation of Occurrences

As explained in section 2, during the transmission process, each server that participates contains a copy of the email, which is represented in the *Occurrence* class of OntoFoCE. Figure 3 illustrates the classes proposed by the ontology to represent the *OCCURRENCE* concept.

Three types of occurrences are identified according to the order or priority they have during transmission: The Sending Occurrence (stored on the sender's device), the Transmission Occurrences (successive copies of the email stored on intermediate servers) and the Receiving Occurrence (stored on the recipient's device). This division is expressed in the ontology with the subclasses *SendingOccurrence*, *TransmissionOccurrence* and *ReceivingOccurrence,* which are part of the *Occurrence* class.

Definition 12. A *THREAD* is defined as the "*Grouping of occurrences related to a recipient's account*." In OntoFoCE this concept allows associating all the occurrences that participate in the transmission process from the sender's account to each recipient's account.

Definition 13. The *SEQUENCE* concept is defined as a "*Series of threads of email occurrences associated with the same email*". This concept allows to identify the threads that belong to each email sent from the same sending account.

Definition 14. A *SENDING_OCCURRENCE* is "*The first Occurrence of the Thread*". All threads of the email will share this type of occurrence.

Definition 15. A *TRANSMISSION_OCCURRENCE* is defined as "*That Occurrence that is prior to another Transmission Occurrence or to a Receiving Occurrence, or that follows a Sending Occurrence*". They are found throughout the transmission process between the first and the last occurrence of the entire process.

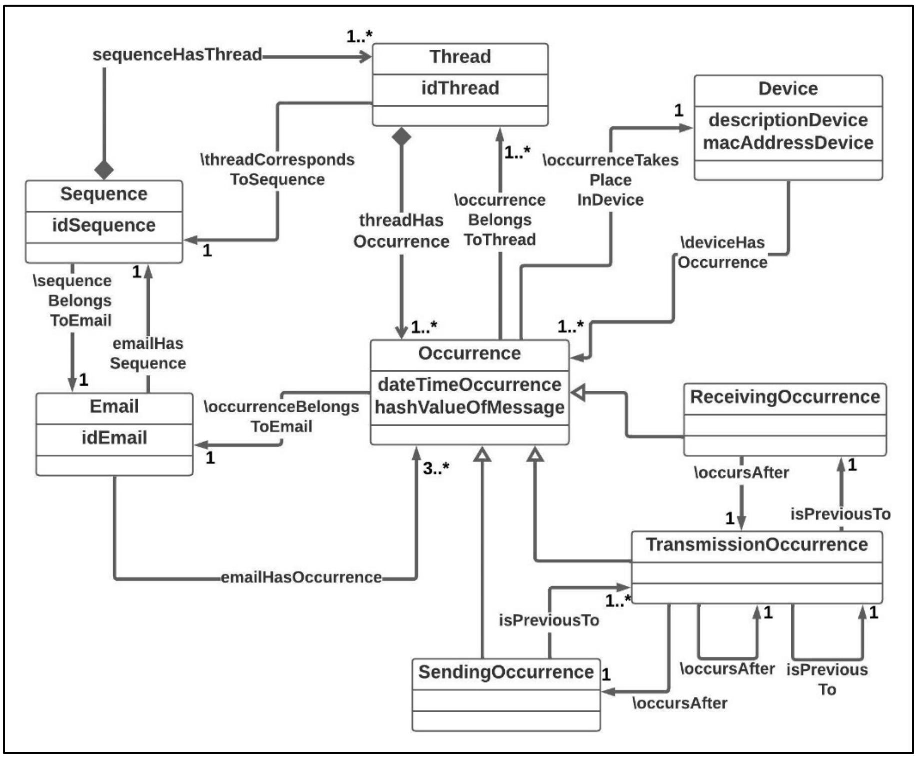


Figure 3: Representation of Occurrences

Definition 16. A *RECEIVING_OCCURRENCE* is defined as "*The last Occurrence of a Thread*".

But how is the order of each occurrence established in the corresponding thread? In each thread, there is a single SendingOccurrence and a single ReceivingOccurrence, which are the first and last, respectively, of that thread. The order of the TransmissionOccurrence in the thread is established through the relationship *isPreviousTo* that links an Occurrence with the one that follows it, and through the relationship *occursAfter*, which is the inverse relationship of the one previously mentioned. Both *isPreviousTo* and *occursAfter* associations favor the definition of the *SendingOccurrence*, *TransmissionOccurrence* and *ReceivingOccurrence* classes, which form the *Occurrence* class. Furthermore, in section 5.5 these concepts are shown with a concrete example of an email sent to three different recipient accounts.

Finally, the OntoFoCE conceptual model includes two complementary concepts: *KEYWORD* and *CASE_FILE*, which, although not directly linked to the email transmission process, they are considered necessary to answer the EISTEEs represented in the competency questions CQ18 to CQ20.

Definition 17: A *KEYWORD* is defined as a "*Word used to search for a topic of interest to the case*". It is linked to the *EmailSubject*, *EmailAttachment* and *EmailBody* classes since those are the parts where the search for the terms of interest for the case can be carried out.

Definition 18: A *CASE_FILE* is defined as a "*legal document containing the digital proof of an email and the items for expertise evidence opinion*." It is linked to the *Account* class since the request for a forensic analysis of the email attached as digital evidence starts there.

4.4 Formalization and Implementation

The semi-formal representations of OntoFoCE were implemented in *Protégé* [Musen, 2015], version 5.5.0, to build the OntoFoCE logic model. OWL was used to define the classes, properties, and attributes of classes and SWRL to write the rules and axioms. In addition, SPARQL query language was used to formalize the competency questions that allow, given an instantiation of the ontology, to obtain the answers to these questions. At https://obe.digilab.ucasal.edu.ar/owl_sin_instancias/ , you can access the OWL implementation of OntoFoCE.

5 OntoFoCE Validation

This section describes how the OntoFoCE validation was carried out, using an integrated methodology that allowed the identification and correction of mistakes in the conceptualization and design of the ontology. In particular, the following aspects were evaluated:

- Correct use of language to evaluate the coding of the ontology based on the rules and characteristics of the language used.
- Accuracy of the taxonomic structure: the taxonomy was analyzed, checking the concepts' and relationships' consistency, completeness and non-redundancy.
- Validity of the vocabulary: the meanings of the terms and concepts were analyzed by experts and with compilations of texts or any other source of knowledge available on the domain.
- Adaptation to requirements: in this phase, it was reviewed if the ontology met the pre-established requirements and if it responded to the competency questions.

The following sections describe each of these activities.

5.1 Validation of the Correct Use of Language

In this first activity, the ontology was verified regarding the representativeness of the ontological model developed, considering the language characteristics used. With the auxiliary tools of the Protégé environment (Hermit reasoner 1.3.8.413), the first consistency checks of the OWL model of OntoFoCE were carried out using an instantiation test bench.

Its development was also validated using semi-automatic tools such as OOPS! to identify errors and bad practices in the OntoFoCE code, which were duly adjusted.

5.2 Accuracy of the Taxonomic Structure

For this evaluation, we resorted to the collaboration of expert users (computer experts) who analyzed and discussed the representativeness of OntoFoCE, with Focus Group activities and in Workshops, using their knowledge of digital forensics. The adjustment proposals developed with these expert users' participation allowed the improvement of

the representativeness of OntoFoCE. These experts made suggestions on the representation of the following concepts and relationships.
These are the experts comments and the corresponding adjustments made to the model:

- Comment 1: "*The concept of USER should not be applied to represent the accounts used since the email identifies the ACCOUNT, not the user who uses that account. From the point of view of forensic analysis, it is not possible to associate an ACCOUNT with a USER by only considering the header of the email*". This recommendation was accepted, and the OntoFoCE model was revised by eliminating the USER concept included in the original ontology model.
- Comment 2: "*What happens when the mail header is not obtained completely? An email can only be validated if its header can be accessed and contains the IP address for sending and receiving the mail.*" Considering that it is not always possible to perform a forensic analysis of an email, as the digital evidence obtained contains incomplete or hidden data generated by malicious scripts that adulterate the content of the header, a new class labeled as *FeasibleEmail* was incorporated into OntoFoCe, making the *Mail* class more specific. This addition was to comply with the minimum requirements that a header must meet to make forensic analysis possible.
- Comment 3: "*The parts of the email identified as HEADER, SUBJECT, BODY AND ATTACHMENT travel with the email and are stored on the devices and pass-through servers; therefore, it is more appropriate to associate them with the concept of OCCURRENCE than with that of MAIL.*" This suggestion also modified the original OntoFoCE model, which associated the attributes with the mail, not with the occurrences.
- Comment 4: "*It may happen that if a domain name (instead of an IP address) is stored in the header during the transmission, it is not possible to be sure what the IP address of that domain is in the process.*" Based on this suggestion, the *IP* and *Hostname* subclasses were added, making the *DeviceIdentification* class more specific to represent both possibilities.

Although there is no formal record of the opinions of the experts who analyzed OntoFoCE, the comments were very useful for properly adjusting the model.

5.3 Vocabulary Validity

In the third validation, the terms included in the ontology were verified compared to a corpus of emails from an independent source of knowledge. The work produced by [Banday, 2011] was selected as the corpus. In this article, the architecture of an email is described, with the definition of the different components and the technical processes of the transmission process.

Based on the work previously mentioned, the "VALIDATION OF THE VOCABULARY OF OntoFoCE.pdf" was developed. It is a document in the ZENODO repository (DOI 10.5281/zenodo.7977291) that describes the different elements of an email, which can be integrated to form the corpus of the domain. This table also illustrates which of these elements or concepts are represented in OntoFoCE, showing the term's identification and name and the definition of the type of element according to the proposal of [Banday, 2011]. Then, a term description is presented to clarify details that the term name itself does not express. Finally, the concepts or attributes associated with each element of the table are indicated. Each row of the table represents a term according to the email architecture proposal of [Banday, 2011].

Thus, according to the function that each one fulfills, different types of elements have been defined:

- Email components: those referring to the email itself (24 elements),
- Transmission Process: those that define the processes involved in the transmission process (11 elements), and
- Protocols and Script: the different protocols and command chains that govern the sending process (11 elements).

From these groups, only 18 elements included in the first section are of interest for the forensic analysis of emails; the rest are not considered because they do not contain data that contribute to the validity or existence of the email.

Considering then the terms of OntoFoCE and those of the reference corpus, two metrics proposed by [Ramos et al., 2009] were analyzed: *Precision* and *Recall*, the former measures the degree of coincidence between the terms of the ontology and the corpus, and the latter indicates how many terms of the corpus are represented in the ontology.

The ***Precision*** metric is defined as the percentage of ontology terms listed in the corpus in relation to the total number of terms in the ontology, and is calculated using this formula:

$$Precision = CO\_C / Conto \qquad (1)$$

Being:

- CO_C = Number of terms repeated in the ontology and the corpus.
- COnto = Total number of terms in the ontology (including classes, subclasses, and attributes).

The corpus and ontology values can also be considered by defining a metric for the corpus. The *Recall* metric is defined as the percentage of terms of the corpus that appear in the ontology in relation to the total number of terms in the corpus. The calculation formula for this metric is this:

$$Recall = CO\_C / CCorpus \qquad (2)$$

Being:

- CO_C = Number of terms repeated in the ontology and the corpus.
- CCorpus = Total number of terms in the corpus.

In the case of OntoFoCE these values are:

- CO_C = 18 Mail Components terms + 4 Transmission Process terms = 22
- COnto = 14 classes + 13 subclasses + 23 class attributes = 50

Thus, applying these values in (1), the precision is expressed as:

Precision = 22/50 = 44%

Concerning the Recall metric, in the case of OntoFoCE these values are:

- CO-C = 22
- CCorpus = 46

Thus, applying these values in (2), the Recall metric is expressed as:

Recall= 22/46 = 48%

These values indicate the validity of the vocabulary. The 44% obtained for ***Precision*** indicates that almost half of the corpus defined for emails has been represented in the ontology, and this is logical since the remaining terms of OntoFoCE were defined only to establish the traceability of an email from the representation of its transmission process. That is, the ontology takes the main elements of an email–those that allow obtaining an accuracy of 44% in relation to the corpus–and was added the remaining terms required to validate, through the traceability of the transmission process, the existence of the email sent. It would have been better if the Precision metric value was higher, but that would only be possible if a more complete corpus were considered, including the terms related to digital forensics.

Moreover, the 48% obtained for ***Recall*** indicates that less than half of the terms are taken from the corpus. However, 22 elements from this corpus refer to the subprocesses, protocols and chains of commands established to perform the transmission process, which is irrelevant to the forensic analysis of email, For example: Mediator (A process that receives, adds, reformulates, and redistributes messages between authors and recipients) or SMTP (Communication protocol for the transfer of mail between the sending computer and the server of the issuing account). If we consider a corpus containing only the terms relevant for digital forensic analysis, the number of total elements in such corpus (CCorpus) will be 24. Thus, applying these values in (2), the *Recall* metric reaches 92%.

Recall= 22/24 = 92%

It can be concluded that the vocabulary defined in OntoFoCE is valid within the constraints found for the corpus defined for emails.

5.4 Adequacy to Requirements

Regarding this criterion, the validation consisted of verifying if the ontology requirements are fulfilled, considering its objectives, and the competency questions.

We worked on the adaptation to the requirements from the specification of examples according to four different scenarios: starting with Scenario 1, with the simplest example (an email with a sender and a recipient); then some more complexity was added in Scenario 2 (an email with one sender and several recipients) and multiplying the number of senders and recipients in Scenario 3.

Figure 4 illustrates these three scenarios.

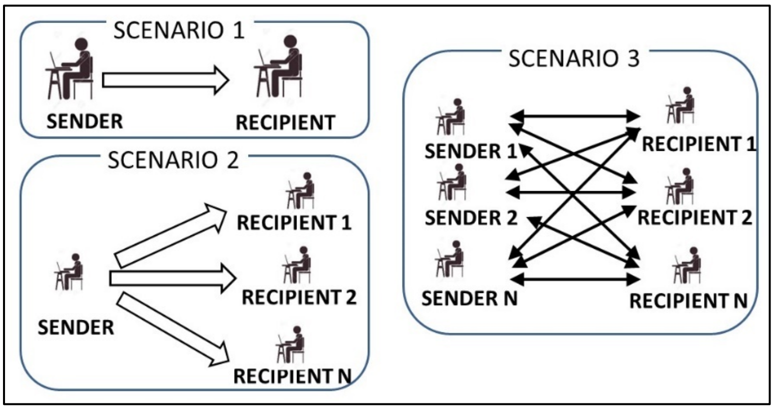


Figure 4: Scenarios for the expert analysis of emails

OWL implementations can be accessed at:
https://obe.digilab.ucasal.edu.ar/owl_escenario_1/,
https://obe.digilab.ucasal.edu.ar/owl_escenario_2/,
https://obe.digilab.ucasal.edu.ar/owl_escenario_3/.

The case in scenario 1 was used to validate the competency questions CQ01 to CQ09, while the competency questions CQ10, CQ14 and CQ15 were validated with scenario 2, and the remaining competency questions (CQ11, CQ12, CQ13, CQ16 to CQ21) were validated with scenario 3. ZENODO (DOI 10.5281/zenodo.7977291) includes the documents "VALIDATION SCENARIO 1.pdf", "VALIDATION SCENARIO 2.pdf" and "VALIDATION SCENARIO 3.pdf", which show the email headers taken as examples for each scenario, the occurrences identified, the competency questions that are answered, and some screenshots of the use of ObE Forensics in the corresponding scenario.

The following section briefly describes scenario 2, which takes as an example the EISTEEs that state the following: "*Determine the existence and veracity of the sending and receiving of the emails detailed*." It is observed that the answers to the competency questions allow the expert to respond with complete certainty to what was requested in the EISTEEs. The responses obtained from the instantiation of the headers clearly identify the data that allow confirmation of *"... the existence and veracity of the sending and receiving...*" as requested.

The answers provided by the competency questions are the same data that the expert would look for in the headers when manually carrying out the expert examination, since the existence of an email is verified when the sender's and the recipient's device and account are identified. These answers can be obtained from the competency questions defined in OntoFoCE.

Regarding the *veracity of the transmission of emails*, these are confirmed when the email's journey can be tracked from the moment it is delivered from the sender's account until it is received in the recipient's account. The answer to competency question CQ09 details the complete transmission process, expressly indicating the servers through with it traveled.

5.5 Example of the OntoFoCE Implementation in a Forensic Scenario

Figure 4 represents the transmission process of an email sent from the account bgallo@ucasal.edu.ar to three recipients' accounts (luzbibiana@gmail.com; enzo.notario@gmail.com and erivetti83@gmail.com). There will be three different threads as there are three recipients' accounts.

Figure 5 shows the corresponding threads (THREAD_1, THREAD _2 and THREAD _3), as well as the Sending Occurrence (SO), the Transmission Occurrences (indicated as TO_1 to TO_21) and the corresponding receiving occurrences (RO_1, RO_2 and

RO_3). The sending device (DEVICE_E), 13 servers (marked as SERVER_1 to SERVER_13) and the corresponding receiving devices (DEVICE_R1, DEVICE _R2 and DEVICE _R3) participate in the sending process.

Note that, in addition to sharing the Sending Occurrence, the threads share a set of common occurrences because all the email accounts have the same domain (Gmail). They are only distinguished in the last transmission occurrence and the corresponding receiving occurrence. Furthermore, it is also observed that there are servers that store more than one transmission occurrence and can be identified with an IP address or with a domain name.

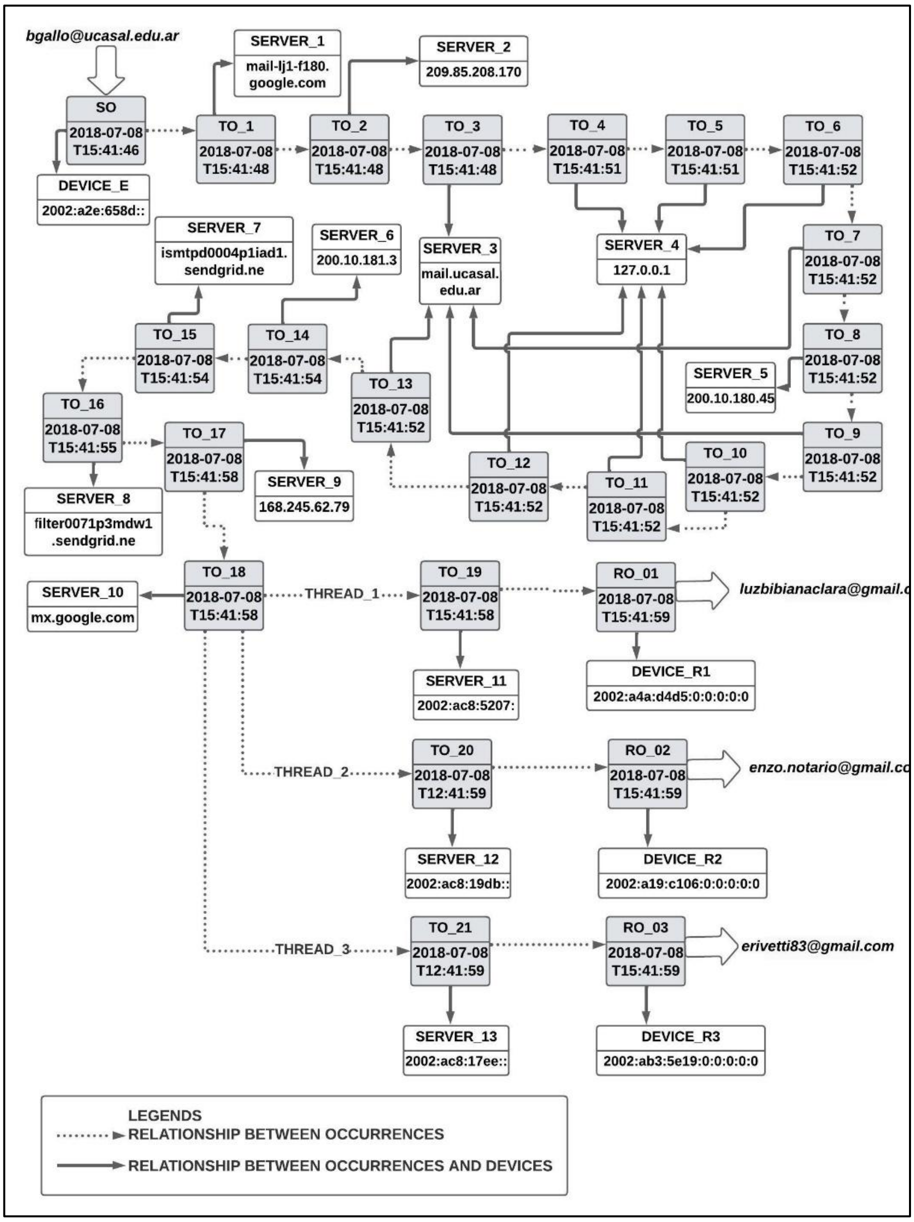
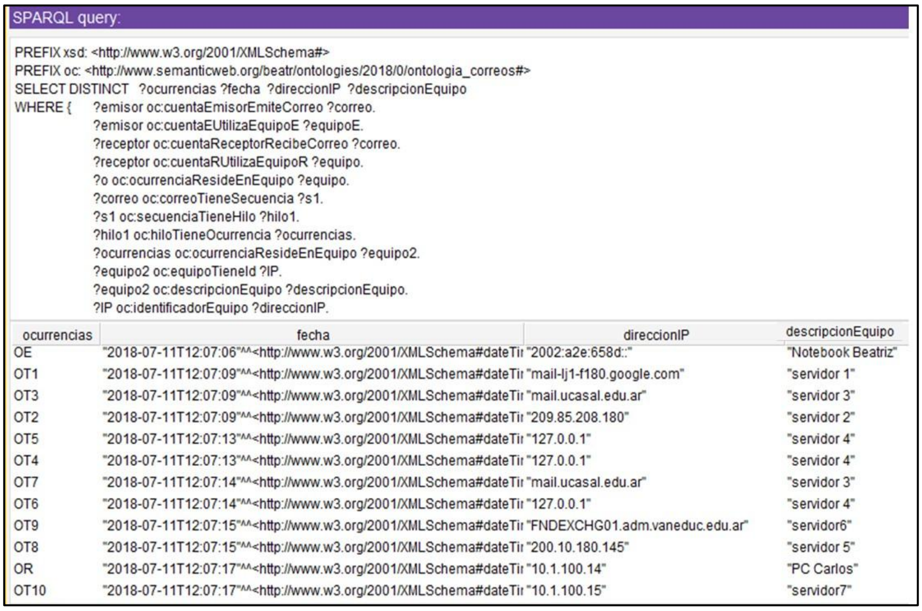


Figure 5: Outline of the Transmission Process of an email to three recipient accounts

A more detailed description of this scenario can be found in the document “VALIDATION OF SCENARIO 2.pdf” in the ZENODO repository (DOI

10.5281/zenodo.7977291). We extracted one competency question from this document to show how the ontology can answer it and, consequently, the EISTEEs.

Specifically, competency question *CQ09: Given an email, a sender S and a recipient R, what is the sequence of devices through which this email traveled?* is considered. Figure 6 shows the SPARQL query that allows answering this competency question, given the instantiation of the ontology with the explained forensic scenario and considering bgallo@ucasal.edu.ar as the sender's account. In the figure, it can be seen that there is an EMAIL_C1 whose subject, the date it was sent, and the three receiving accounts are shown.



```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX oc: <http://www.semanticweb.org/beatr/ontologies/2018/0/ontologia_correos#>
SELECT DISTINCT  ?ocurrencias ?fecha  ?direccionIP  ?descripcionEquipo
WHERE {     ?emisor oc:cuentaEmisorEmiteCorreo ?correo.
            ?emisor oc:cuentaEUtilizaEquipoE ?equipoE.
            ?receptor oc:cuentaReceptorRecibeCorreo ?correo.
            ?receptor oc:cuentaRUtilizaEquipoR ?equipo.
            ?o oc:ocurrenciaResideEnEquipo ?equipo.
            ?correo oc:correoTieneSecuencia ?s1.
            ?s1 oc:secuenciaTieneHilo ?hilo1.
            ?hilo1 oc:hiloTieneOcurrencia ?ocurrencias.
            ?ocurrencias oc:ocurrenciaResideEnEquipo ?equipo2.
            ?equipo2 oc:equipoTieneId ?IP.
            ?equipo2 oc:descripcionEquipo ?descripcionEquipo.
            ?IP oc:identificadorEquipo ?direccionIP.
```

ocurrencias	fecha	direccionIP	descripcionEquipo
OE	"2018-07-11T12:07:06"^^<http://www.w3.org/2001/XMLSchema#dateTi	"2002:a2e:658d::"	"Notebook Beatriz"
OT1	"2018-07-11T12:07:09"^^<http://www.w3.org/2001/XMLSchema#dateTi	"mail-lj1-f180.google.com"	"servidor 1"
OT3	"2018-07-11T12:07:09"^^<http://www.w3.org/2001/XMLSchema#dateTi	"mail.ucasal.edu.ar"	"servidor 3"
OT2	"2018-07-11T12:07:09"^^<http://www.w3.org/2001/XMLSchema#dateTi	"209.85.208.180"	"servidor 2"
OT5	"2018-07-11T12:07:13"^^<http://www.w3.org/2001/XMLSchema#dateTi	"127.0.0.1"	"servidor 4"
OT4	"2018-07-11T12:07:13"^^<http://www.w3.org/2001/XMLSchema#dateTi	"127.0.0.1"	"servidor 4"
OT7	"2018-07-11T12:07:14"^^<http://www.w3.org/2001/XMLSchema#dateTi	"mail.ucasal.edu.ar"	"servidor 3"
OT6	"2018-07-11T12:07:14"^^<http://www.w3.org/2001/XMLSchema#dateTi	"127.0.0.1"	"servidor 4"
OT9	"2018-07-11T12:07:15"^^<http://www.w3.org/2001/XMLSchema#dateTi	"FNDEXCHG01.adm.vaneduc.edu.ar"	"servidor6"
OT8	"2018-07-11T12:07:15"^^<http://www.w3.org/2001/XMLSchema#dateTi	"200.10.180.145"	"servidor 5"
OR	"2018-07-11T12:07:17"^^<http://www.w3.org/2001/XMLSchema#dateTi	"10.1.100.14"	"PC Carlos"
OT10	"2018-07-11T12:07:17"^^<http://www.w3.org/2001/XMLSchema#dateTi	"10.1.100.15"	"servidor7"

Figure 6: SPARQL Query Viewer for CQ09 COMPETENCY QUESTION

6 OntoFoCE Applications

For OntoFoCE to be used by an expert to perform the forensic analysis, a web tool called ObE Forensics (Ontology based Email Forensics) was built. To use this tool, the expert has to load the header of one or more emails, and a report with the answers to the competency questions will be generated. This report supports the expert's conclusions.

Each time a user ran ObE Forensics to perform a forensic analysis, it was necessary to implement an ETL method (Extract, Transform and Load) expressly defined for the preprocessing of email headers on OntoFoCE, based on the following criteria:

The morphological analysis of the email header is required, although not based on the meaning of the words, but rather on their position in the text concerning the header structure indicated by the RFC 822 standard.

It is important to provide the procedure with maximum automation because of the volume of data that can be entered during the forensic analysis of the headers of an email account.

ObE Forensics integrates four components in the same environment: a) the Ontology Instance Manager, responsible for creating the instances; b) the EISTEEs Analyzer, responsible for displaying the answers to the competency questions; c) the OntoFoCE ontology, which represents the knowledge applied to email forensic analysis; and d) the SPARQL Endpoint service that allows storing the instances and making queries about them.

The processing architecture used for the development, storage, and processing of data of ObE Forensics, as well as for the implementation of user interfaces and issues related to security, are shown in Figure 7:

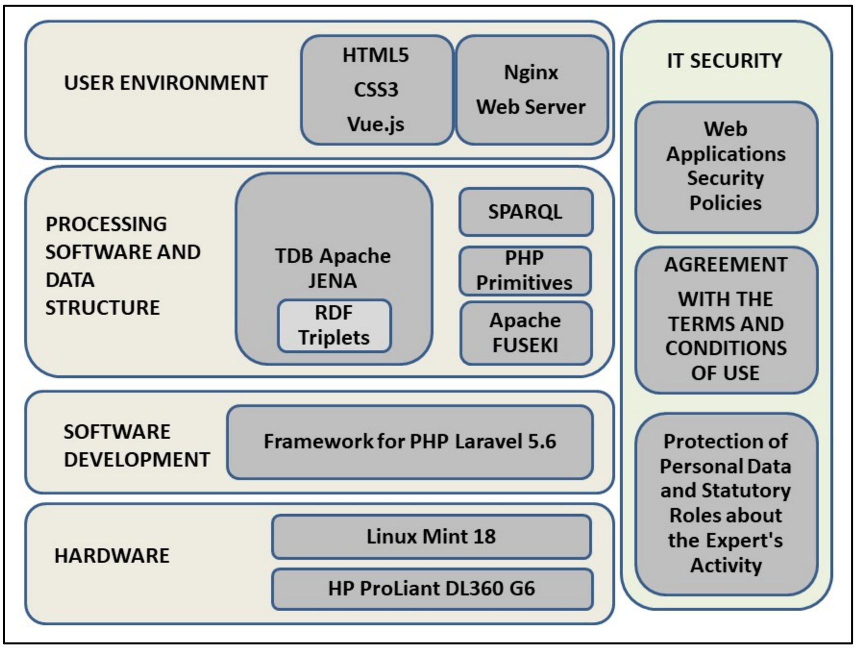

USER ENVIRONMENT
HTML5
CSS3
Vue.js
Nginx
Web Server
PROCESSING SOFTWARE AND DATA STRUCTURE
TDB Apache JENA
RDF Triplets
SPARQL
PHP Primitives
Apache FUSEKI
SOFTWARE DEVELOPMENT
Framework for PHP Laravel 5.6
HARDWARE
Linux Mint 18
HP ProLiant DL360 G6
IT SECURITY
Web Applications Security Policies
AGREEMENT WITH THE TERMS AND CONDITIONS OF USE
Protection of Personal Data and Statutory Roles about the Expert's Activity

Figure 7: ObE Forensics Processing Architecture

The RDF triple that makes OntoFoCE is stored in TDB Apache Jena, and Apache Fuseki is used as an API to communicate TDB Apache Jena with the application developed in Laravel.

The ObE Forensics application was also validated by expert users (IT experts) who were invited to test the application's performance in real cases of email experts' reports. This activity was extremely helpful in adjusting the web application according to the observations and suggestions provided by the experts.

Readers can enter the following URL https://obe.digilab.ucasal.edu.ar/ and request the corresponding access credentials to test the application. In addition, the documents stored in ZENODO with the description of the scenarios mentioned in section 5.4 show screenshots of this application's use in each scenario and the expert report that ObE Forensics has generated in each case.

It is important to note that, as it is based on an ontology, ObE Forensics complies with the requirements related to the implementation of scientific methodologies and principles, which is a mandatory requirement so that the forensic analysis process is not questioned. This condition is particularly observed in the answers to the EISTEEs provided by the tool, which are supported by the ontology competency questions.

7 CONCLUSIONS

During the development of OntoFoCE, it was necessary to direct the study and research toward the use of semantic technologies in digital forensics, addressing the issue from various perspectives.

The context of application and experimentation of the subject under study was circumscribed according to the criteria of scope, depth, opportunity, and access to real problems of digital forensics, identifying the basic problems when carrying out the forensic analysis of an email. Thus, the research was based on the following criteria: a) the concept of email traceability is taken as the common thread for forensic analysis to address the existence of the email; b) the EISTEEs, which act as a guide for the forensic analysis of the email, are represented in the ontology in terms of competency questions, so that, when answered, it is possible to respond to the EISTEEs requested by the judge; and c) the created ontology must allow the representation of the domain, whatever the structure of the email analyzed, even in those cases in which said structure does not conform to the standards of rigor.

The work carried out highlights the use of ontological engineering in the development of a supporting tool for forensic analysis, which allows scientific support for the non-repudiation condition of the digital evidence when it comes to an email, giving rise to the two most important contributions: the OntoFoCE ontology and the ObE Forensics application.

The OntoFoCE ontology allows the representation of the traceability of an email in order to verify its authenticity as digital evidence and, consequently, create the non-repudiation condition of the evidence.

The ObE Forensics application, which is based on OntoFoCE, constitutes a supporting tool for the IT expert's task, adding benefits to their work. Among these benefits are the efficiency in the forensic activity and greater precision in the analysis due to automatic evidence processing that helps avoid human errors resulting from visual analysis. The ObE Forensics application provides an answer to the most common EISTEEs for the forensic analysis of emails, and these answers are supported by the OntoFoCE competency questions, ensuring the application of "scientific principles" in the forensic analysis process.

It is possible to take advantage of the representation characteristics of OntoFoCE traceability with a specification of this ontology to model other situations that use transmission processes. For example, the *EmailAttachment* class can be made more specific to validate the process of sending an invoice or other legal or commercial document, in which traceability requirements must be met.

References

[Apoorva & Sangeetha, 2020] Apoorva, K. A., & Sangeetha, S. (2020). Forensic Analysis of E-mail for Authorship Attribution : Research Perspective. In *Proceeding of First Doctoral Symposium on Natural Computing Research* (pp. 281–292).

[Arshad et al., 2020] Arshad, H., Jantan, A., Hoon, G. K., & Abiodun, I. O. (2020). Formal knowledge model for online social network forensics. *Computers and Security*, *89*, 101675. https://doi.org/10.1016/j.cose.2019.101675

[Atkinson, 2008] Atkinson, C. (2008). *A Detailed Comparison of UML and OWL A Detailed Comparison of UML and OWL*. 1–58.

[Banday, 2011] Banday, M. T. (2011). Techniques and Tools for Forensic Investigation of E-Mail. *International Journal of Network Security & Its Applications (IJNSA)*, *3*(6), 227–241. https://doi.org/10.5121/ijnsa.2011.3617

[Cafferata Nores & García, 2003] Cafferata Nores, J. I., & García, G. (2003). *La prueba en el proceso penal* (LexisNexis, Ed.; 5a Edición). Depalma.

[Corcho et al., 2005] Corcho, O., Fernandez, M., Gomez, A., & Lopez-Cima, A. (2005). Construcción de ontologías legales con la metodología METHONTOLOGY y la herramienta WebODE. *Law and the Semantic Web: Legal Ontologies, Methodologies, Legal Information Retrieval, and Applications*, 142–157.

[Crocker, 1982] Crocker, D. H. (1982). *RFC822 Internet Message Format*. https://www.rfc-editor.org/rfc/rfc822

[Dada et al., 2019] Dada, E. G., Bassi, J. S., Chiroma, H., Abdulhamid, S. M., Adetunmbi, A. O., & Ajibuwa, O. E. (2019). Machine learning for email spam filtering: review, approaches and open research problems. *Heliyon*, *5*(6). https://doi.org/10.1016/j.heliyon.2019.e01802

[De Nicola, A. et al., 2009] De Nicola, A., Missikoff, M., & Navigli, R. (2009). A software engineering approach to ontology building. *Information Systems*, *34(2)*, 258–275.

[Dimitriadis et al., 2022] Dimitriadis, A., Lontzetidis, E., Kulvatunyou, B., Ivezic, N., Gritzalis, D., & Mavridis, I. (2022). Fronesis : Digital Forensics-Based Early Detection of Ongoing Cyber-Attacks. *IEEE Access*, *11*(December 2022), 728–743. https://doi.org/10.1109/ACCESS.2022.3233404

[Ellison, 2020] Ellison, D. (2020). Ontology for Reactive Techniques in Digital Forensics. *IEEE Conference on Application, Information and Network Security*, 83–88.

[Fang et al., 2020] Fang, Y., Zhao, C., Huang, C., & Liu, L. (2020). SankeyVis: Visualizing active relationship from emails based on multiple dimensions and topic classification methods. *Forensic Science International: Digital Investigation*, *35*, 300981. https://doi.org/10.1016/j.fsidi.2020.300981

[Fensel et al., 2000] Fensel, D., Harmelen, F. Van, Klein, M., Akkermans, H., Schnurr, H., Studer, R., Hughes, J., Krohn, U., & Davies, J. (2000). On-To-Knowledge : Ontology-based Tools for Knowledge Management. *Proceedings of the EBusiness and EWork*, 18–20.

[Gómez-Pérez et al., 2007] Gómez-Pérez, A., Fernández-López, M., & Corcho, O. (2007). Ontological Engineering: with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web. *Springer Science & Business Media*.

[Gopinath et al., 2022] Gopinath, M., Chakkaravarthy, S., & Ph, S. (2022). A comprehensive survey on deep learning based malware detection techniques. *Computer Science Review*, *47*, 100529. https://doi.org/10.1016/j.cosrev.2022.100529

[Gruber, 1993] Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, *5*(2), 199–220. https://doi.org/10.1006/knac.1993.1008

[Gruninger & Fox, 1995] Gruninger, M., & Fox, M. S. (1995). *Methodology for the Design and Evaluation of Ontologies 1 Introduction 2 Motivating Scenarios*. 1–10.

[Hina et al., 2021] Hina, M., Ali, M., Javed, A. R., & Member, G. S. (2021). SeFACED : Semantic-Based Forensic Analysis and Classification of E-Mail Data Using Deep Learning. *IEEE Access*, *9*, 98398–98411. https://doi.org/10.1109/ACCESS.2021.3095730

[ISO 9001:2015, 2015] *ISO 9001:2015*. (2015). https://www.iso.org/obp/ui#iso:std:iso:9000:ed-4:v1:en

[Karim et al., 2020] Karim, A., Azam, S., Shanmugam, B., & Kannoorpatti, K. (2020). Efficient Clustering of Emails into Spam and Ham: The Foundational Study of a Comprehensive Unsupervised Framework. *IEEE Access*, *8*, 154759–154788. https://doi.org/10.1109/ACCESS.2020.3017082

[Kebande et al., 2020] Kebande, V. R., Karie, N. M., Ikuesan, R. A., & Venter, H. S. (2020). Ontology-driven perspective of CFRaaS. *WIREs Forensic Science*, *2*(5), 1–18. https://doi.org/10.1002/wfs2.1372

[Keshavarzi & Ghaffary, 2023] Keshavarzi, M., & Ghaffary, H. R. (2023). An ontology-driven framework for knowledge representation of digital extortion attacks. *Computers in Human Behavior*, *139*(October 2022), 107520. https://doi.org/10.1016/j.chb.2022.107520

[Krause et al., 2019] Krause, T., Uetz, R., & Kretschmann, T. (2019). Recognizing Email Spam from Meta Data only. *2019 IEEE Conference on Communications and Network Security, CNS 2019*, 178–186. https://doi.org/10.1109/CNS.2019.8802827

[Mehta, 2017] Mehta, R. (2017). *SEMANTIC E-MAIL ADDRESSING USING.*

[Méndez et al., 2019] Méndez, J. R., Cotos-Yañez, T. R., & Ruano-Ordás, D. (2019). A new semantic-based feature selection method for spam filtering. *Applied Soft Computing Journal*, *76*, 89–104. https://doi.org/10.1016/j.asoc.2018.12.008

[Msongaleli, 2018] Msongaleli, D. L. (2018). Electronic Mail Forensic Algorithm for Crime Investigation and Dispute Settlement. *In Digital Forensic and Security (ISDFS), 2018 6th International Symposium*, 1–5.

[Musen, 2015] Musen, M. (2015). The Protege Project: A Look Back and a Look Forward. *AI Matters*, *1*(4), 4–12. https://doi.org/10.1126/science.1249098.Sleep

[Nightingale, 2017] Nightingale, S. (2017). *NIST Technical Note 1945 Email Authentication Mechanisms : DMARC , SPF and DKIM.*

[Palmer, 2001] Palmer, G. (2001). A road map for digital forensic research. *Proceedings of the Digital Forensic Research Conference, DFRWS 2001 USA*, iii–42.

[Parra, B et al., 2019] Parra, B., Vegetti, M., & Leone, H. (2019). Advances in the application of Ontologies in the area of Digital Forensic Electronic Mail. *IEEE Latin America Transactions*, *17*(10), 1694–1705. https://doi.org/10.1109/TLA.2019.8986448

[Parra, 2019] Parra, H. B. (2019). *Una Ontología del Correo Electrónico y su Trazabilidad como Soporte para la Forensia Digital* [Universidad Tecnológica Nacional]. https://digilab.ucasal.edu.ar/tesis

[Peppes et al., 2020] Peppes, N., Alexakis, T., Adamopoulou, E., Remoundou, K., & Demestichas, K. (2020). A semantic engine and an ontology visualization tool for advanced crime analysis. *Procedia Computer Science*, *176*, 1829–1838. https://doi.org/10.1016/j.procs.2020.09.222

[Ramos et al., 2009] Ramos, E., Nuñez, H., & Casañas, R. (2009). Esquema para evaluar ontologías únicas para un dominio de conocimiento. *Enlace*, *6*(1), 1–11.

[Reedy, 2023] Reedy, P. (2023). Interpol review of forensic firearm examination 2019 – 2022. *Forensic Science International: Sinergy*, *6*(December 2022), 100305. https://doi.org/10.1016/j.fsisyn.2022.100305

[Robledo, 2015] Robledo, M. M. (2015). La aportación de la prueba pericial científica en el proceso penal. *Gaceta International de Ciencias Forenses [Revista Electrónica]*, *15*(1), 5–12.

[Saidani et al., 2020] Saidani, N., Adi, K., & Allili, M. S. (2020). A semantic-based classification approach for an enhanced spam detection. *Computers and Security*, *94*, 101716. https://doi.org/10.1016/j.cose.2020.101716

[Schreiber et al., 1995] Schreiber, G., Amsterdam, V. U., Wielinga, B. J., & Jansweijer, W. N. H. (1995). The KACTUS View on the ' O ' Word. *IJCAI Workshop on Basic Ontological Issues in Knowledge Sharing*, 159–168.

[Shukla et al., 2020] Shukla, S., Misra, M., & Varshney, G. (2020). Identification of Spoofed Emails by applying Email Forensics and Memory Forensics. *ICCNS 2020*, 109–114. https://doi.org/10.1145/3442520.3442527

[Sikos, 2021] Sikos, L. F. (2021). AI in digital forensics: Ontology engineering for cybercrime investigations . *WIREs Forensic Science*, *3*(3), 1–11. https://doi.org/10.1002/wfs2.1394

[Soni, 2020] Soni, A. N. (2020). *Spam e-mail detection using advanced deep convolution neural network algorithms*. *May 2019*.

[Srimukh & Shridevi, 2020] Srimukh, P. V, & Shridevi, S. (2020). Development of Ontology on Crime Investigation process. *Journal of Physics: Conference Series*, *1716*, 012053. https://doi.org/10.1088/1742-6596/1716/1/012053

[Studer et al., 1998] Studer, R., Benjamins, V. R., & Fensel, D. (1998). KNOWLEDGE ENGINEERING: Principles and Methods. *Data & Knowledge Engineering*, *25*, 161–197. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.110.8406&rep=rep1&type=pdf

[Suárez-Figueroa, 2010].Suárez-Figueroa, M. del C. (2010). *NeOn Methodology for Building Ontology Networks : Specification , Scheduling and Reuse* (Issue June). Universidad Politécnica de Madrid (España).

[Swartout et al., 1997] Swartout, B., Knight, K., Russ, T., & Rey, M. (1997). t Toward Distributed Use of Large-Scale Ontologies. *AAAI Technical Report SS-97-06*, 138–148.

[Szulman & Biébow, 2002] Szulman, S., & Biébow, B. (2002). Structuration de terminologies à l'aide d'outils de TAL avec TERMINAE. *Revue Traitement Automatique Des Langues*, *43*, 103–128.

[Tchakounté et al., 2020] Tchakounté, F., Molengar, D., & Ngossaha, J. M. (2020). *A Description Logic Ontology for Email Phishing*. *9*(1), 44–63.

[Uschold et al., 1996] Uschold, M., & Gruninger, M. (1996). Ontologies : Principles , methods and applications. *Knowledge Engineering Review*, *11*(2).

[Uschold & Gruninger, 1996] Uschold, M., King, M., House, S. B. R., Moralee, S., & Zorgios, Y. (1996). The Enterprise Ontology. *The Knowledge Engineering Review*, *13*(August), 31–89. https://dl.acm.org/citation.cfm?id=976226