2023

Ontology-Based Case Study Management Towards Bridging Training and Actual Investigation Gaps in Digital Forensics

Hung Q. Ngo
Technological Unuiversity Dublin, Ireland, hung.ngo@tudublin.ie

Nhien-An Le-Khac
University College Dublin, Ireland

Funder: This research received no external funding

Forensic Science International: Digital Investigation 47 (2023) 301621

Contents lists available at ScienceDirect

Forensic Science International: Digital Investigation

journal homepage: www.elsevier.com/locate/fsidi

Ontology-based case study management towards bridging training and actual investigation gaps in digital forensics

Hung Q. Ngo [a,*], Nhien-An Le-Khac [b,*]

[a] Technological University Dublin, Dublin, Ireland
[b] University College Dublin, Dublin, Ireland

ARTICLE INFO

Keywords:
Open source intelligence
Digital forensic intelligence
OSINT investigation
Forensic ontology
Case study management

ABSTRACT

The training programs in digital forensics have contributed many case study models to guide digital forensic analyses. However, they only account for a small number of real cases and they are usually too abstract while actual cybercrime investigations are more diverse and complex. This gap leads to difficulties in giving immediate and straightforward actions for law enforcement during cybercrime investigations. In this paper, we propose an ontology-based knowledge map model, which is a foundation model for building a case study management system for Digital Forensic Intelligence (DFINT) and Open Source Intelligence (OSINT) in digital forensics. The main idea of this proposed model is to encode specific training cases of cybercrime into knowledge map representations, then the system uses the knowledge from the ontology to provide more information on the context and enrich them to match actual cybercrime scenes. Therefore, this approach can be used to bridge the gap between training case studies and the actual investigation environment. To illustrate our approach, we build a DFOSINT ontology for DFINT and OSINT domain; develop a prototype of the case study management system, and evaluate it in two aspects, ontology validation and case study validation with existing case studies of digital investigations.

1. Introduction

Digital Forensic Intelligence (DFINT) and Open Source Intelligence (OSINT) are inevitable trends of monitoring, early detection, and investigation in digital forensics. OSINT is a process to identify, harvest, process, analyze and report data obtained from open data sources such as the mass media, social networks, forums and blogs, websites, public government data, publications, or commercial data. In general, there are three primary goals of digital forensics: 1) collect electronically stored information in a sound and defensible manner, 2) analyze the results of the collections, and 3) present the findings in formal legal proceedings or less formally to inform a client. The three main goals are reflected in the three main steps of the digital forensics process, including data acquisition, data analysis, and presentation of the results.

In digital forensics, Hunton (2009) first introduced a four-stack cybercrime execution model for the investigation process. Then, this author proposed eight logical stages of a technical cybercrime investigation (including initiation, modeling, assessment, impact/risk, planning, tools, and acquisition) Hunton (2011). Recently, ISO (the International Organization for Standardization) and IEC (the International Electrotechnical Commission) proposed standard ISO/IEC 27043:2015[1] as a worldwide standardization for incident investigation principles and processes. These cybercrime execution models and ISO standards provide guidelines for common investigation processes across various investigation scenarios. They are a solid foundation for DFINT+OSINT investigations, which play an increasingly important role in digital forensics Quick and Choo (2018).

In general, there are several studies, which provide frameworks for DFINT and OSINT in digital forensics. Quick and Choo (2018) described relations of DFINT and OSINT, and intelligence analysis techniques with digital forensic data. They also proposed a framework for the DFINT and OSINT processes with 11 steps, including commerce (scope/tasking), preparing anticipated equipment required and expertise; identifying and collecting data; data reduction from devices and media; quick analysis and entity extraction; performing the OSINT process; creating entity chart; inference development based on information findings; producing written/verbal report, and finally finalizing the matter. For DFINT, Weiser et al. (2006) developed a national repository as a plat-

* Corresponding authors.
E-mail addresses: hung.ngo@tudublin.ie (H.Q. Ngo), an.lekhac@ucd.ie (N.A. Le-Khac).

[1] https://www.iso.org/obp/ui/#iso:std:iso-iec:27043:ed-1:v1:en.

https://doi.org/10.1016/j.fsidi.2023.301621
Received 26 January 2023; Received in revised form 2 July 2023; Accepted 26 August 2023
Available online 13 September 2023

form for law enforcement agencies and bureaucratically diverse groups to share information about cyber crimes and digital investigations. In their study, the DFINT model has three aspects, including foreign intelligence, criminal intelligence, and cyber intelligence. On the other hand, Tabatabaei and Wells (2016) addressed an OSINT process in the context of cybersecurity with four primary elements, including data acquisition, data enrichment, data analysis, and dissemination.

However, the rapid development of technology has led to a disconnect between practitioners and researchers in digital forensics. The lack of an established theoretical foundation leads to the gap between theory and practice Sremack (2007). For example, researchers can quickly grasp the various new technologies, which are used in image/video investigations through their research, while practitioners only master the trained technology and they normally find the difficulty in adapting or using of untrained technologies. Therefore, there are challenges for the practitioners when dealing with untrained cases in the real-world investigations. Mohammed et al. (2019) also reported that there are gaps between training cases and real-world investigations, and even the gap between legislation, investigation, and prosecution. It requires investigators to have good knowledge of adapting to different directions to collect and analyze information. Similarly, Hunton (2011) also stated the gap between technology examination and law enforcement investigation in cybercrime investigations. Previous studies in DFINT and OSINT have provided several necessary definitions and processes for digital forensic, however, it still needs high-quality knowledge bases to assist law enforcement during investigations. For example, when photos are found during the investigation, they can contain information about objects, taken date, taken device, geo-locations, etc, and there are also investigated tools. It also lacks of knowledge representation and management systems to handle investigation cases in this domain. Moreover, Pastor-Galindo et al. (2020) addressed several limitations of OSINT in digital forensics, including the complexity of data management, unstructured information, misinformation, data source reliability, and strong ethical/legal considerations. These gaps lead our study to build a case-study management system to support handling case studies of OSINT investigations and training law enforcement in digital forensics.

In general, OSINT has several different approaches, and each department has a different approach to real-world case investigation. For example, Wells and Gibson (2017) stated that different UK police forces apply different OSINT processes and best practices for the collection of data prior to achieving a directed surveillance authority (DSA) under the Regulation of Investigatory Powers Act (RIPA). In digital forensics, ontology approaches are widely used to solve digital forensics tasks. Firstly, several studies focus on building ontologies for digital forensics as a knowledge base. For example, Karie and Venter (2014) introduces an ontology for digital forensics, including computer forensics (server, laptop, and desktop forensics), software forensics (operating systems, application software, and database forensics), multimedia forensics, device forensics, and network forensics. Their study provided a hierarchy of digital forensics and lacked details of each component in the ontology, as well as the relationships between them. In another study, Grigaliunas et al. (2017) proposed an ontology-based transformation model for modeling in the digital forensics domain. Chabot et al. (2015) proposed an ontology-based approach for reconstructing and analyzing automatically the events related to a digital incident, while respecting legal requirements. Sharma et al. (2019) presented different video forgery detection techniques for forensics as an ontology; however, their study does not provide a clear ontology with a structure or a set of concepts, instances, and relationships in the domain. Recently, Grant et al. (2020) proposed an ontology for intelligence, surveillance, target acquisition, and reconnaissance (ISTAR) activities. This cyber ISTAR ontology has a general framework with common concepts and relationships, which are identified by OSINT activities; however, the number of these concepts and relationships can be much higher in real investigations. In addition, Bielska et al. (2020) published a handbook on existing resources and tools for OSINT. It is a taxonomy that includes a wide range of OSINT tasks and tools; however, each record only has a concept name and links of tools or resources. In this context, a case study management based on a pre-defined ontology is a suitable approach to assist in training and then to support investigations by enabling the liaison and comparison between training and actual cases.

In this paper, we propose an ontology-based case study management to represent and handle case studies of investigations in digital forensics. These case studies include concepts, evidence, and investigating tools, which are used and found during OSINT activities. This proposed model aims to bridge training and actual investigation gaps in digital forensics, which has been produced previously from training cases and investigation reports. The main contribution of this model is to support law enforcement in understanding, managing, and using training cases for decision making. The pre-defined ontology in this model also provides more information about finding evidence and more options to choose investigation tools for law enforcement during their work. This ability of the model helps to bridge the gap in the rapid development of technology. The proposed model is much more robust and inclusive to model any type of investigation cases. In addition, this study also built a prototype for a knowledge repository that can hold up to investigation cases extracted from training cases and investigation reports. Finally, we developed an innovative Case Study Browser to identify case studies by concepts and roles.

The next section gives an overview of the requirements for DFINT and OSINT processes. Section 2 describes the details of the proposed model, including the proposed architecture and implementation. Section 3 presents the ontology building as the knowledge resource of the architecture, while Section 4 deals with several validations of the proposed model. Finally, the paper gives a conclusion and several future works in Section 5.

2. Ontology-based case-study management

2.1. Investigation process in digital forensics

In this study, we propose a new framework to assit the cybercrime investigation. This framework is based on Hunton's 6-stage process Hunton (2011) and a predefined digital forensic ontology (Fig. 1). In which, this framework supports finding similar case studies from the repository and also finding information on each instance occurring during the investigation. It moreover supports searching for suitable toolkits and techniques for each step of the investigation process.

Roles of the ontology-based case study management module in Hunton's 6-stage investigation process can be listed as follows:

- Modeling: Finding similar cases from the repository to classify the type of case studies.
- Assessment: Cross-verifying by similar cases, such as comparing evidence and finding hidden/potential information.
- Impact/Risk: If the ontology includes warning messages for each type of evidence, it can support investigators to pay attention when examining the evidence. However, our ontology does not include this information to support.
- Planning: Organizing investigation steps based on similar cases and existing resources. For example, if investigators found suspect's photos, they can plan to use EXIF tools to extract metadata.
- Tools: Finding suitable investigation tools based on the list of investigation tools for each piece of evidence during investigations.
- Action: The ontology and system do not assist in the action during investigations, however, it supports having more action plans when investigations can know about the context and potential tools. For example, they can choose to use pre-paid 3G sim card or VPN service to anonymize.

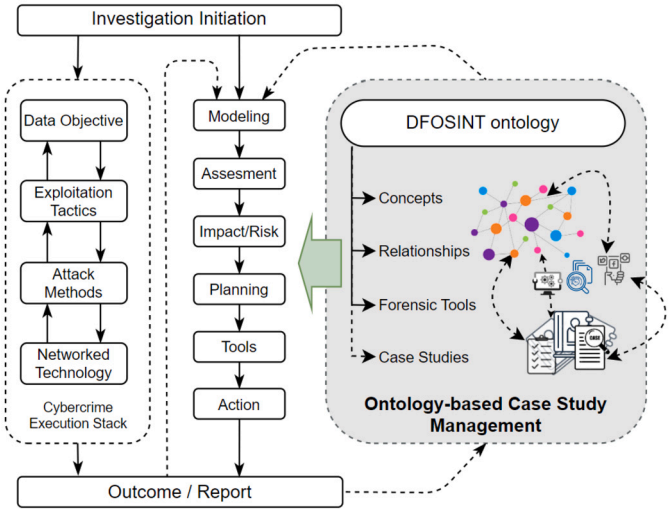

Fig. 1. Role of DFOSINT ontology and Case Study Management for 4-Stack Execution Model and 6-stage Process in Cybercrime Investigation.

In general, law enforcement is trained before being appointed to investigate. Education programs also ensure that they learn the full range of key concepts, skills, and tactics. However, real-world case investigations are diverse and technology is always changing and developing rapidly. The demand for up-to-date information to support investigations and case study management is great and very useful for investigators throughout their work. Therefore, a pre-defined ontology provides knowledge-base and information of toolkits for examining the evidence, while the ontology-based case study management system supports finding similar case studies during investigations. A rich pre-defined ontology provides a wide range of options in the investigation process. It fills the gap between training and actual investigation. Moreover, the system containing more case studies will provide more knowledge to the investigators.

2.2. System architecture

Based on the proposed assistant framework for cybercrime investigation, we propose an ontology-based knowledge map (OKM) model to represent case studies obtained from training programs, investigations, or experts in digital forensics. This model allows knowledge handling and exploitation in a flexible and scalable way.

Fig. 2 shows the architecture of our proposed ontology-based case study management system for digital forensics. Investigation cases are gathered from forensic reports, training case studies, or study models. These cases are transformed into independent RDF graphs and stored in the RDF storage module. RDF (Resource Description Framework) is a standard model for data interchange on the Web and linked data.

Like other knowledge management systems, the knowledge acquisition module decides which data types the system can accept and manage it. In digital forensics, these data can be forensics reports from real-world investigation cases, training cases from training programs, and study models from research projects. All of them are represented as knowledge maps in the system and each knowledge map represents a case study, moreover, the instances of each knowledge map are linked to pre-defined concepts in the ontology. In this study, a knowledge map contains the findings of the investigations/case studies, the concepts of their origin in the ontology, information about the investigation tools, and the relationships between these entities.

The storage module stores these knowledge maps in a graph database. These maps can be a big graph when they link together through concepts in the ontology or they also can be individual graphs if they just only link to the root concepts.

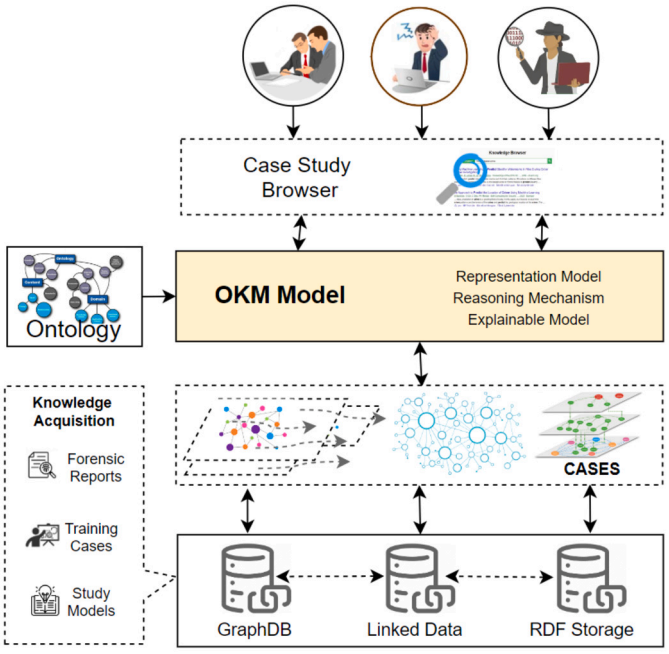

Fig. 2. An Architecture of Ontology-based Case Study Management for Digital Forensics.

The application browser layer can access case studies (RDF graphs) in the RDF storage module, and then represent them to end-users. In the OKM model, the explainable engine will retrieve domain knowledge from the pre-defined ontology. All instances and concepts from the repository and ontology (stored in the RDF storage module) have a Universal Resource Identifier (URI) and they link together based on their URIs.

In this study, the ontology-based knowledge map model is used to build an ontology-based case study management system to handle investigation cases in digital forensics and support law enforcement. In addition, a prototype of this proposed model has been applied successfully to handle mined knowledge from data mining Ngo et al. (2020, 2022).

2.3. Implementation

To implement and validate the proposed model experimentally in the context of digital forensics, the system is designed and developed in three major phases: 1) building a digital forensic ontology; 2) developing an ontology-based case study management system as a digital forensics repository; and 3) selecting and transforming case studies from existing resources into the repository.

The proposed model is based on a graph database and linked data technology. In this study, Apache Jena[2] is used for the graph database as native knowledge graph storage. This graph database server supports RDF triple storage and the SPARQL[3] protocol for query. In addition, it also includes Fuseki[4] for SPARQL Endpoint. All of these techniques are used to build a Digital Forensic Case Study Repository (DFKMaps).

One of the case study management applications is a search engine, which supports finding investigation cases from the DFKMaps repository. This web-based search engine is used to explore investigation cases (from the RDF storage) and provide explanations (with visualization techniques).

[2] https://jena.apache.org/index.html.
[3] https://www.w3.org/TR/sparql11-query/.
[4] https://jena.apache.org/documentation/fuseki2/.

Due to data of the proposed model being stored in the RDF storage module, all instances, concepts, and relationships have a URI as a unique identifier of objects. They also follow the FAIR (Findable, Accessible, Interoperable, Reusable) principles. Furthermore, the knowledge base of the predefined ontology and case studies of the input data are separated by two different prefixes (DFOSINT[5] and DFKMaps[6]). This way of data structure supports the reasoning mechanism that allows one-way inference from case studies to the knowledge base in the ontology.

3. Knowledge resources

3.1. Ontology design

In computer science, ontology is a knowledge model that represents a domain, the objects in that domain, and the relations between them. In other words, ontology is a formal representation of the domain knowledge by a set of concepts within a domain and the relationships between those concepts Gomez-Perez et al. (2006). Ontologies are built manually by using toolkits, such as Protégé,[7] CmapTools Ontology Editor[8] (COE), TopBraid Composer,[9] and many other toolkits.[10] Protégé is the most well-known toolkit for academics because it is a free, open-source visual ontology editor and knowledge-base framework and it has online and offline versions. As a result, the ontologies are stored in OWL, RDF, or XML format. This ensures that these ontologies containing domain knowledge can be accessed and further processed by computers.

The objective of ontology requirements is the identification of the scope of the ontology, the definition of possible scenarios, and the competence of the ontology Uschold and Gruninger (1996); Bravo et al. (2019). These requirements are used to support the construction and conceptualization of the ontology. They are as follows:

- **Scope of ontology**: Specify the motivation of the ontology. This ontology is used to represent case studies in DFINT and OSINT.
- **Scenarios of ontology**: Clarify the possible scenarios, users, and applications that will benefit. Due to the wide range of investigations in digital forensics, this version of ontology only focused on representing online investigation case studies from DFINT and OSINT activities in digital forensics.
- **Competency of ontology**: Specify the competency of the ontology for assisting the ontology development. To provide the competency of the ontology, a list of competency questions (CQs) is produced and the ontology must be capable to answer using its axioms Uschold and Gruninger (1996); Grüninger and Fox (2015).

The ontology supports represent four main dimensions related to forensic investigation, including criminal cases, location of evidence, potential context of finding pieces of evidence, and forensic resources. Based on the scope of the ontology and descriptions of the proposed model, this study proposes a list of six CQs, as well as assumptions for each CQ, which are used to support the construction and conceptualization of the ontology. These are as follows.

- **CQ1**. What types of knowledge are captured and represented in the ontology?
  Assumption: Common case studies of OSINT and DFINT, such as *website investigations*, *social networks investigations*.

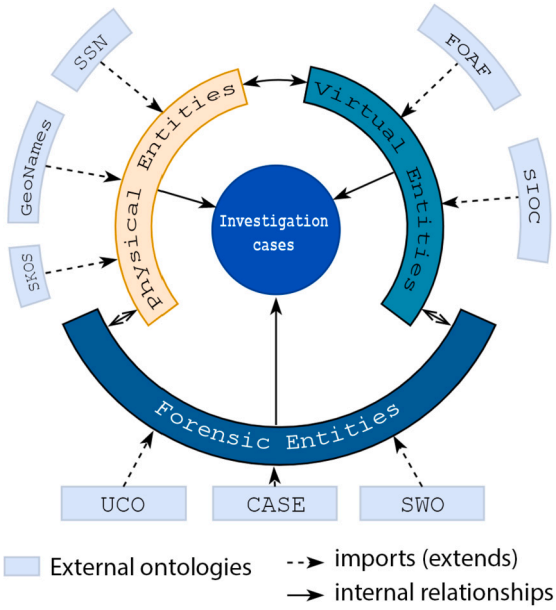

Fig. 3. The architecture of DFOSINT ontology.

- **CQ2**. What types of elements are represented in each case study?
  Assumption: Common real-world objects and related information can be found on the internet and digital tools to find these objects during investigations.
- **CQ3**. What concepts of a given domain are represented in case studies?
  Assumption: Digital Forensic concepts are nested in *VirtualEntity* and *PhysicalEntity* in the ontology, while the digital forensic tools are nested in *ForensicEntity* (as shown in Fig. 3).
- **CQ4**. What types of relationships are represented in the system?
  Assumption: Common relationships between represented digital forensic concepts, between forensic tools, and between forensic concepts and forensic tools.
- **CQ5**. How do case studies record, transform, transmit, and explain data?
  Assumption: Case studies are captured as a set of concepts, relations (the way to process data), and values. Then, they are recorded as RDF triples and can be transformed into other formats, such as OWL, or XML. Finally, they are stored in an OWL file or an RDF storage.
- **CQ6**. What types of case studies arise from the external aspects of the system and its improvement stages?
  Assumption: The type of case studies can be extended with other tasks in digital forensics, such as computer or mobile investigations.

The DFOSINT ontology has three main components, including *PhysicalEntity* for physical objects, *VirtualEntity* for virtual or online objects, and *ForensicEntity* for digital forensic entities (shown in Fig. 3). The design of DFOSINT ontology supports including existing ontologies related to social networks, geographical domains, or physical observations, such as FOAF[11] (Friend of A Friend), SOIC[12] (Semantically Interlinked Online Communities), SKOS[13] (Simple Knowledge Organization System RDF Schema), GeoNames,[14] SSN[15] (Semantic Sensor Network),

[5] Using *http://aseados.ucd.ie/DFOSINT#* as prefix for ontology.
[6] Using *http://aseados.ucd.ie/DFKMaps#* as prefix for knowledge maps.
[7] https://protege.stanford.edu/.
[8] https://cmap.ihmc.us/.
[9] https://www.topquadrant.com/products/topbraid-composer/.
[10] https://www.mkbergman.com/904/listing-of-185-ontology-building-tools/.
[11] http://xmlns.com/foaf/spec/.
[12] https://www.w3.org/Submission/sioc-related/.
[13] http://www.w3.org/2008/05/skos.
[14] https://www.geonames.org/ontology/documentation.html.
[15] https://www.w3.org/TR/vocab-ssn/.

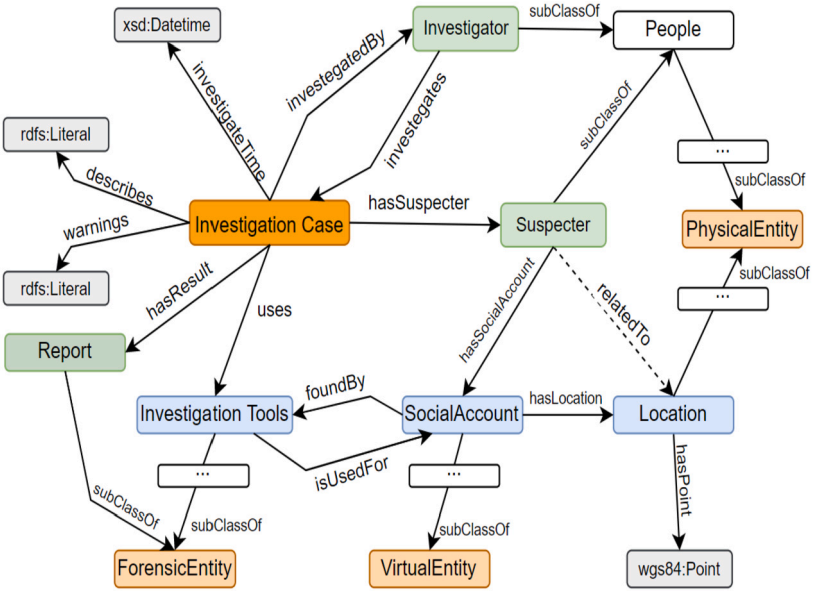

Fig. 4. The core structure of DFOSINT ontology.

UCO[16] (Unified Cyber Ontology), CASE[17] (Cyber-investigation Analysis Standard Expression), or SWO[18] (Software Ontology) ontology. This extension supports inheriting more pre-defined concepts and information to represent and explain forensic evidence.

Fig. 4 provides the core structure of the DFOSINT ontology and representations of case studies, which are represented and handled in the system. Each case study of investigations is an investigation case entity that links to suspects and other objects (including physical and virtual entities). It also links to investigation tools, which are used to investigate and find facts.

To support findability and accessibility of the model, the design of the ontology and stored case studies need to follow the FAIR principles. Moreover, every concept and relation in the ontology have been defined with at least two attributes, *rdfs:label* and *rdfs:comment*, for the title and description of each concept. *rdfs:isDefinedBy* and *rdfs:seeAlso* also provide external references for further information for each concept. These attributes are considered basic information for each entity in the ontology. They are used to provide definitions and detailed information of concepts and forensic tools, which are defined in the ontology.

- *URI* - Universal Resource Identifier.
- *rdfs:label* - the name of concepts or instances.
- *rdfs:comment* - description of concepts, instances, relations, or transformations for explanation purposes.
- *dc:identifier* - formula, expression, or function to calculate and transform data if it has.
- *rdfs:isDefinedBy* - sources or creators of the concepts.
- *rdfs:howToUse* - the way to use this tool.
- *rdfs:howToFind* - the way to find this piece of information.

This requirement is reflected in both concepts of the core ontology and entities in the system. Moreover, the relation *rdfs:howToUse* and *rdfs:howToFind* are used to assist investigators in finding suitable toolkits and potential pieces of information of evidence during practional investigations.

[16] https://unifiedcyberontology.org/.
[17] https://ontology.caseontology.org/.
[18] https://www.ebi.ac.uk/ols/ontologies/swo.

Table 1
Retrieval descriptions for the ontology's concepts.

	Count	Explanation
Wikipedia	1,711/7,366	Extract introduction and descriptions of concepts/tools.
OpenAI	5,824/7,366	Extract definitions of concepts/tools.
GoogleSearch	4,000/7,366	Extract screenshots or related images of tools.

3.2. Ontology vocabulary and enrichment

First of all, Bielska et al. (2020) provided a list of 7,366 resources in 370 categories in DFINT and OSINT. Each resource includes its name and its web address. This handbook[19] is a valuable resource to build the ontology with three steps of ontology enrichment (shown in Table 1).

In addition, there are many maintained collections of free actionable resources for those conducting OSINT investigations. They can be used to provide references or materials for building the ontology, such as OSINT_Collection,[20] AsINT_Collection,[21] Awesome OSINT,[22] Jungla OSINT,[23] CASE Ontology.

From the ontology hierarchy and vocabulary, concepts in the ontology are enriched semi-automatically with several steps as follows:

- Using Wikipedia to extract definitions for each definition in the ontology.
- Using GPT-3 model based on OpenAI[24] to the answer for query "What is definition of <*concept*>?".
- Using Google Search to find snapshots of instances if they are software.

After the enrichment process, all concepts and records with enriched information (shown in Table 1) were manually reviewed before being imported into the DFOSINT ontology.

[19] Open Source Intelligence tools and resources handbook.
[20] https://github.com/Ph055a/OSINT_Collection.
[21] https://start.me/p/b5Aow7/asint_collection.
[22] https://github.com/jivoi/awesome-osint.
[23] https://start.me/p/7k48PK/jungla-osint-por-ra1000.
[24] https://openai.com/.

4. Validation

To validate the proposed model, this study implements two validations, one for the ontology and the other for the case study representations. The ontology validation step ensures that the proposed ontology matches the requirements described in Section 3.1 and can be used in the proposed model. Moreover, the case study validation step helps to evaluate the proposed model to illuminate that it can be used to represent and handle OSINT and DFINT case studies in digital forensics.

4.1. Ontology validation

Many different criteria are used for the evaluation of ontologies, including precision, adaptability, clarity, completeness/incompleteness, consistency/inconsistentness, conciseness, computational efficiency, expandability, sensitiveness, redundancy, and transparency Vrandečić (2009); Staab and Studer (2010). Different approaches have different groups of these criteria, such as Gómez-Pérez used a group of five criteria (consistency, completeness, compactness, expandability, and sensitiveness) Gómez-Pérez (2004), while J. Bandeira, et al. Bandeira et al. (2016) proposed FOCA methodology with a group of 6 criteria (Completeness, Adaptability, Consistency, Conciseness, Computational Efficiency, and Clarity).

One of the first approaches to evaluating ontologies is using competency questions to measure competency criteria and to see if the designed ontology satisfies the requirements Grüninger and Fox (2015). The evaluation may be performed automatically, semi-automatically, or manually depending on the competency questions represented formally, specific heuristics, or human judgment, respectively.

To evaluate the DFOSINT ontology, this study divides the ontology evaluation process into two parts, validation and verification tests. Ontology validation examines the developed ontology to determine whether the correct ontology has been developed. In addition, ontology verification examines the developed ontology to determine whether the ontology has been developed correctly. Both evaluation tests are based on the scope of ontology, six competency questions (CQs) as described in Section 3.1 and the set of criteria to review the ontology.

In this study, two approaches are used to implement the test process. In the first test (ontology validation), the content of the ontology is evaluated based on five criteria, including consistency, completeness, clarity, expandability, and sensitiveness (proposed by A. Gómez-Pérez Gómez-Pérez (2004)). Each criterion is evaluated manually to answer *Yes* or *No*. The ontology will pass the content evaluation of this validation test if all criteria have been answered *Yes*. After checking the DFOSINT ontology with the five criteria and their guides, the proposed ontology has the answer *Yes* for all criteria, as shown in Table 2. In the first part of the evaluation of the ontology, the DFOSINT ontology has been shown to have the ability to be used in the model as a core background knowledge for the Forensic Case Study repository. It is also evaluated with five criteria to ensure that the development of the ontology is performed correctly. Each criteria in the list of five criteria also combines with the CQs when evaluating the ontology. For example, discovering EXIF data of images is one of common techniques in OSINT. This EXIF data might contain camera settings, date and time, and location information. Based on CQ2 (*What types of elements are represented in each case study?*) with its assumption (*Common real-world objects and related information can be found on the internet and digital tools to find these objects during investigations.*), the DFOSINT ontology needs EXIF data, image, date and time, location, and camera concepts to represent case studies related to EXIF data scanning. In addition, criteria *Completeness* (C2) requires that DFOSINT ontology reflects the scope of the ontology and assumptions of CQs for representing case studies in digital forensics. Therefore, the DFOSINT ontology passes the criteria when the ontology contains common concepts to represent them it.

In the second part of the test (ontology verification), the ontology is examined with a group of three criteria (including inconsistency, incompleteness, and redundancy) based on taxonomy evaluation metrics. The result in Table 3 has shown that there is no inconsistency, incompleteness, and redundancy error in the DFOSINT ontology. This result reflects the detail and carefulness of the ontology development process. Each concept in the digital forensics and computing domain has been reviewed and located in the ontology. Three criteria in the ontology verification can be verified by the design of the ontology and resources for building the ontology. For example, the definition and detailed information of concepts (including rdfs:label, rdfs:comment, dc:identifier, etc.) ensure that the concepts in the DFOSINT ontology are not inconsistency and incompleteness. Moreover, the main resource for building the DFOSINT ontology is a handbook on existing resources and tools for OSINT Bielska et al. (2020).

Overall, the DFOSINT ontology evaluation has passed two tests, ontology validation and ontology verification, with a set of 8 criteria (including consistency, completeness, clarity, expandability, sensitiveness, and without inconsistency, incompleteness, and redundancy). The ontology validation test ensures that the DFOSINT ontology is consistency, completeness, clarity, expandability, and sensitiveness to use in the proposed model to represent and handle case studies of OSINT investigations in digital forensics. On the hand, the ontology verification confirms that the proposed ontology does not include the concepts of inconsistency, incompleteness, and redundancy.

Table 2
Content Evaluation Metrics.

No.	Criteria	Explanation	Result
C1	**Consistency**	DFOSINT ontology contains all over 300 concepts and 5,300 investigation tools, which are consistent.	Yes
C2	**Completeness**	DFOSINT ontology reflects the scope of the ontology and assumptions of CQs for representing case studies in digital forensics.	Yes
C3	**Conciseness**	DFOSINT ontology is free of any needless concepts or redundancies between concepts.	Yes
C4	**Expandability**	DFOSINT ontology is a well-defined and scalable ontology.	Yes
C5	**Sensitiveness**	Small changes in DFOSINT ontology are not observant of the current concepts.	Yes

Table 3
Taxonomy Evaluation Metrics.

Criteria	Explanation	Check
C1. Inconsistency		
SC1. Circularity errors	All concepts are stated as specializations themselves.	No
SC2. Definition errors	No wrongly define concepts.	No
C2. Incompleteness		
SC3. Semantic errors	Incorrect semantic classification.	No
SC4. Incomplete concept classification	All main concepts are provided clear definitions and reviewed manually.	No
SC5. Partition errors	A partition between a set of concepts is omitted.	No
C3. Redundancy		
SC6. Grammatical redundancy	More than one explicit definition.	No
SC7. Identical formal definition	Concepts with the same formal definition.	No

4.2. Case study validation

To demonstrate the practical implementation and use of the case study mapping of the crime scene represented in the DFKMaps reposi-

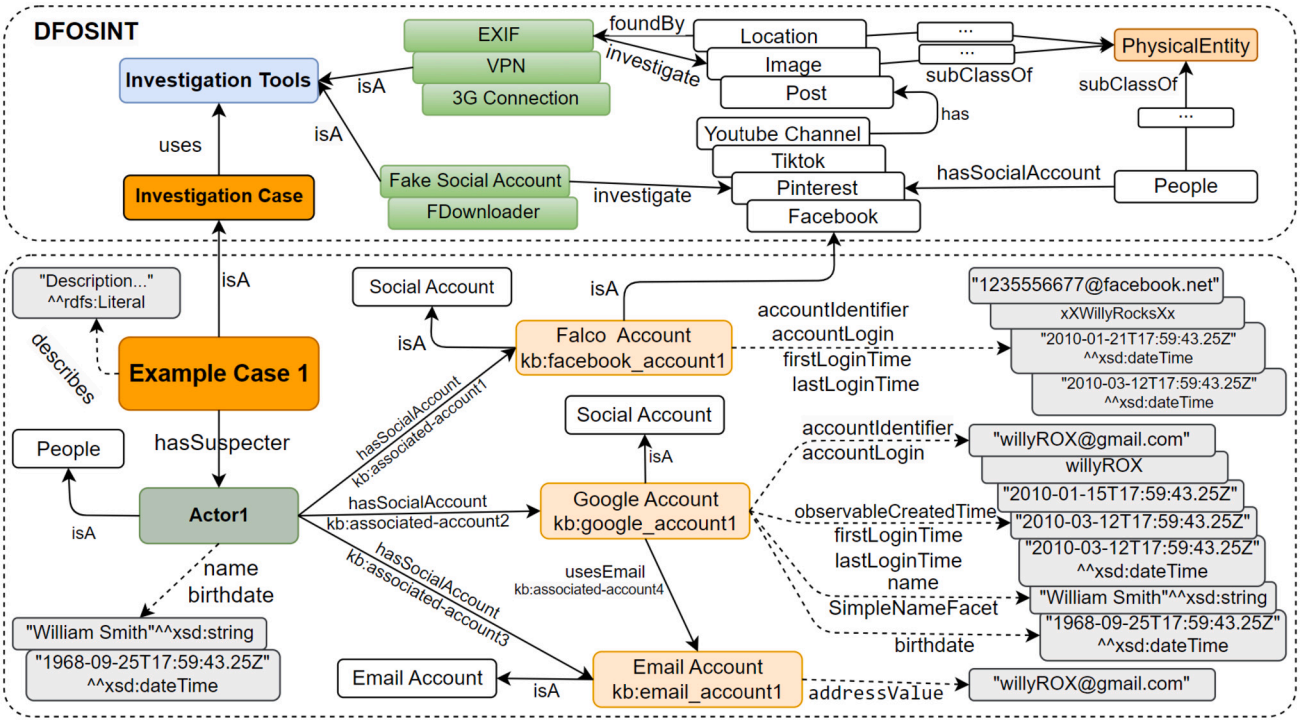

Fig. 5. An Example of Account Case from CASE Example repository, presented under DFOSINT.

tory, we evaluated it in two steps. First, we enter the case studies into the knowledge repository and trace them as knowledge maps. Then, these case studies can be retrieved in some forms.

In general, original case studies of forensic investigations can be presented in different ways, such as structured data (as CASE[25] examples from the DFRWS Workshop[26]) or investigation reports from the investigation process or training program. Due to the system being based on a graph database and linked data technologies to store and retrieve (as mentioned in Section 2.3), these case studies are collected, modeled, and transformed semi-automatically into the case study repository (DFKMaps repository). This repository supports to search existing/similar case studies or present case studies in suitable formats (graph- or text-based visualization). Moreover, different data sources will need different data-wrappers to extract instances and relationships from given case studies, then model and convert them into the RDF format before transforming them into the DFKMaps repository. These case studies are samples to validate the ability of our proposed model in representing, storing, and retrieving in a management system.

All case studies are transformed and stored in the DFKMaps repository, which is based on our proposed ontology-based case study management model and the DFOSINT ontology. Listing 1 presents the SPARQL query to list a collection of case studies in the DFKMaps repository.

Case Study 1:

Recently, the DFRWS Workshop proposed CASE (Cyber-investigation Analysis Standard Expression) as a community-developed standard to support reporting of digital traces, exchanging of digital traces, analysis of digital traces, and tool validation in digital forensic science and criminal justice. The first case study comes from CASE[27] Example repository. This case study provides a small investigation case of social accounts for *Actor1* "William Smith" (Fig. 5). Listing 2 presents a part of the raw triples of the case study after modeling. It includes basic entities and relationships extracted from the case study and stored as RDF triples. In the next step of the model, the application layer like web-based search is used to represent case studies in human-readable formats, such as web pages or graphs with interactions.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-
    syntax-ns#>
PREFIX DFOSINT:  <http://aseados.ucd.ie/DFOSINT#>
PREFIX DFKMaps: <http://aseados.ucd.ie/DFKMaps#>
SELECT  ?subject  ?sublabel
WHERE {
    ?subject  rdf:type   DFOSINT:InvestigationCase
        .
    ?subject  rdf:label  ?sublabel .
}
```

Listing 1: Collection of Case Studies.

In general, Fig. 5 provides entities and relationships occurring in the case study. However, the presentation of the given case study not only includes account names, emails, and several messages from the original document, but also includes assistant information from DFOSINT ontology (stored in the DFKMaps repository), such as potential accounts from other social networks, potential posts on social networks, and how to investigate them in the terms of OSINT investigation. When representing this case study based on our proposed model, these entities are linked to root concepts and investigation tools from DFOSINT ontology. Therefore, their representations in DFKMaps repository also provide necessary information of entities, context, and investigation tools. Without the DFKMaps repository and DFOSINT ontology, the representation has only entities and finding information, and it does not include the general context as well as investigation tools if they are not mentioned in the case study. This proposed approach assists to enrich necessary definitions of concepts and related investigation tools for investigators.

Case Study 2:

During the Master program in Forensic Computing and Cybercrime Investigation (MSc. FCCI[28]) of University College Dublin, students are

[25] https://github.com/casework/CASE-Examples.
[26] https://dfrws.org/presentation/case-workshop/.
[27] https://github.com/casework/CASE-Examples.
[28] https://www.ucd.ie/courses/msc-forensic-computing-cybercrime.

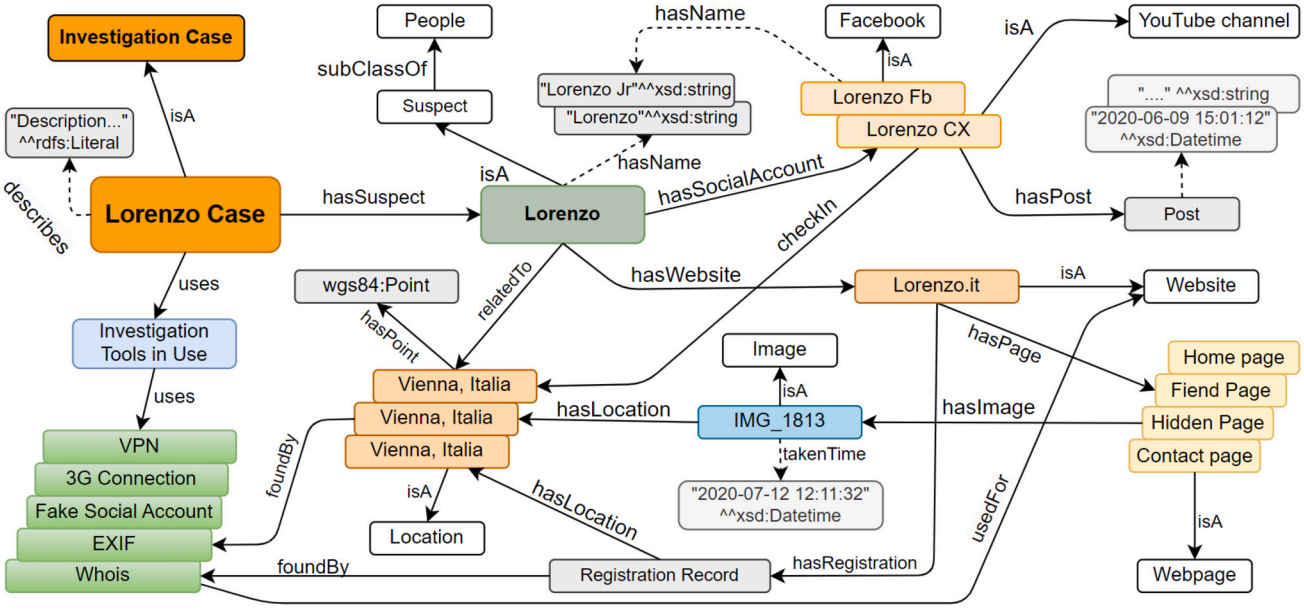

Fig. 6. An Example of Case Studies in OSINT, presented under DFOSINT.

```
DFKMaps:Case_Example_1 rdf:type owl:NamedIndividual ,
                          DFOSINT:Investigation_CASE;
        DFOSINT:hasSuspect DFKMaps:People_Actor1 ;
        rdfs:label "CASE Example 1" .
DFKMaps:People_Actor1 rdf:type owl:NamedIndividual ,
                          DFOSINT:Person ;
        DFOSINT:name "William Smith" .
        DFOSINT:birthdate "1968-09-25T17:59:43.25Z"
                          ^^xsd:dateTime ;
        DFOSINT:hasSocAccount DFKMaps:Email_account1 ,
                          DFKMaps:Facebook_account1 ,
                          DFKMaps:Google_account_1 ;
DFKMaps:Email_account1 rdf:type owl:NamedIndividual ,
                           DFOSINT:EmailAccount ;
                    DFOSINT:addressValue
                           "willyROX@gmail.com"
                           ^^xsd:string ;
```

Listing 2: Part of raw triples of Case_Example_1.

provided with many case studies related to online investigations. One of these case studies is Lorenzo case study, which starts from a warning message posted on Lorenzo's personal website. The message mentions a "big event" on a coming date and he recommends that his friends should not visit that place at that time. Students are required to investigate from open sources, such as web pages, social network accounts, etc, then make a report of the potential hazard and warnings. The report also includes information of the investigation and the toolkits used.

When representing Lorenzo case study in this model, extracted entities and relationships from investigation reports are transformed and stored in the DFKMaps repository. It includes information of suspects, their social accounts, posted messages, images, related locations, etc. These entities are linked to root concepts in the DFOSINT ontology and used toolkits during the investigation. Fig. 6 shows an overview of Lorenzo case study in the FCCI program. This case study is used in the training of law enforcement, however, similar cases can happen in their real-world investigations. This case study can be the same or similar information of each step and action. For example, the image "*IMG_1813*" can be found in the *hidden page* or on other pages of the suspect's website. Furthermore, the image can be found at any step of the investigation process. Then, it can be examined as an *Image* object using *Forensic Tools*, such as using the EXIF tools to explore the meta-data of the image. When the case study is linked to the DFOSINT ontology, investigators can find more EXIF information related to images (such as camera settings, date and time, locations, etc) and forensic tools. In addition, the knowledge from the ontology can show that there are many EXIF tools to use in investigations (Listing 3). Therefore, law enforcement can be trained to use *exiftool* to scan EXIF images, however, they can choose other options from a list of similar tools during practional investigations.

Discussion:

During the training period over the past few years, there were hundreds of investigation reports for this Lorenzo case study from FCCI's students. However, there is a huge difference between these reports on both sides, content and quality. Several reports have only about 10 pages with basic tools and simple finding clues, while several other reports have over 100 pages including information about the given case and using many toolkits during the investigation. In addition, several students have difficulty knowing what is complete or what information is missing. These difficulties and differences come from four main aspects:

- The diversity of the information sources that investigators can access on the internet;
- The variety of investigation tools and constant change of technology;
- The difference of OSINT models that investigators have used at their departments; and
- The knowledge and experience of investigators.

```
[EXIF]
    exiftool    https://exiftool.org/
    EXIF.tools  https://exif.tools/
    ExifyMe     http://www.xtrasimplicity.com/
    Exif Fixer  https://exiffixer.com/
    XnView      https://www.xnview.com/en/
    PhotoME     https://www.photome.de/
    ...
[Whois]
    Who.is      https://who.is/;
    Whois       https://www.whois.com/whois/
    ICANN Lookup    https://lookup.icann.org/en
    Whois Lookup    https://mxtoolbox.com/whois.aspx
    Whois Lookup    https://whois.domaintools.com/
    Domain Lookup   https://www.godaddy.com/en-ie/whois
    ...
```

Listing 3: Collection of Tools for EXIF/Whois.

The challenge is training tutorials can not provide information about the context for all pieces of clues and all toolkits, which can be used to investigate. It might lead to differences in practional investigations in the future. Therefore, this proposed model will be suitable not only to provide the framework to represent the case studies but also to provide the context and related information of concepts and finding entities for further usage in the future. Moreover, the information of over 7,000 resources in digital forensics from Bielska et al. (2020) and other online resources (they are imported into the DFOSINT ontology, and then stored in the DFKMaps repository) provide suitable tools for investigations.

In addition, Jeroen investigation is another case study used to train the MSc FCCI students, like Lorenzo case study. Although this case has different detailed information, they have a same format, which includes social accounts, personal web pages, sharing messages, posted images, and related locations. As a result, they can have a similar knowledge map when represented in the DFKMaps repository. It is especially helpful for investigators, who only need to be trained in fewer case studies, while they can handle many similar cases in practional investigations with supporting information.

In general, practical cases have shown that the ability of the proposed architecture can support the representation of case studies during the investigation and training processes. It also helps investigators to have a wider context and more information related to finding clues, then they can have more options for actions.

5. Conclusions and future work

In this study, we have implemented a fully operational prototype around the DFOSINT ontology to provide domain knowledge in digital forensics to store knowledge and support efficient knowledge exploration and retrieval. The current implementation is adaptive and easy to use. The study also built an open source and intelligent digital forensics ontology with thousands of concepts and instances, which will support the investigation process and training activities in digital forensics. In addition, the DFKMaps repository based on the proposed model can provide a flexible representation for investigation reports, including a brief or full detailed information representation. It also can support having more recommendations of potential toolkits related to finding clues during investigations. Finally, it can provide a same map for similar investigation cases.

The prototyped ontology-based case study management is very promising and has the potential to be extended to other application domains. Despite the flexibility, an efficient storage system, and knowledge retrieval, the proposed model is heavily based on the quality of inputs and repositories (previous ontology). We plan to design an evaluation methodology to evaluate not only the model performance, robustness, and scalability, but also the quality of input case studies of the knowledge repository.

Due to the vast scope of investigation cases in digital forensics, it is still challenging to include all possible cases with the latest digital devices, applications, investigation tools into the system, and the ontology must contain descriptions of all concepts during the investigation. However, the established framework should allow for the placement of these concepts in one of the aforementioned categories. Further research in small-scale digital forensics will examine the various forms of evidence and the procedures associated with each categorized investigation case, then they are considered to transform into the system for further references in training and even real investigations. In addition, we also plan to allow the investigators to enrich knowledge of the ontology with the information they found on their own on the digital crime scene in order to improve interactivity and collaboration between the DFKMaps repository and the investigators.

Declaration of competing interest

The authors declare that they have no conflict of interest.

Data availability

The data that has been used is confidential.

References

Bandeira, J., Bittencourt, I.I., Espinheira, P., Isotani, S., 2016. FOCA: a methodology for ontology evaluation. Appl. Ontol. 3.

Bielska, A., Kurz, N.R., Baumgartner, Y., Benetis, V., 2020. Open source intelligence tools and resources handbook. i-Intelligence.eu, 17–509.

Bravo, M., Hoyos Reyes, L.F., Reyes Ortiz, J.A., 2019. Methodology for ontology design and construction. Contad. Adm. 64.

Chabot, Y., Bertaux, A., Nicolle, C., Kechadi, T., 2015. An ontology-based approach for the reconstruction and analysis of digital incidents timelines. Digit. Investig. 15, 83–100.

Gómez-Pérez, A., 2004. Ontology evaluation. In: Handbook on Ontologies. Springer, pp. 251–273.

Gomez-Perez, A., Fernández-López, M., Corcho, O., 2006. Ontological Engineering: with Examples from the Areas of Knowledge Management, e-Commerce & the Semantic Web. Springer Science & Business Media.

Grant, T., van't Wout, C., van Niekerk, B., 2020. An ontology for cyber istar in offensive cyber operations. In: ECCWS 2020 20th European Conference on Cyber Warfare and Security, Academic Conferences and Publishing Limited, p. 117.

Grigaliunas, S., Toldinas, J., Venckauskas, A., 2017. An ontology-based transformation model for the digital forensics domain. Elektron. Elektrotech. 23, 78–82.

Grüninger, M., Fox, M.S., 2015. Methodology for the design and evaluation of ontologies. In: IJCAI95 Workshop on Basic Ontological Issues in Knowledge Sharing, pp. 88–98.

Hunton, P., 2009. The growing phenomenon of crime and the Internet: a cybercrime execution and analysis model. Comput. Law Secur. Rev. 25, 528–535.

Hunton, P., 2011. The stages of cybercrime investigations: bridging the gap between technology examination and law enforcement investigation. Comput. Law Secur. Rev. 27, 61–67.

Karie, N.M., Venter, H.S., 2014. Toward a general ontology for digital forensic disciplines. J. Forensic Sci. 59, 1231–1241.

Mohammed, K.H., Mohammed, Y.D., Solanke, A.A., 2019. Cybercrime and digital forensics: bridging the gap in legislation, investigation and prosecution of cybercrime in Nigeria. Int. J. Cybersecurity Intell. Cybercrime 2, 56–63.

Ngo, Q.H., Kechadi, T., Le-Khac, N.A., 2020. OAK: ontology-based knowledge map model for digital agriculture. In: Future Data and Security Engineering: 7th International Conference (FDSE). Springer, pp. 245–259.

Ngo, Q.H., Kechadi, T., Le-Khac, N.A., 2022. Knowledge representation in digital agriculture: a step towards standardised model. Comput. Electron. Agric. 199, 107127.

Pastor-Galindo, J., Nespoli, P., Mármol, F.G., Pérez, G.M., 2020. The not yet exploited goldmine of OSINT: opportunities, open challenges and future trends. IEEE Access 8, 10282–10304.

Quick, D., Choo, K.K.R., 2018. Digital forensic intelligence: data subsets and open source intelligence (DFINT+ OSINT): a timely and cohesive mix. Future Gener. Comput. Syst. 78, 558–567.

Sharma, H., Kanwal, N., Batth, R.S., 2019. An ontology of digital video forensics: classification, research gaps & datasets. In: 2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE), IEEE, pp. 485–491.

Sremack, J.C., 2007. The gap between theory and practice in digital forensics. In: Annual ADFSL Conference on Digital Forensics, Security and Law.

Staab, S., Studer, R., 2010. Handbook on Ontologies. Springer Science & Business Media.

Tabatabaei, F., Wells, D., 2016. Osint in the context of cyber-security. Open Source Intell. Investig., 213–231.

Uschold, M., Gruninger, M., 1996. Ontologies: principles, methods and applications. Knowl. Eng. Rev. 11, 93–136.

Vrandečić, D., 2009. Ontology evaluation. In: Handbook on Ontologies. Springer, pp. 293–313.

Weiser, M., Biros, D.P., Mosier, G., 2006. Development of a national repository of digital forensic intelligence. J. Digit. Forensics, Secur. Law 1, 1.

Wells, D., Gibson, H., 2017. OSINT from a UK perspective: considerations from the law enforcement and military domains. In: Proceedings Estonian Academy of Security Sciences, 16: From Research to Security Union, vol. 16, pp. 84–113.