Sikos Leslie (Orcid ID: 0000-0003-3368-2215)

WIREs Wiley Interdisciplinary Reviews

Article Title: AI in Digital Forensics: Ontology Engineering for Cybercrime Investigations

Article Type:

- ○ OPINION
- ○ PRIMER
- ◉ OVERVIEW
- ○ ADVANCED REVIEW
- ○ FOCUS ARTICLE
- ○ SOFTWARE FOCUS

Authors:

First author
Leslie F. Sikos, 0000-0003-3368-2215, Edith Cowan University, l.sikos@ecu.edu.au,
no conflict of interest

Abstract

In parallel with the exponentially growing number of computing devices and IoT networks, the data storage and processing requirements of digital forensics are also increasing. Therefore, automation is highly desired in this field, yet not readily available, and many challenges remain, ranging from unstructured forensic data derived from diverse sources to a lack of semantics defined for digital forensic investigation concepts. By formally describing digital forensic concepts and properties, purpose-designed ontologies enable integrity checking via automated reasoning and facilitate anomaly detection for the chain of custody in digital forensic investigations. This paper provides a review of these ontologies, and investigates their applicability in the automation of processing traces of digital evidence.

Graphical/Visual Abstract and Caption

Attached.

1. INTRODUCTION

A *digital forensic investigation* is an investigative process of a crime related to, or enabled by, computing devices, with the ultimate goal of finding digital evidence admissible in court. A typical digital forensic investigation process consists of four main phases: device seizure, data acquisition, data analysis, and reporting. What make digital forensics particularly challenging are the network resources, which, when involved, can make the complexity of any case explode—think of the variety of data formats during transmission, the encryption schemes used, online account permissions, data ownership, cloud storage location, etc. In today's interconnected world in which ubiquitious computing is commonplace, digital crime may be committed using a variety of, and multiple, devices, whether personal or corporate, including devices belonging to various organizations having geographically remote IT infrastructures, all with potentially different operating systems and file systems. While corporate workstations cannot be taken out from their assigned office space, and might be subject to CCTV surveillance, smartphones with 4G/5G mobile Internet connection can be used anywhere within the coverage area of their service provider, although the location of the device at a particular point in time does not necessarily mean that the user or owner of the device was definitely there at the same time (unless being on a—trackable—call at the time; or a geotagged photo of the user was taken, which is still available on, can be restored from, or can be carved from, the microSD card of the device, or was transmitted via the mobile Internet connection at the time and available elsewhere). Online profiles used for committing a crime may or may not be directly linked to a particular individual, potentially resulting in contradictory statements and uncertain knowledge regarding a case, misleading evidence traces, or missing evidence.

There is a continuing research effort to find ways to at least partially automate digital forensic investigations, with most approaches in the literature focusing on a particular phase of the process. For example, modeling the behavior of suspect executables in the form of finite state automata, in which each state represents behavior resulting in an observable modification to the victim system, was introduced for automated malicious event reconstruction (Shosha, James, Liu, & Gladyshev, 2012). A Perl tool called *Fast Modular Profiling Utility (FMPU)* was designed for the *Semi-Automated Crime-Specific Digital Triage Process Model*, which can profile computers by volume and user directory, and uses and generates HTML for digital forensic triage (Cantrell & Dampier, 2013). The automation of the detection of information linkage between distinct digital devices has been proposed by Brennan, Udris, and Gladyshev (2013). A framework to automate the post-incident analysis of mobile devices has been developed, which aggregates data by using semistructured files (XML) to store heterogeneous mobile data (Andriotis et al., 2014). Despite these and similar works, holistic approaches that cover the automation of most or all phases of the investigative process are yet to be developed.

While standardized ways for digital forensic evidence acquisition and preservation are readily available (ISO/IEC 27037:2012,[1] ISO/IEC 27043:2015[2]) (Karie, Kebande, Venter, & Choo, 2019), there

[1] https://www.iso.org/standard/44381.html

is a growing need for making at least part of digital investigative processes autonomous. The enormous growth of digital forensics data volume creates new challenges, making it necessary to employ data mining, data reduction, distributed processing, and other advanced data processing methods, or utilize artificial intelligence (Quick & Choo, 2014). In parallel with the increasing data volume and storage space on storage devices and cloud platforms, the number and variety of computing devices (servers, workstations, laptops, smartphones, IoT devices, etc.) involved in digital forensic investigations is also increasing. These are associated with higher levels of data versatility and heterogeneity, which create more and more barriers to the automation of digital investigations, particulary when proprietary formats are used to store data. One of the options to overcome the processing challeges of heterogeneous digital forensic data is to formally define the concepts of digital forensic investigation processes and the concepts of domain-specific expert knowledge, preferably with their properties, and the relationships between these. Purpose-designed knowledge organization systems can indeed be well utilized in digital forensic investigations to represent case data in machine-interpretable form. This in turn facilitates complex queries and enables automated reasoning, though which software agents can deduct conclusions or abduct the most likely explanation, identify anomalies, and support the creation of case timelines.

The semantics of cyberspace concepts can be efficiently captured using formal knowledge representation, a field of AI, which facilitates complex querying mechanisms and automated reasoning (Sikos, 2018). However, not all data models are suitable for capturing the rich semantics of digital forensic concepts and relationships—see concepts maps, for example, which can visualize concept correlations (Tanner & Dampier, 2009), but do not utilize namespaces.

The main types of formal knowledge organization systems are thesauri, taxonomies, knowledge bases, ontologies, and datasets (Sikos, 2017b). Among these, two types are of particular interest in digital forensics: taxonomies, which define automatically extractable concepts categorized in a concept hierarchy, and ontologies, which are formal conceptualizations of the knowledge domain of interest (here, the digital forensics domain), with complex relationships and rules to be able to automatically generate explicit statements based on implicit knowledge. Such systems enable software tools to achieve task automation for complex processes like digital forensic investigative processes that cannot be otherwise automated. For example, software agents cannot reason over network packet captures from standard (unstructured) packet capture file formats, unless the rich semantics of the associated concepts are captured accurately, such as in the form of machine-interpretable ontology definitions, as seen with the purpose-designed *Packet Analysis Ontology, PAO* (Sikos, 2019). [3] This ontology covers the terminology of packet analysis (aligned with that of

[2] https://www.iso.org/standard/44407.html

[3] https://purl.org/ontology/pao/

Wireshark), including concepts and properties, as well as the associated datatype restrictions, which can be used for automated reasoning in network forensics.

Between the two, there are machine-readable, semistructured representations as well, typically in XML or JSON, such as the XML vocabulary of Garfinkel (2009), however, these are limited compared to those representations that constitute structured data.

Sidebar 1

Formal knowledge representation relies on standards that define data models and file formats so that expert knowledge can be recorded in a uniform manner, processable by software agents. The *Resource Description Framework (RDF)* is a fundamental standard of the Semantic Web for writing structure data, making it possible to create machine-interpretable statements that utilize arbitrary terms from decentralized knowledge organization systems. For this, RDF provides a vocabulary and constructors, a graph-based data model, formal semantics (meaning) and interpretation, and several serialization syntaxes. For creating classes, subclass-superclass relationships, and custom datatype definitions, the *RDF Schema Language (RDFS)* extends the RDF vocabulary. A semantic extension of RDF and RDFS is the Web Ontology Language (OWL), which provides constructors for class relationships from set theory (union, intersection, etc.), property cardinality restrictions (minimum, maximum, exact number), characteristic of properties (transitive, symmetric, etc.), and domain and range restrictions for properties (Sikos, 2015).

The following section reviews milestones of taxonomies and ontologies in digital forensics.

2. KNOWLEDGE ORGANIZATION SYSTEMS IN DIGITAL FORENSICS

Digital forensic investigations, similar to traditional forensic investigations, rely on expert knowledge covering concepts in a taxonomical structure (Pollitt, 2008). However, defining subclass-superclass relationships for concepts in a taxonomy is not sufficient for all application scenarios, many of which require the declaration of property domain and range, complex relationships, datatype declarations, etc. These are only available in ontologies,[4] which can be well utilized in areas such as packet analysis, one of the primary traceback techniques in network forensics (Sikos, 2020). To complement digital forensic ontologies, relevant domain ontologies can be used, such as legal ontologies that cover legal concepts and intellectual property rights (Casellas, 2011).

Although the forensics knowledge organization system of Brinson, Robinson, and Rogers (2006) was originally designed for cyber-forensics teaching, it can be used in other application areas as well. Technically it is a taxonomy, not a fully featured ontology as its name suggests, however, it covers

[4] Note that not all ontologies described in the literature constitute structured data, and as such, are not machine-interpretable, because not all have been defined in a language adequate for this (such as OWL).

digital forensics concepts of different aspects, both technological (hardware and software) and professional (law, academia, military, private sector).

The *Cyber Forensics Ontology for Cyber Criminal Investigation* is an OWL ontology that defines concepts such as cybercrime, law, crime case, criminal, and evidence (Park, Cho, & Kwon, 2009). This ontology provides comprehensive, objective information for criminal investigations of cyber-crimes, and can be used for data mining, classification of crimes by type, extracting similar crime cases and criminals, and detecting similar methods of criminal investigations.

As part of a semantic framework for modeling, analyzing, and reusing digital forensic knowledge, several ontologies have been proposed, including the *Digital Investigation Ontology (DIALOG)* and its sub-ontologies: the *Crime Case Ontology*, the *Information Ontology*, the *Information Location Ontology*, and the *Forensic Resource Ontology* (Kahvedžić & Kechadi, 2009). All of these, as their names suggest, focus on different aspects of digital forensics.

Alzaabi (2013) proposed the combination of domain ontologies and an application ontology in a framework for the forensic analysis of mobile devices. These ontologies can be used to capture the semantics of contacts, social network accounts, IM accounts, email addresses, phone numbers, user messages (SMS, IM, and social network messages), and call logs. Once mobile data is extracted and matched with concepts and relationships of the model, a knowledge base can be generated, which captures not only the concepts of the mobile computing environment being analyzed, but also correlations relevant to the case. This way, the metadata from various evidence resources is superimposed, covering the entire evidence space, i.e., the forensic image of the device being investigated.

The class hierarchy of Karie and Venter (2014) is a taxonomy for the digital forensics domain. It covers concepts of computer, software, database, multimedia, device, and network forensics in a taxonomical structure.

Eden et al. (2015) proposed a forensic taxonomy for SCADA system investigations as a part of a forensic incident response model. This taxonomy covers concepts of traditional forensic triage. In addition, it considers three types of SCADA forensic artifacts: information at the enterprise level, information at the information level, and information at the physical level. SCADA system assets can be described as safety-critical (process/timed/location), mission-critical, and business-critical, which can be used for asset prioritization.

The *Digital Forensic Analysis eXpression (DFAX) Ontology* leverages industry standard CybOX[5] terms (Casey, Back, & Barnum, 2015). It can be used to describe digital forensic cases with concepts such as

[5] Since then, CybOX has been integrated into version 2 of the Structured Threat Information eXpression (STIX 2.0).

victim, subject, examiner, investigator, attorney, authorization, forensic action, evidence record, etc., defined in a taxonomical structure and with typed links. DFAX is efficient is describing digital evidence, actions, action patterns, changes, and the absence of digital evidence.

The *ParFor Ontology*[6] (stands for "Parallax Forensics") is an OWL ontology that defines concepts and properties for various aspects of digital forensics (Turnbull & Randhawa, 2015). For file systems, it defines files, directories, and permissions, for communications, the types of communications. It can also be used to describe computing devices, operating systems, installed software, and system events.

For the forensic investigation of Android communication apps, the *Android Communication App Forensic Taxonomy* has been proposed, defining concepts of 30 popular apps in a concept hierarchy (Azfar & Choo, 2016). This can be used to categorize Android artifacts by exchanged message type (text, voice, group), by timestamps (for request to/from contact and message sent/delivered/received), by group chat membership, and by duration of voice calls. This taxonomy also captures the relationships between Android concepts and properties, such as phone numbers and contact IDs, or users and authentication tokens.

The *Cyber-investigation Analysis Standard Expression (CASE)* ontology[7] aligns with and extends the Unified Cyber Ontology (UCO),[8] and provides more flexibility than DFAX (Casey et al., 2017). Examples of what can be represented using CASE include, but are not limited to, SMS messages, crime scene location, person identity, a software tool and version used for data acquisition (e.g., XRY Logical 8.2.4), and the file system type in a disk partition. The CASE framework supports annotations for labeling, grouping, and custom notes, confidence values for properties and relationships, and provenance data to store the origin of traces.

By increasing the specificity of a taxonomy, the efficiency of investigating a particular type of application might be significantly improved. For example, Cahyani, Choo, Rahman, and Ashman (2018) developed a forensic taxonomy for the systematic classification of forensic artifacts of Windows Mobile 8 dating apps. The artifact categories of this taxonomy cover timestamps, exchanged messages, matching information, location, and user, partner, and app information. These can be utilized in smartphone forensic investigations in event reconstruction and creating the case timeline.

Yaqoob, Hashem, Ahmed, Kazmi, and Hong (2019) developed a taxonomy based on forensic concepts, enablers, networks, evidence sources, investigation modes, forensic models, forensics

[6] https://github.com/benjaminturnbull/ParFor/blob/master/ontology/ParForOntology.ttl

[7] https://github.com/casework/CASE/blob/master/case.ttl

[8] https://github.com/ucoProject/UCO

layers, forensics tools, and forensics data processing, to enable the description of smart objects and their vulnerabilities to IoT cyberattacks. It can help to mitigate privacy risks, integrate IoT data, and preserve user privacy during digital forensic investigations.

The *Web Services Forensic Ontology (WSFO)* is the OWL mapping and extension of the XML Schema-based Incident Object Description Exchange Format (IODEF) defined in IETF RFC 5070[9] (Akremi, Sriti, Sallay, & Rouached, 2019). It covers incidents, hardware, software, personnel qualifications, security, attribution, web services, chain of custody, and weighting for digital forensic concepts and properties. The implementation potential of the ontology has been demonstrated in the form of the prototype software tool Fi4SOA, which allows investigators to set and view preferences on admissibility requirements so that the system can automatically compute the degree of admissibility.

An ontology-based framework has been proposed to enhance the data extraction, conservation, analysis, and documentation of forensic investigations (Amato, Cozzolino, Moscato, & Moscato, 2019). This framework can be used to correlate evidences, thereby providing enhanced retrieval and reasoning for digital forensic investigations. To do so, it instantiates the associated ontology in the form of RDF assertions. As a result, relationships between various concepts, such as file, text, image, person, and location, or registry value, event, person, and file, can be described, which can serve rule-based reasoning.

The *Event-Based Forensic Integration Ontology for Online Social Networks (EFIOSN)* is the implementation of a knowledge model specifically designed for social network forensics (Arshad, Jantan, Hoon, & Abiodun, 2020). In this model, evidence-entity relationships can be characterized using interaction-based, temporal, subject-object, and rule-based correlations, allowing the automation of extraction, analysis, and interpretation of digital forensic data in the social networks' context. The temporal sequence of events can be utilized in efficient timeline construction and analysis.

3. ONTOLOGY-BASED DIGITAL FORENSIC DATA PROCESSING

Utilizing semantics in digital forensic investigations can support the discovery and analysis of patterns of fraudulent activities from diverse data sources through data integration. This requires correlating digital evidence found using various digital forensic tools, and enhance information retrieval and automated reasoning for digital investigations, as seen with the methodology of Amato, Castiglione, Cozzolino, and Narducci (2020). This methodology is compliant with the main phases of digital investigations described in the ISO/IEC 27037:2012 standard,[10] i.e., identification,

[9] https://tools.ietf.org/html/rfc5070

[10] https://www.iso.org/standard/44381.html

collection, acquisition, conservation and transport, analysis, evaluation, and presentation, and utilizes a reference ontology for semantic annotations in RDF, complemented by *SWRL*[11] rules.

The semantic modeling of digital forensic evidence concepts in an ontology makes it possible to represent in a machine-interpretable way why a particular evidence is considered important for a case (Kahvedžić & Kechadi, 2010). The knowledge formally defined in ontologies can be complemented by typed links to related *Linked Open Data (LOD)* datasets for semantic enrichment.

In digital forensic analysis, knowledge engineering can be utilized to capture the following types of knowledge:

- Digital forensics knowledge: knowledge specific to computer forensics, network forensics, smartphone forensics, cloud forensics, etc.
    - Technical knowledge: IT concepts, properties, and entities in the context of digital forensics
        - Background knowledge: ground truth, core definitions
            - Hardware knowledge: concept definitions for and properties of computer components, such as CPU, RAM, etc. (e.g., using the *WICUS Hardware Specs Ontology*)[12]
            - Communication network knowledge: concept definitions for and properties of autonomous systems, network devices, etc. (e.g., using the *Communication Network Topology and Forwarding Ontology (CNTFO)*)[13]
            - IoT knowledge: concept definitions for and properties of IoT networks, sensors, etc. (e.g., using the *IoT Ontology*)[14]
            - Packet analysis knowledge: network packet analysis concept definitions (e.g., using the *Packet Analysis Ontology (PAO)*)[15]

[11] Semantic Web Rule Language, https://www.w3.org/Submission/SWRL/

[12] http://vocab.linkeddata.es/wicus/hwspecs/hwspecs.owl

[13] https://purl.org/ontology/network/

[14] http://ai-group.ds.unipi.gr/kotis/sites/default/files/iot-ontology.owl

[15] https://purl.org/ontology/pao/

        - Digital forensic investigation process knowledge, such as the order of tasks in FTK Imager for data acquisition
        - Cybersecurity knowledge: cybersecurity concepts, such as threat, threat actor, vulnerability, attack vector, etc. (e.g., using the *Unified Cybersecurity Ontology (UCO)*)[16]
        - Case-specific knowledge, such as the MAC address of the computing device of the suspect(s)
    - Legal knowledge, such as the concept of admissible evidence
- World knowledge
    - Commonsense knowledge bases, upper ontologies
    - Temporal knowledge (e.g., using *OWL Time*[17] and the *SWRL Temporal Ontology*)[18]

Chu, Deng, and Chao (2009) developed an ontology-based model for digital forensic investigations in the context of ubiquitous computing. This model can represent complex case scenarios by creating the union of the digital evidence pieces via four facets of the computing environment: identification, location, sensing, and connectivity. In this model, identification covers properties such as MAC address and email account, operating system, SIM card, etc.; location considers IP address and DNS resolution; sensing captures Wi-Fi, 3G, IrDA, Bluetooth, and so on; and connectivity describes available Wi-Fi networks, external drives connected to the system, etc. As an example, a cyber-espionage scenario has been presented by the authors using this ontological model, considering the available wireless networks in the vicinity of the suspect's device at the time of committing the crime, files deleted from the suspect's USB drive depicting the blueprint of the company's new product, email correspondence about an agreement to provide the design specifications of this new product, and that the access required for the highly confidential file of this blueprint was not granted to the suspect, only the CEO.

Dosis, Homem, and Popov (2013) defined ontologies for the digital forensics domain, such as the *Storage Media Ontology* and the *Network Traffic Ontology*. The Storage Media Ontology defines concepts such as media device image, partition, file system, and file, and properties to describe file creation time, size, and MD5 checksum. The Network Traffic Ontology defines, on top of core

[16] https://github.com/Ebiquity/Unified-Cybersecurity-Ontology/blob/master/uco2.ttl

[17] https://www.w3.org/2006/time

[18] http://swrl.stanford.edu/ontologies/built-ins/3.3/temporal.owl

networking concepts seen in other network domain ontologies, concepts that are utilized in digital forensics, such as packet capture file, TCP flow, IP address, etc.

The *Semantic Analysis of Digital Forensic Cases (SADFC)* approach combines a knowledge model and an investigation process model with an ontology-powered architecture, including inference mechanisms and analysis algorithms (Chabot, Bertaux, Nicolle, & Kechadi, 2014). This is based on the formal modeling of digital forensic incidents with subjects, objects, events, and footprints. The entities of the model are linked with composite relationships that can be used for event reconstruction and analysis by utilizing the *Ontology for the Representation of Digital Incidents and Investigations (ORD2I)*, an expressive OWL 2 ontology (Chabot, Bertaux, Nicolle, & Kechadi, 2015).

Cuzzocrea and Pirrò (2016) created an ontology that defines general digital forensics concepts, such as case, related concepts, such as cybercrime and storage, classifies evidence as volatile or non-volatile, etc. Digital evidence annotated with concepts of this ontology are utilized in the integration layer of a framework, semantically enriched with *Linked Data* (structured data with typed links published according to best practices). A cyberattack, for example, can be analyzed using this framework by describing the Wireshark packet capture of the relevant Internet traffic in RDF (using terms of a network domain ontology), the disk image with a disk image ontology, and the firewall logs using a firewall ontology. The semantic integration of these can be used, among other things, for creating RDF statements in the case dataset only if they are supported by multiple data sources, such as an IP address is retrieved from a Wireshark packet capture and also from the firewall log. This can be formulated in the SWRL as follows: `trOnt:hasIP(?x, ?y)^FirewOnt:hasIP(?w, ?z)^swrlb:stringEqual(?y, ?z)→kb:IPtoFwl(?x,?w)`. By querying datasets that utilize terms from the aforementioned ontologies using the query language SPARQL, hypotheses of a case can be tested, such as to verify whether communication between a victim and a host from a known malicious network took place:

```
SELECT ?wirTraffic ?IPDaddr ?network

WHERE {

  ?wirTraffic :toIP ?destIP.

  ?destIP :value ?IPDaddr.

  ?destIP :IPtoWHOIS ?whoisAddr.

  ?whoisAddr :isInRange ?range.

  ?whoisAddr :WHOISAddrToSuspicious ?suspIP.

  ?suspIP :IPToHost ?hostSusp.

  ?hostSusp rdf:type :SuspiciousHost.
```

```
  ?range :inIn ?value.
  ?value :netName ?network.
}
```

In the WHERE clause, first the URI and the IP are extracted, then the data checked on WHOIS (via a user-defined function), the IP range is determined, and the URI of the suspicious addresses identified. The IP is checked against the `SuspiciousHost` class, then the value is obtained from the range, and finally the network name is queried.

Amato, Cozzolino, and Mazzocca (2016) proposed a methodology for the semantic integration, correlation, and querying of digital forensic data sources with the purpose of case reconstruction. They relate three phases of analysis and reconstruction, namely data fusion, correlation, and validation, to Semantic Web technologies. They utilize RDF statements and OWL ontology axioms for fusing heterogeneous digital forensic data derived from diverse sources, SWRL rules for data correlation, logic and proof for validation, and SPARQL for querying. The semantic parser of this methodology reuses existing ontologies as per Semantic Web best practices, but utilizes additional ontologies to integrate referenced domain ontologies with new classes and additional constraints if needed (Amato, Barolli, Cozzolino, Mazzeo, & Moscato, 2018). How deep the analysis of digital incidents can be with such a framework depends on the range of data sources integrated (Amato, Cozzolino, Mazzeo, & Moscato, 2018).

The graphical representation of structured data can be for visualization as well as data fusion. A prime example is *Property Graph Event Reconstruction (PGER)*, which is a data normalization and event correlation system that stores event data in a native graph database for index-free traversal (Schelkoph, Peterson, & Okolica, 2018). It can be used for digital forensic event graph reconstruction. The timeline of cyber-incidents can be efficiently generated using the ontological approach of Bhandari and Jusas, which utilizes purpose-designed ontologies for Windows, Android, and iOS (Bhandari & Jusas, 2020).

Some notable ontology-based digital forensic software tools include the following:

- *Zero Skills Analysis Tool (ZSAT):* allows a detective at a crime scene to scan a computer owned or used by a suspect for possibly suspicious data (Slay & Schulz, 2006). This approach proved successful in finding files and filtering mass amounts of forensic investigation data. However, ZSAT was published as a proof of concept only, rather than a fully featured software tool.

- The ontology-based forensic analysis tool of Alzaabi et al. was designed for analyzing content retrieved from Android smartphones (Alzaabi, Jones, & Martin, 2013). It is part of a framework utilizing multiple domain ontologies and an application ontology. It can be used for the semantic organization of the evidence domain, and to represent associations

between entities, such as to formally describe that a call was established between two persons' phones, or that the phones of two persons were at a particular location at the same time. The RDF-based representation of smartphone data, such as contact lists, call logs, SMS messages, social media messages, etc., can be efficiently queried using SPARQL.

- The ontology-based digital forensics analysis tool of Akremi et al. utilizes the aforementioned WSFO ontology in a GUI tool for reasoning based on user-defined validation rules and queries (Akremi et al., 2019). This tool can be used to generate comprehensive digital forensic reports to present to a court.

4. AUTOMATED REASONING OVER DIGITAL FORENSICS DATA: AN EXAMPLE

To demonstrate how digital forensics data can be formally represented using RDF, consider the following example. Assume that data exfiltration has been committed using desktop computer Nr. 52 of a company (DESKTOP52), and the evidence found on its SSD (SSD1DP52). By declaring the name for the suspect's computer and considering that the domain of the "has name" property has been defined in an ontology to be the "suspect computer" class, formally (in RDF Turtle serialization),

```
case:DESKTOP52 :hasName "WKS52" .
```

```
:hasName rdfs:domain df:SuspectComputer .
```

and using the second rule of standard RDFS entailment (rdfs2),[19] it can be automatically inferred that desktop computer Nr. 52 is a suspect's computer, i.e.,

```
case:DESKTOP52 rdf:type df:SuspectComputer .
```

which was not explicitly stated in the knowledge base. Similarly, by declaring that the SSD of the suspect's computer contains evidence, and using the definition of the "contains evidence" property being the inverse property of "found on," i.e.,

```
case:DP52SSD1 df:containsEvidence case:EVIDENCE1 .
```

```
df:containsEvidence owl:inverseOf df:foundOn .
```

and using the OWL 2 reasoning rule for inverse properties (prp-inv1),[20] it can be inferred that

```
case:EVIDENCE1 df:foundOn case:DP52SSD1 .
```

which, again, was not explicitly stated in the knowledge base of the case.

[19] https://www.w3.org/TR/rdf11-mt/#patterns-of-rdfs-entailment-informative

[20] https://www.w3.org/TR/owl2-profiles/#Reasoning_in_OWL_2_RL_and_RDF_Graphs_using_Rules

As seen in this example, digital forensic data can be efficiently represented formally using the combination of ontological background knowledge and case-specific knowledge derived from the knowledge base of the case, and automated reasoning performed using the standard RDFS/OWL entailment regimes.

Challenges and Future Direction

While ontology-based digital forensic investigation approaches have real benefits in terms of querying, data aggregation, and data fusion, they rely on structured data to reach their full potential. Considering the wide and continuously expanding range of computing devices, however, data acquisition often produces forensically sound data without capturing the associated semantics. Therefore, one of the major challenges is how to automatically generate structured data from hybrid data efficiently. This requires a holistic approach that takes into account operating system differences, and the de facto standard semistructured formats implemented on many IoT devices.

Another challenge is the standardization of digital forensic ontologies, because most approaches and frameworks seen in the literature apply custom application and/or domain ontologies. Proprietary implementations prevent widespread use, and ultimately, global deployment.

Nevertheless, the ever-growing volume, variety, and velocity of data related to, and generated by, computing devices urges task automation, for which a promising direction is the development of ontology-based investigation tools that mainstream automation in digital investigation pipelines. These call for new ways to represent investigation data, such as in the form of knowledge graphs, and employ machine learning to find anomalies and identify patterns in digital forensic investigation data.

Sidebar 1

Semantic Web Standards, Backed by Mathematics

To ensure that meaning can be associated with every statement, RDF and OWL rely on model-theoretic semantics.
The RDF data model has its roots in graph theory, as each collection of RDF statements intrinsically represents a labeled, directed multigraph.
OWL ontologies can be formally grounded in description logics, most of which are decidable fragments of first-order logic. The letters in description logic names indicate the available mathematical constructors/relationships from set theory (U – concept union, I – inverse roles, Q – qualified cardinality restrictions, etc.)—the more the available constructors, the higher the expressivity and reasoning complexity (Sikos, 2017a). Such well-understood computational properties make it possible, among other things, to ensure decidability so that reasoning engines will

never run in an infinite loop.

Conclusions

While automation is highly desired in digital forensic investigations facing an exponential increase in data volume, variety, and complexity, only a few research fields offer solutions to overcome the associated challenges, one of which is knowledge representation, a field of artificial intelligence. The semantic representation of digital forensic case data facilitates data integration and the partial automation of digital forensic investigations via automated reasoning, event correlation analysis, knowledge consistency checking, anomaly detection, and timeline generation. Purpose-designed knowledge organization systems, and digital forensic ontologies in particular, provide the formal definition for concepts and properties in this domain, focusing either on the digital forensic investigation process, on the actual digital artifacts of a particular case, or both. Other ontologies codify expert knowledge by defining concepts and properties for specific computing areas, e.g., that of the packet analysis domain (to be used for network forensic investigations, such as downloading or distribution of illegal material), or the communication network domain (to be used for cyberattack investigations). These all differ in terms of scope and specificity, which determines the primary application area for each.

By formally capturing the semantics of the objects of digital forensic investigations, 1) heterogeneous investigation data can be represented uniformly, ready to be processed by software agents, allowing complex queries to be executed on case datasets; 2) implicit knowledge can be made explicit automatically via reasoning; and 3) data integration becomes not only possible, but even efficient. This requires preprocessing that calls for the standardization of data formats of computing devices, and IoT devices in particular, but well worth the effort, because it enables other automation mechanisms to aggregate and fuse data, such as via rule-based systems and machine learning.

References

Akremi, A., Sriti, M.-F., Sallay, H., & Rouached, M. (2019). Ontology-Based Smart Sound Digital Forensics Analysis for Web Services. *International Journal of Web Services Research, 16*(1). doi:10.4018/IJWSR.2019010104

Alzaabi, M. (2013). *Ontology-Based Forensic Analysis of Mobile Devices.* Paper presented at the 20th International Conference on Electronics, Circuits, and Systems Abu Dhabi, UAE.

Alzaabi, M., Jones, A., & Martin, T. A. (2013). *An Ontology-Based Forensic Analysis Tool.* Paper presented at the ADFSL ConferenceonDigital Forensics, Security and Law, Richmond, VA, USA.

Amato, F., Barolli, L., Cozzolino, G., Mazzeo, A., & Moscato, F. (2018). *Improving Results of Forensics Analysis by Semantic-Based Suggestion System.* Paper presented at the 6th International Conference on Emerging Internetworking, Data & Web Technologies, Tirana, Albania.

Amato, F., Castiglione, A., Cozzolino, G., & Narducci, F. (2020). A Semantic-Based Methodology for Digital Forensics Analysis. *Journal of Parallel and Distributed Computing, 138*, 172–177. doi:10.1016/j.jpdc.2019.12.017

Amato, F., Cozzolino, G., Mazzeo, A., & Moscato, F. (2018). *An Application of Semantic Techniques for Forensic Analysis.* Paper presented at the 32nd International Conference on Advanced Information Networking and Applications Workshops, Krakow, Poland.

Amato, F., Cozzolino, G., & Mazzocca, N. (2016). *Semantic Integration and Correlation of Digital Evidences in Forensic Investigations.* Paper presented at the 11th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing, Asan, Korea.

Amato, F., Cozzolino, G., Moscato, V., & Moscato, F. (2019). Analyse Digital Forensic Evidences through a Semantic-Based Methodology and NLP Techniques. *Future Generation Computer Systems, 98*, 297–307. doi:10.1016/j.future.2019.02.040

Andriotis, P., Tryfonas, T., Oikonomou, G., Li, S., Tzermias, Z., Xynos, K., . . . Prevelakis, V. (2014). *On the Development of Automated Forensic Analysis Methods for Mobile Devices.* Paper presented at the International Conference on Trust and Trustworthy Computing, Heraklion, Crete, Greece.

Arshad, H., Jantan, A., Hoon, G. K., & Abiodun, I. O. (2020). Formal Knowledge Model for Online Social Network Forensics. *Computers & Security, 89*. doi:10.1016/j.cose.2019.101675

Azfar, A., & Choo, K. K. R. (2016). An Android Communication App Forensic Taxonomy. *Journal of Forensic Sciences, 61*(5), 1337–1350. doi:10.1111/1556-4029.13164

Bhandari, S., & Jusas, V. (2020). An Ontology Based on the Timeline of Log2timelineand Psort Using Abstraction Approach in Digital Forensics. *Symmetry, 12*(4). doi:10.3390/sym12040642

Brennan, F., Udris, M., & Gladyshev, P. (2013). *An Automated Link Analysis Solution Applied to Digital Forensic Investigations.* Paper presented at the International Conference on Digital Forensics and Cyber Crime, Moscow, Russia.

Brinson, A., Robinson, A., & Rogers, M. (2006). A Cyber Forensics Ontology: Creating a New Approach to Studying Cyber Forensics. *Digital Investigation, 3*, 37–43. doi:10.1016/j.diin.2006.06.008

Cahyani, N. D. W., Choo, K. K. R., Rahman, N. H. A., & Ashman, H. (2018). An Evidence-based Forensic Taxonomy of Windows Phone Dating Apps. *Journal of Forensic Sciences, 64*(1), 243-253. doi:10.1111/1556-4029.13820

Cantrell, G., & Dampier, D. (2013). *Evaluation of the Semi-automated Crime-Specific Digital Triage Process Model.* Paper presented at the IFIP International Conference on Digital Forensics, Orlando, FL, USA.

Casellas, N. (2011). Legal Ontologies. In *Legal Ontology Engineering: Methodologies, Modelling Trends, and the Ontology of Professional Judicial Knowledge*. Dordrecht: Springer.

Casey, E., Back, G., & Barnum, S. (2015). Leveraging CybOX™ to Standardize Representation and Exchange of Digital Forensic Information. *Digital Investigation, 12*, S102–S110. doi:10.1016/j.diin.2015.01.014

Casey, E., Barnum, S., Griffith, R., Snyder, J., Beek, H. v., & Nelson, A. (2017). Advancing Coordinated Cyber-Investigations and Tool Interoperability Using a Community Developed Specification Language. *Digital Investigation, 22*, 14–45. doi:10.1016/j.diin.2017.08.002

Chabot, Y., Bertaux, A., Nicolle, C., & Kechadi, M.-T. (2014). A Complete Formalized Knowledge Representation Model for Advanced Digital Forensics Timeline Analysis. *Digital Investigation, 11*, S95–S105. doi:10.1016/j.diin.2014.05.009

Chabot, Y., Bertaux, A., Nicolle, C., & Kechadi, T. (2015). An Ontology-Based Approach for the Reconstruction and Analysis of Digital Incidents Timelines. *Digital Investigation, 15*, 83–100. doi:10.1016/j.diin.2015.07.005

Chu, H.-C., Deng, D.-J., & Chao, H.-C. (2009). An Ontology-Driven Model for Digital Forensics Investigations of Computer Incidents under the Ubiquitous Computing Environments. *Wireless Personal Communications, 56*, 5–19. doi:10.1007/s11277-009-9886-x

Cuzzocrea, A., & Pirrò, G. (2016). *A Semantic-Web-Technology-Based Framework for Supporting Knowledge-Driven Digital Forensics.* Paper presented at the 8th International Conference on Management of Digital EcoSystems.

Dosis, S., Homem, I., & Popov, O. (2013). Semantic Representation and Integration of Digital Evidence. *Procedia Computer Science, 22*, 1266–1275. doi:10.1016/j.procs.2013.09.214

Eden, P., Blyth, A., Burnap, P., Cherdantseva, Y., Jones, K., & Soulsby, H. (2015). *A Forensic Taxonomy of SCADA Systems and Approach to Incident Response*. Paper presented at the 3rd International Symposium for ICS & SCADA Cyber Security Research 2015, Germany.

Garfinkel, S. L. (2009). *Automating Disk Forensic Processing with SleuthKit, XML and Python.* Paper presented at the Fourth International IEEE Workshop on Systematic Approaches to Digital Forensic Engineering, Berkeley.

Kahvedžić, D., & Kechadi, T. (2009). DIALOG: A Framework for Modeling, Analysis and Reuse of Digital Forensic Knowledge. *Digital Investigation, 6*, S23–S33. doi:10.1016/j.diin.2009.06.014

Kahvedžić, D., & Kechadi, T. (2010). *Semantic Modelling of Digital Forensic Evidence.* Paper presented at the Second International Conference on Digital Forensics and Cyber Crime, Abu Dhabi, UAE.

Karie, N. M., Kebande, V. R., Venter, H. S., & Choo, K.-K. R. (2019). On the Importance of Standardising the Process of Generating Digital Forensic Reports. *Forensic Science International: Reports, 1*. doi:10.1016/j.fsir.2019.100008

Karie, N. M., & Venter, H. S. (2014). Toward a General Ontology for Digital Forensic Disciplines. *Journal of Forensic Sciences, 59*(5), 1231–1241. doi:10.1111/1556-4029.12511

Park, H., Cho, S., & Kwon, H.-C. (2009). *Cyber Forensics Ontology for Cyber Criminal Investigation.* Paper presented at the International Conference on Forensics in Telecommunications, Information, and Multimedia, Adelaide, Australia.

Pollitt, M. (2008). Applying Traditional Forensic Taxonomy to Digital Forensics. In S. S. Indrajit Ray (Ed.), *Advances in Digital Forensics IV* (pp. 17–26). Boston, MA: Springer.

Quick, D., & Choo, K.-K. R. (2014). Impacts of Increasing Volume of Digital Forensic Data: A Survey and Future Research Challenges. *Digital Investigation, 11*(4), 273–294. doi:10.1016/j.diin.2014.09.002

Schelkoph, D. J., Peterson, G. L., & Okolica, J. S. (2018). *Digital Forensics Event Graph Reconstruction.* Paper presented at the 10th International EAI Conference, New Orleans.

Shosha, A. F., James, J. I., Liu, C.-C., & Gladyshev, P. (2012). *Towards Automated Forensic Event Reconstruction of Malicious Code*. Paper presented at the International Workshop on Recent Advances in Intrusion Detection, Amsterdam, The Netherlands.

Sikos, L. F. (2015). Knowledge Representation. In *Mastering Structured Data on the Semantic Web* (pp. 13–57). Berkeley, CA: Apress.

Sikos, L. F. (2017a). Description Logics: Formal Foundation for Web Ontology Engineering. In *Description Logics in Multimedia Reasoning*. Cham: Springer.

Sikos, L. F. (2017b). Knowledge Representation with Semantic Web Standards. In *Description Logics in Multimedia Reasoning* (pp. 11–49). Cham: Springer.

Sikos, L. F. (2018). OWL Ontologies in Cybersecurity: Conceptual Modeling of Cyber-Knowledge. In L. F. Sikos (Ed.), *AI in Cybersecurity* (pp. 1–17). Cham: Springer.

Sikos, L. F. (2019). *Knowledge Representation to Support Partially Automated Honeypot Analysis Based on Wireshark Packet Capture Files.* Paper presented at the 11th KES International Conference on Intelligent Decision Technologies, St. Julians, Malta.

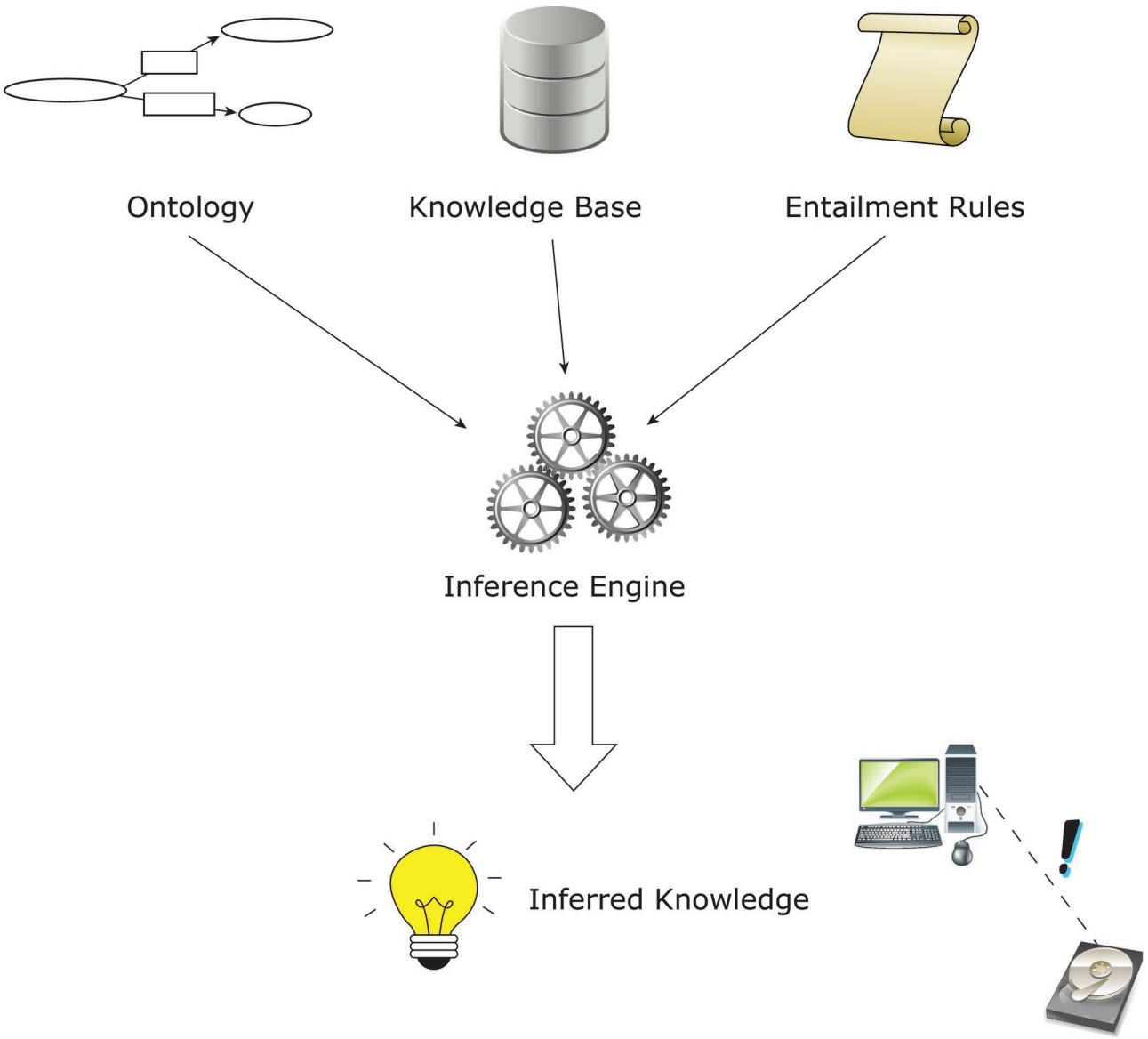
Sikos, L. F. (2020). Packet Analysis for Network Forensics: A Comprehensive Survey. *Forensic Science International: Digital Investigation, 32*, 200892. doi:10.1016/j.fsidi.2019.200892

Slay, J., & Schulz, F. (2006). Development of an Ontology-Based Forensic Search Mechanism: Proof of Concept. *The Journal of Digital Forensics, Security and Law, 1*(1). doi:10.15394/jdfsl.2006.1002

Tanner, A., & Dampier, D. (2009). *Concept Mapping for Digital Forensic Investigations.* Paper presented at the 2009 IFIP International Conference on Digital Forensics, Orlando, FL, USA.

Turnbull, B., & Randhawa, S. (2015). Automated Event and Social Network Extraction from Digital Evidence Sources with Ontological Mapping. *Digital Investigation, 13*, 94–106. doi:10.1016/j.diin.2015.04.004

Yaqoob, I., Hashem, I. A. T., Ahmed, A., Kazmi, S. M. A., & Hong, C. S. (2019). Internet of Things Forensics: Recent Advances, Taxonomy, Requirements, and Open Challenges. *Future Generation Computer Systems, 92*, 265–275. doi:10.1016/j.future.2018.09.058



wfs2_1394_wire-ai_ in_digital_forensics_visual_abstract.eps