Ontology in the Digital Forensics Domain: A Scoping Review

Martin Morgenstern[1], Johannes Fähndrich[2] and Wilfried Honekamp[3]

Abstract: A need for the use of automation in digital forensics is imminent. With the overwhelming workload for analysing collected digital evidence, law enforcement agencies are no longer meeting the quality investigations we expect. For automation to work (e. g., integrating heterogeneous data sources, structuring unstructured data, or drawing conclusions from structured data), formalisation must be specified. Specification of formalisation includes as a first step to describe concepts in the domain of digital forensics. In this contribution, we analyse the state of the art of ontological formalisation in the domain of digital forensics via a scoping review. There are some attempts to formalise the technical domain of digital forensics in ontologies, but they do not cover essential context, like technical annotations or acquisition modelling. Future work will be to use the existing foundations and extend them with facts and rules to enable reasoning.

Keywords: Formalisation, Digital Forensics, Ontology.

1 Introduction and Background

When collecting evidence in criminal cases, digital traces must be identified, recorded and evaluated. Due to the quantity and heterogeneity (see Figure 1) of forensically relevant digital data, it is no longer feasible for investigating authorities to perform a high-quality manual analysis of the same. In order to automate the analysis of extensive and heterogeneous data in a meaningful way, for example by using artificial intelligence, a generally applicable ontology is needed for the field of digital forensics. This paper aims to highlight the current state of development on the topic of ontologies for general IT forensics, as well as the resulting research gaps to provide a basis for the development of forensic AI systems. The overarching goal is a proposal for a uniform use of technical language. Therefore, it will be investigated which ontological prerequisites have to be created in IT forensics to enable a meaningful use of artificial intelligence for IT forensic analysis, and it will be reviewed, which of these prerequisites are already fulfilled.
In order to solve crimes, facts have to be found out and documented in a way that can be used in court. Very often, this requires the use of forensic experts. Forensics is the science that deals with the identification, analysis and reconstruction of criminal acts.

[1] Hochschule Stralsund, IACS, Zur Schwedenschanze 15, Stralsund, 18435, martin.morgenstern@hochschule-stralsund.de
[2] Hochschule für Polizei Baden-Württemberg, Johannesfaehndrich@hfpol-bw.de
[3] Hochschule Stralsund, IACS, Zur Schwedenschanze 15, Stralsund, 18435, wilfried.honekamp@hochschule-stralsund.de, https://orcid.org/0000-0003-2931-7047

doi:10.18420/inf2022_05

Forensics is divided into several subfields. One subfield is computer forensics, which is a synonym for digital forensics [Ch09]. [Si18]

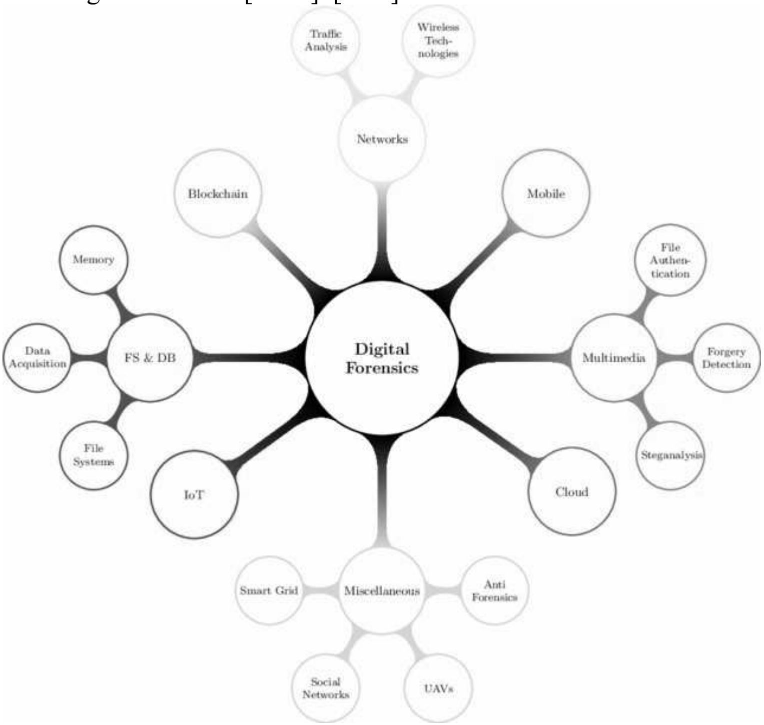

Figure 1: Sub-areas of digital forensics [Ca21]

Figure 1 shows an overview of identified sub-areas of digital forensics. However, the sub-areas can only ever be a snapshot, as information technology is constantly changing and evolving. The authors consider it problematic that they are classified according to data origin. Povalej et al. already drew attention to the problem in 2021 and asked whether such a division is still up-to-date or whether the classification should be done according to the type of digital traces. [Po21].

Another term for digital forensics is IT forensics. In this paper, the definition of digital forensics used is that of the German Federal Office for Information Security (BSI). According to this definition, digital forensics is "the strictly methodical analysis of data on data carriers and in computer networks to clarify incidents, including the possibilities of strategic preparation, especially from the point of view of the system operator of an IT system." [BS11]

The complexity and quantity of digital forensics is increasing every year. On the one hand, the amount of data is permanently increasing, and on the other hand, so is the number of possible data sources. One reason for this are rapidly falling prices for large data media. [Al13] In traditional IT forensics, data carriers are removed, copied for forensic purposes and analysed. Supporting tools are used for the evaluation. The use of artificial intelligence could significantly facilitate IT forensic work. In addition to legal hurdles, the lack of ontological integration of heterogeneous data is also an obstacle to the use of AI in digital investigations [Po21], [Ho22]. Errors in the evaluation of digital traces can have fatal, e. g. financial or social, consequences. One cause of evaluation errors can be ignorance of exactly how forensic evaluation software works or the assumption that the software always works without errors. In 2016, for example, a suspect was nearly acquitted because the forensic evaluation software failed to process a memory area on the smartphone being evaluated. In individual areas of computer forensics, e. g. mobile forensics, cooperation with external service providers is necessary. For the cooperation with external service providers, the standardisation of forensic procedures is useful. [He21]

For successful cooperation between different organisations, in addition to the standardisation of processes, a common description of the same is also necessary. However, in the field of digital forensics, there is a lack of a generally accepted general ontology [Gr17]. Ontologies are used to represent and provide a basis for logical reasoning in a domain, or for objectifying knowledge (for example, in a knowledge graph [So99]). There are several definitions for the term ontology. [Ka14], [He05].

According to Gruber's definition, one of the most well-known definitions of ontology, an ontology is a common specification to share concepts. Through an ontology a knowledge domain, a domain, is described. Another definition of ontology that may be helpful here is "An ontology is a formal, explicit specification of a shared conceptualisation." [St98]. In this description, a standardised terminology is used, and relationships and derived rules are defined. In many knowledge domains there are multiple or competing ontologies. An ontology gains value the more it is recognised in the professional community. [He05]

For example, a technical application of ontologies is illustrated by Linked Open Data [Mc20]. Currently, while there are many digital forensics ontologies, none are widely used. In the past, there have been several approaches for the introduction of a general ontology in digital forensics, but these do not yet meet the current requirements. The requirements for an ontology for digital forensics differ depending on the literature. From the literature research some requirements emerge, which can be found again and again. In particular, ontologies should be able to describe data from heterogeneous sources. In a modern ontology for digital forensics, cross-agency and cross-national work should also be possible. The authors assume that a common and standardised ontology will improve communication between investigative and law enforcement agencies and, if necessary, external forensics experts, thus contributing to an increase in efficiency in law enforcement. Ontologies are used to support the investigation of data from different types of sources in various application areas of digital forensics [Av10]. One of the main challenges is to find a way to automatically create structured data from different data sources [Si21]. The need for formalisation in digital forensics was noted as early as 2004 by Bruschi, Monga, and Marignoni. At the digital forensic Research Conference, a hypothesis-based scheme for formalising contexts in IT forensic analysis was presented. However, no ontology was developed in this work. [Br04]

Model driven engineering (MDE) abstracts models domains and implementations for formalisation and normalisation [Fa04]. With MDE one idea is to domain-specific modelling language to aid “formalisation the application structure, behaviour and requirements” [Sc06]. MDE enables the description of self-adapting systems [Tr18]. For self-adaptation to be useful, the semantics of the models used can support the adaptation process. Based on this idea of MDE, semantic service description can be seen as a machine-readable model including meta-models, transformations and semantics. An ontology formalizes the concepts used in a domain to describe the parameters and results of services. „High quality domain ontologies are essential for successful employment of semantic Web services“ [Sa05] (p. 1).

The idea of this formalisation is to formally extract shared knowledge in order to facilitate a common explicit understanding of the field of digital forensics. The idea is to create a common language, structured documentation and tools, which stand up in curt to structure forensic work. Much of the knowledge about digital forensics is implicit in experts, and is built upon experience. The aim is to share this knowledge. First formation of a vocabulary is part of the formalisation with an ontology. We see this as an extension of the Linked Open Data, where knowledge of different domains has been collected and formalised for others to use.

Similar to Euzenat and Shavaiko [Eu07] we define an Ontology the following way:

$$\mathbf{o} = \{C, I, R, \leq, \perp, \in, =\}$$

With C is the set of classes, I is the set of Individuals, R is the set of relationships, $\leq$ is a relation on $(C \times C) \cup (R \times R) \cup (T \times T)$ called specialization, $\perp$ is a relation on $(C \times C) \cup (R \times R) \cup (T \times T)$ called exclusion, $\in$ is a relation over $((I \times C) \cup (V \times T)$ called instantiation, = is a relation over $I \times (I \cup V)$ called assignment.

Euzenat and Shavaiko [Eu07] (p. 39), define that expression can be done and are satisfied by an interpretation if it is coherent. We then are able to create formulars over this ontology with e.g., an implementation like Web Ontology Language (OWL): This gives us basic entities, which can have relation with other entities or themselves. We will use this formal definition to ground further work in OWL, regarding Resource Description Framework (RDF) triplets and the formulation of entities and relations as well as the instantiation by e.g., datatypes and Individuals. Thus, we will use this kind of ontology to state state facts (Individuals and their relationships) about entities (TBox) and facts about the abstract classes (ABox) (concepts and their relationships). This can be seen as a formal grounding of Cyber-investigation Analysis Standard Expression (CASE) or Unified Cyber Ontology (UCO) and the reuse of this first formalization. The benefits of formalising digital forensics were presented in detail in Dewald's dissertation. In this, digital traces were formalised based on an abstract model. Through the formalisation it was possible to show if and which relationship exists between different traces. [De12]

2 Method

To capture the current state of development of ontologies in digital forensics, a scoping review was conducted. A scoping review is a suitable method to obtain an overview of

the evidence of a research topic. Scoping review can identify research gaps in the area under investigation. [Vo19]
Alternativ zum Scoping Review gibt es weitere wissenschaftliche Methoden um einen Überblick über die relevante Literatur, sowie dem Stand der Forschung zu erhalten. Alternativ hätte z. B. eine strukturierte Literaturrecherche gewählt werden können. Der Vorteil des Scoping Reviews ist die hohe Effienz, da mit wenig Zeitaufwand aktuelle wissenschaftliche Fragestellungen identifiziert werden können. [Vo19]
Only German and English language sources were reviewed for the scoping review. Google and Google Scholar were used to search for relevant literature. Search terms were 'ontology digital forensics' as well as synonyms in German and English. For both Google and Google Scholar, only sources on the first five pages of search results were used. Furthermore, the Google Scholar function "cited by" was used to find further relevant documents. All searches took place in May 2022.

3 Results

In the following, the results of the literature review from 2009 until today are presented chronologically. The need for ontology for digital evidence analysis was identified at least in 2009, when Kahvedžić and Kechadi described the DIALOG framework for modelling knowledge in the field of digital forensics (see Figure 2). For a detailed description, the sub-ontologies criminal case, information, information location, and forensic resource were used. The example of the Windows Registry was used to illustrate the use of DIALOG. digital forensics is the central element of the ontology. The first level of sub ontologies are criminal case, information, location information and forensic recovery. The authors of DIALOG have indicated in their work that further work will be needed to consider additional use cases, such as file system analysis or mobile phone forensics. [Ka09]

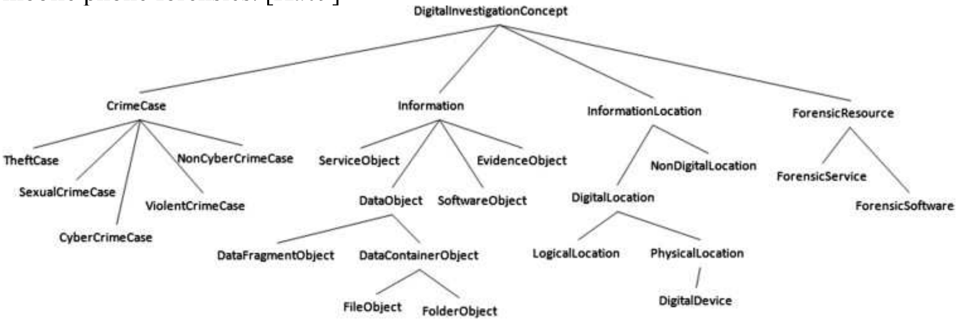

Figure 2: Hierarchy of the DIALOG framework [Ka09]

Back in 2011, Poisel and Tjoa noted that there has been progress in the area of IT forensics ontology, but its unification of standardisation is the biggest challenge. [Po11] Especially the comparability of created reports was seen by Bariki, Hashmi, and Baggili as a problem to be solved. Therefore, an XML-based proposed solution was published.

The development or extension of a standard that also includes other areas such as memory or network forensics was seen as a future challenge. [Ba11] Similarly, in 2012, a conference paper from n Ćosić and Ćosić detailed the need for a unified ontology for digital forensics [Co12].

In 2013, a framework for IT forensic analysis based on semantic web technologies was presented that was based on an ontology. Although semantic ontologies for the digital forensics domain already existed at that time, according to the authors, they focused only on evidence management and forensic reporting. However, the focus of the new framework was on digital evidence analysis. The goal was to develop a framework that would make it possible to link information from different sources in a way that would result in new information. [Al13]

In 2015, Casey et al found that there was no widely accepted standard representation in the digital forensics domain. The existing Structed Thread Information eXpression (Stix) and other models for exchanging legally relevant information could not specifically address the specifics of the digital forensics domain. As a result, it would not be possible to compare investigative results across tools and agencies. Various standards, such as digital forensic XML, have been developed to enable the exchange of IT forensic information. However, none of these standards has gained widespread acceptance. The development of the digital forensic Analysis eXpression (DFAX) ontology was intended to solve the problems described above. Included in DFAX were many improvements over previous standards, such as including chain of custody evidence and victim and investigator actions [Ca15].

Ćosić and Baca developed the Digital Evidence Framework in 2015. The Developed Ontology was intended to solve recognised problems of previous ontologies. One of the innovations was that the integrity of digital evidence should also be considered in the ontology. Lack of standards and processes for sharing information between different agencies and states was identified as a problem to be solved [Co15].

As of October 2016, DFAX is no longer under development. Since that time, the project page states that information about a new project under development will be linked there. [Ba16] Even for basic terms like artifact, there were different definitions. This problem also highlighted the need for a universally accepted and recognised ontology for digital forensics. Harichandran et al. sought a formal definition for digital forensic artifacts [Ha16].

There was still a need for a standard to represent digital evidence and its interrelationships. However, DFAX did not meet all the requirements. For example, it was not possible to represent deleted data in this schema. A new standard was developed in the form of the Cyber-investigation Analysis Standard (CASE). The development of the new standard incorporated experience from many previous standards, including DFXML and DFAX. Improvements were made for national and international exchange of information [Ca17].

Semantic and ontological linking of forensically relevant information has received increasing attention from scholars in recent years [Bh20]. For example, in 2019, Amato et al. presented a semantic methodology based on Natural Language Processing (NLP) that used log data as an example to represent unstructured data in a standardised format. [Am20] OliveiraJr et al. published a paper on experimental digital forensics. The goal was to find a method to describe experiments in the domain of digital forensics in a repeatable way. The development of an ontology for experimental digital forensics was identified as a future task. [Ol20]

In digital forensics today, a variety of different data sources and types are evaluated. In addition to traditional data carriers, online profiles, cloud data, IoT devices with different operating systems and file systems, smartphones, and other devices often need to be evaluated today. In addition to the quantity of data and devices to be evaluated, the complexity is also increasing, for example due to encryption. For many areas of digital forensics, there were 2020 ontologies, e.g., for social networks or web services forensics. While technical capabilities for analysing data continue to evolve, there is little research in digital forensics ontology. There is a need for standardisation of digital forensics ontologies, as existing ontologies are not widely used or belong to subset of digital forensics, e.g., cloud forensics ontologies [Ke20]. The ontology to be developed should be developed globally and have wide user acceptance. Standardisation of data is necessary to process it in an automated way, such as rule-based or machine learning [Si21].

Recently, Solanke and Biasiotti published a paper on Formalisation of AI in digital forensics. In it, they presented techniques for analysing the effectiveness of classification and regression in digital forensics. Further development of methods for evaluating AI techniques in digital forensics is seen as a promising development [So22].

4 Discussion and Conclusions

There are first attempts to formalise the technical domain of digital forensics in an ontology. However, only CASE as an extension of UCO has a description logic class of ALUO(D). Here Object-Properties, data properties and individual types are defined. As a formalisation, this means that a solid foundation has been established. However, in order to make such an ontology usable for automation, for example, technical annotations or acquisition modelling is still needed. For example, an IP address is modelled, but not yet what type of IP exists on a system (e.g., public or private) and where that IP appears in an investigation, for example, the IP of the victim or the perpetrator. Another annotation missing is the tools of the operating systems by means of which an IP can be collected, documented, and used in an investigation. So, for the example IP, ipconfig or ifconfig should be described for the local view, nslookup for the resolution, ping or tracert/traceroute for the classification in a network and so on. Criticism of the previous approaches is that mostly only the classification (like the introduction of IPv4 and IPv6 as a subclass of the IP address) has been modelled here, but hardly any further forensic knowledge has been formalised. The technical knowledge that an IPv4 is also an IP address is basic but not yet forensic. What is important for an IP address in an investigation (deletion periods, resolution at ISP, individualisation through assignment to MAC address, etc.) is missing here. We implemented an example of the CASE and UCO ontology to show how the implementation of OWL Object-Properties and Individuals can be used to describe knowledge of the utilisation of tools, their parameters, and the interrelation of the described entities in the ontology. This has the goal to automate the appropriate tasks to make the work in digital forensics manageable in the future. We used an example of an IP address, a MAC address, its assignment to the network interface, and the appropriate tools.

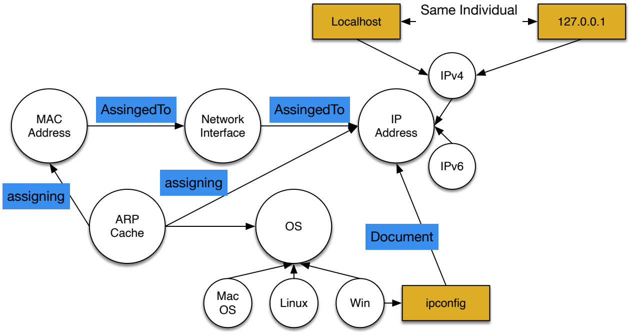

Figure 3: Extension of the CASE ontology by example Individuals and Object-Properties.

Figure 3 shows how the relationship to tools like ipconfig and the use of Object-Properties to externalise the knowledge that the ARP cache holds information about which IP address has been assigned to which MAC address. The aim is to use reasoners and semantic service descriptions to automate some of the information extraction processes in digital forensics cases.

So future work will be to use the existing basis of UCO and CASE and extend it with rules of the Semantic Web Rule Language, and for example Linear Time Logic, to be able to formalise facts like "IP addresses are collected by the ISP until the deletion period of data retention is reached and therefore the connection owner must be documented with a request for information until then". Reasoning could then take place on this, which, for example, sends requests independently or prioritises for which trace situation which necessary next steps must be initiated.

Bibliography

[AL13] Alzaabi, M., Jones, A., Martin, T.A.: An Ontology-Based Forensic Analysis Tool. ADFSL Conference on Digital Forensics, Security and Law, 2013.

[Am20] Amato, F., Castiglione, A., Cozzolino, G., Narducci, F.: A semantic-based methodology for digital forensics analysis. J. Parallel Distrib. Comput. 138, 172–177, 2020.

[Av10] Avni, O., Knierim, T.: Carving und semantische Analyse in der digitalen Forensik. Frauenhofer IGD-A8 Sicherheitstechnologie. 2010.

[Ba16] Back, G.: dfax (DEPRECATED). DFAX. 2016. github.com/DFAX/dfax (accessed 5.23.22).

[Ba11] Bariki, H., Hashmi, M., Baggili, I.: Defining a Standard for Reporting Digital Evidence Items in Computer Forensic Tools. In: Baggili, I. (eds) Digital Forensics and Cyber Crime. ICDF2C 2010. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol 53. Springer, Berlin, Heidelberg. 2011.

[Bh20] Bhandari, S., Jusas, V.: An Ontology Based on the Timeline of Log2timeline and Psort Using Abstraction Approach in Digital Forensics. Symmetry 12, 642. 2020.

[Br04] Bruschi, D., Mongra, M., Martignoni, L.: How to Reuse Knowledge about Forensic Investigations, The Digital Forensic Research Conference. Baltimore. 2004.

[BS11] BSI, 2011. Leitfaden IT-Forensik. Bundesamt für Sicherheit in der Informationstechnik. Bonn. 2011

[Ca15] Casey, E., Back, G., Barnum, S.: Leveraging CybOXTM to standardize representation and exchange of digital forensic information. Digit. Investig. 12, pp. 102-110. 2015.

[Ca17] Casey, E., Biasiotti, M.A., Turchi, F.: Using Standardization and Ontology to Enhance Data Protection and Intelligent Analysis of Electronic Evidence. 2017.

[Ca21] Casino, F., Dasaklis, T., Spathoulas, G., Anagnostopoulos, M., Ghosal, A., Borocz, I., Solanas, A., Conti, M., Patsakis, C.: A cross-domain qualitative meta-analysis of digital forensics: Research trends, challenges, and emerging topics., 2021.

[Ch09] Chu, H.-C., Deng, D.-J., Chao, H.-C.: An Ontology-driven Model for Digital Forensics Investigations of Computer Incidents under the Ubiquitous Computing Environments. Wirel. Pers. Commun. 56, pp. 5-19. 2009.

[Co15] Ćosić, J., Baca, M.: Leveraging DEMF to Ensure and Represent 5ws&1h in Digital Forensic Domain 13, 5. 2015.

[De12] Dewald, A.: Formalisierung digitaler Spuren und ihre Einbettung in die Digitale Forensik. Universität Erlangen-Nürnberg. 2012.

[Eu07] Euzenat, J., & Shvaiko, P. (2007). Ontology matching (Vol. 18). Heidelberg: springer.

[Fa04] Favre, J. M. (2004, October). Towards a basic theory to model model driven engineering. In 3rd workshop in software model engineering, wisme (pp. 262-271).

[Gr17] Grigaliunas, S., Toldinas, J., Venckauskas, A.: An Ontology-Based Transformation Model for the Digital Forensics Domain. Elektron. Ir Elektrotechnika 23, pp. 78-82. 2017

[Ha16] Harichandran, V.S., Walnycky, D., Baggili, I., Breitinger, F.: CuFA: A more formal definition for digital forensic artifacts. Digit. Investig. 18, p. 125-137. 2016.

[He21] Heller, P.: Probleme bei digitalen Ermittlungen - Wenn Forensik-Software zu Fehlurteilen führt. Deutschlandfunk. 2021. URL www.deutschlandfunk.de/probleme-bei-digitalen-ermittlungen-wenn-forensik-software-100.html (accessed 5.27.22).

[He05] Hesse, W.: Ontologie(n). 2005. URL gi.de/informatiklexikon/ontologien (accessed 5.24.22).

[Ho22] Honekamp, W., Povalej, R., Rittelmeier, H., Fähndrich, J., Berner, S., Labudde, D.: Technologiegetriebene Polizeiausbildung im Umgang mit Digitalen Spuren, in: Rüdiger, T.-G., Bayerl, P. S. (Eds.) Handbuch Cyberkriminologie. Springer. 2022.

[Ka09] Kahvedžić, D., Kechadi, T.: DIALOG: A framework for modeling, analysis and reuse of digital forensic knowledge. Digit. Investig. 6, pp. 23-33. 2009.

[Ka14] Karie, N.M., Venter, H.S.: Toward a General Ontology for Digital Forensic Disciplines. J. Forensic Sci. 59, pp. 1231-1241. 2014.

[Ke20] Kebande, V.R., Karie, N.M., Ikuesan, R.A., Venter, H.S.: Ontology-driven perspective of CFRaaS. WIREs Forensic Sci. 2, e1372. 2020.

[Mc20] McCrae, J. P.: The Linked Open Data Cloud. 2020. URL lod-cloud.net (accessed 5.27.22).

[Ol20] OliveiraJr, E., Zorzo, A.F., Neu, C.V.: Towards a conceptual model for promoting digital forensics experiments. Forensic Sci. Int. Digit. Investig. 35, 301014. 2020

[Po11] Poisel, R., Tjoa, S.: Forensische Untersuchung multimedialer Daten, in: 9. Sicherheitskonferenz Krems. Donau-Universität Krems, 2011.

[Po21] Povalej, R., Rittelmeier, H., Fähndrich, J., Berner, S., Honekamp, W., Labudde, D.: Die Enkel von Locard: Analyse digitaler Spuren in der forensischen Informatik. Inform. Spektrum 44, pp. 355–363. 2021

[Sa05] Sabou, M., Wroe, C., Goble, C., & Stuckenschmidt, H. (2005). Learning domain ontologies for semantic web service descriptions. Journal of Web Semantics, 3(4), 340-365.

[Sc06] Schmidt, D. C. (2006). Guest editor's introduction: Model-driven engineering. Computer, 39(02), 25-31.

[Si21] Sikos, L.F.: AI in digital forensics: Ontology engineering for cybercrime investigations. WIREs Forensic Sci. 3. 2021

[Si18] Siller, H.: Definition: Forensik. 2018. URL wirtschaftslexikon. gabler.de/definition/forensik-53390/version-276483 (accessed 5.27.22).

[So22] Solanke, A.A., Biasiotti, M.A.: Digital Forensics AI: Evaluating, Standardizing and Optimizing Digital Evidence Mining Techniques. KI - Künstl. Intell. 2022.

[So99] Sowa, J.F.: Knowledge representation: logical, philosophical, and computational foundations. Brooks/Cole, Pacific Grove. 1999.

[St98] Studer, R., Benjamins, V.R., Fensel, D.: Knowledge engineering: Principles and methods. Data Knowl. Eng. 25, pp. 161–197. 1998.

[Tr18] Trollmann, F., Fähndrich, J., & Albayrak, S. (2018, May). Hybrid adaptation policies: towards a framework for classification and modelling of different combinations of adaptation policies. In Proceedings of the 13th International Conference on Software Engineering for Adaptive and Self-Managing Systems (pp. 76-86).

[Vo19] von Elm, E., Schreiber, G., Haupt, C.C.: Methodische Anleitung für Scoping Reviews (JBI-Methodologie). Z. Evid. Fortbild. Qual. Gesundh. wesen 143. 2019.