2015

The Use of Ontologies in Forensic Analysis of Smartphone Content

Mohammed Alzaabi

Thomas Anthony Martin

Kamal Taha

Andy Jones
Edith Cowan University

THE USE OF ONTOLOGIES IN FORENSIC ANALYSIS OF SMARTPHONE CONTENT

Mohammed Alzaabi[1], Thomas Anthony Martin[1], Kamal Taha[1], and Andy Jones[2]

[1]Khalifa University of Science, Technology and Research, United Arab Emirates
[2]Edith Cowan University, Australia
{mohammed.alzaabi, thomas.martin, kamal.taha}@kustar.ac.ae
andy1.jones@btinternet.com

ABSTRACT

Digital forensics investigators face a constant challenge in keeping track with evolving technologies such as smartphones. Analyzing the contents of these devices to infer useful information is becoming more time consuming as the volume and complexity of data are increasing. Typically, such analysis is undertaken by a human, which makes it dependent on the experience of the investigator. To overcome such impediments, an automated technique can be utilized in order to aid the investigator to quickly and efficiently analyze the data. In this paper, we propose F-DOS; a set of ontologies that models the smartphone content for the purpose of forensic analysis. F-DOS can form a knowledge management component in a forensic analysis system. Its importance lies in its ability to encode the semantics of the smartphone content using concepts and their relationships that are modeled by F-DOS.

Keywords: digital forensics, forensic analysis, ontology

1. INTRODUCTION

Digital investigation involves the process of preserving, collecting, analyzing, and reporting evidence gathered from digital devices. Among these processes, the analysis of evidence has become the focus of a great deal of academic and industrial research. This is primarily due to the continuously evolving challenges that are associated with this process. The exponential growth of storage volumes is one of the main challenges that hinder digital investigations along with the inability of current forensic tools and methodologies to keep pace. In addition, analyzing content from these devices is becoming a more time consuming process as the complexity of data is increasing. That is, the wide range of formats, in which data is structured causes forensic investigators to spend much of their time understanding these structures instead of identifying relevant evidence.

We believe that the development of an automated system that can handle the above mentioned problems will lead to a more efficient analysis process for cases that involve large volumes of data. The system should be able to automatically identify items of evidence and their relationships. Such a technique not only helps investigators to identify relevant evidence, but also to shed light on

some hidden patterns and trends that can be difficult for a human to detect using manual analysis techniques.

Towards this goal, we propose an ontology-based analysis system to analyze content gathered from smartphones. Ontology is one of the major concepts used in Semantic Web applications, and it is used to model particular concepts in the real world. In our proposed approach, ontologies are used to model the environment of a smartphone through a set of concepts and their relationships. For instance, a Contact and a Message are two concepts in a smartphone which can have the relation 'hasSent' representing that a contact has sent a message. Such a representation allows a formal description of the smartphone content, hence allowing systems to correctly interpret different concepts involved in the modeled environment. Applying this representation to all of the extracted content will result in a solid, interconnected knowledge base that permits investigators to explore evidence objects and how they are related.

This paper focuses on an essential part of the system, namely the design of the ontologies. These ontologies have been developed for the purpose of forensic analysis. They are designed to be flexible to allow other researchers and forensic tool vendors to utilize them. Therefore, this paper aims at contributing to forensic science by presenting a set of ontologies, so-called Forensic-Driven Ontologies for Smartphones (F-DOS), that are specifically designed to be used by forensic analysis systems to conduct analysis on smartphones.

2. RELATED WORK

The idea of the Semantic Web was mainly introduced to have organized, integrated, and consistent Web content (Fensel et al., 2002). To achieve this, data models need to be built for different domains of interest which in turn built using what is called *Ontology*. Gruber (Gruber, 1995) defines an ontology as "an explicit specification of a conceptualization". A typical ontology consists of a finite number of terms (or vocabulary) and the relationships between them. These terms denote some concepts of a particular domain in the world. For instance, in a library management system, the domain that would be described here is the library, where concepts such as Book, Author, Publisher, and Borrower are terms relevant to that domain. These concepts are also called Classes. An example of a relationship between these concepts is: X-Publisher *published* X-Book; which indicates that a publisher with instance name "X-Publisher" has published a book with instance name "X-Book". Resource Description Framework Structure (RDFS) and Web Ontology Language(OWL) are two popular ontology languages.

One of the core technologies used in the Semantic Web is the Resource Description Framework (RDF). It provides a formal method to encode information about Web content in a graph-based data model. The syntactic construct of any RDF expression is what is called a triple. Each triple consists of three elements, namely subject, predicate, and object. The publisher example given above can be formed in RDF as in Listing 1.

Listing 1: Example of an RDF triple

```
@prefix library:<http://digitalLibraryabc.com/1#>.
library:X-Publisher
library:published library:X-Book.
```

The use of ontologies in digital investigation has been considered, but are still only seen limited application. Its employment has ranged from being used to manage the investigation processes, represent forensic artifacts, and to examine and analyze these artifacts. The author in (Luthfi, 2014) has

applied ontologies in a framework that can assist an investigator in the process of digital investigation by suggesting proper software and hardware systems that can be used in different stages of the process. With a similar approach, the authors in (Cosic, Cosic, & Baca, 2011) have defined an ontology to manage the chain of custody process. The ontology models aspects such as details related to the investigation process, information related to the acquired devices, and methods used to maintain the integrity of digital evidence. An ontology has been developed in (Park, Cho, & Kwon, 2009) that can categorize the various types of cyber crimes which can be used with data mining techniques.

With regards to works related to the representation and analysis of digital evidence, the authors in (Dosis, Homem, & Popov, 2013) have presented ontologies that model concepts related to computer storage media and network traffic data. The former is used to encode knowledge related to different types of storage media (such as hard disks, USB sticks, and SD cards) and file systems. The ontology defines a set of relationships to model how concepts such as File, File System, and Disk Partition can be linked together. For instance, a specific file is contained in a file system, and the file system in turn is contained in a disk partition. The network traffic ontology is used to model information such as IP addresses, TCP requests, and ports. Similarly, a system is proposed in (Kota, 2012) that utilizes an email ontology to represent concepts and relationships related to emails for forensics analysis. In (Ahmed, 2014), a set of ontologies were used to reason about network attacks and to reconstruct attack scenarios.

It is worth mentioning that the approach presented in (Dosis et al., 2013) is similar to our approach from the "representation of evidence" point of view. That is, in both approaches the ontology is used to model an environment under investigation (specifically a computer in Dosis's approach and smartphone in our approach). In both cases, the data is extracted from the intended device and modeled using the proposed set of ontologies. However, the ontologies defined in Dosis's approach are limited and model only a small part of a computer system. Thus, the approach did not show how concepts from different ontologies can be related. In addition, Dosis's approach has not described how the ontologies are designed, which is essential to ensure a consistent design of domain ontologies. Also, Dosis's approach lacks concepts related to the management of digital evidence such as case-related information, investigator's information, and the integrity of the constructed knowledge base. We cover all of these areas with F-DOS.

3. F-DOS DESCRIPTION

F-DOS contains a set of ontologies that formally model the smartphone content for the purpose of forensic investigation. Such a model supports a common and shared understanding of a domain of discourse among people and most essentially among software systems. This is achieved through a common vocabulary that is formally described using names for classes (concepts), attributes of classes, and relationships between classes. In addition, axioms are also defined to provide constraints on the interpretation of the model.

Before going into detail about how these ontologies are designed, it is vital to understand how F-DOS can be coupled with a forensic analysis system. F-DOS forms an abstract layer between the underlying resources that are extracted from a smartphone and the interface that is used by the

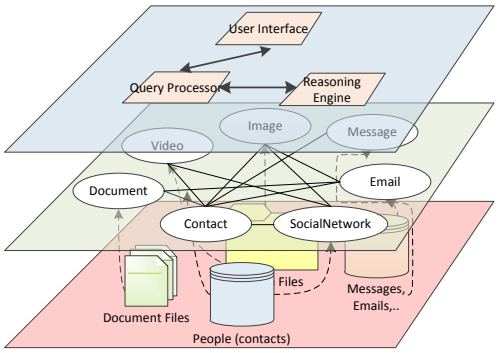


Figure 1: A simplified layered model which illustrates how F-DOS can be used in a forensic analysis system.

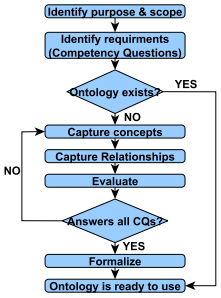


Figure 2: The methodology used to develop F-DOS.

investigator. Figure 1 illustrates this layered model. The F-DOS layer provides unified and transparent access to the underlying resources by hiding their various implementation details. Such a characteristic increases the interoperability of the system in the sense that the system can analyze content acquired from different smartphones with different operating systems. This feature is particularly important in forensic investigation when several devices with different operating systems are to be examined.

As depicted in Figure 1, the F-DOS layer proposes a set of vocabularies that generally represents classes and their relationships. Part of the system is to align the extracted contents with the appropriate classes and relationships that are modeled by the ontologies. This process results in a knowledge base, where the contents of the analyzed devices are transformed from being normal data that represents nothing but a series of bytes to objects (called instances), that represent concepts modeled by the ontologies. Such a representation allows the system, and possibly other systems that utilize the same ontologies, to correctly interpret the object and to understand how it is connected to other objects within or outside the same ontology. The User Interface layer provides the investigator with a way to query the generated knowledge base. A Reasoning Engine can also be involved to infer new knowledge.

3.1 Ontology Development Methodology

To ensure a proper and consistent modeling of the ontologies, we adopt a modified version of an ontology development methodology called Methontology (Fernndez-Lpez, Gmez-Prez, & Juristo, 1996). Although this methodology was proposed some time ago, it is still considered to be one of the most comprehensive and detailed methodologies (Fernndez-Lpez & Gmez-Prez, 2002). This is primary due to the clear description of the various aspects of the ontology's life cycle as well as its independence to any specific application. Methontology is intended to be used for large-scale, enterprise ontologies, thus, a simplified version is used. The methodology is illustrated in Figure 2.

The process depicted in Figure 2 is used to develop all ontologies in F-DOS. It starts by identifying the general purpose and scope of each ontology. This is essential to limit the scope of the modeled domain and to assess in the identification of concepts in a later stage of the process. The second step is to identify the requirements of the ontology. That is to identify what problems/questions the ontol-

ogy is supposed to solve/answer. As with Methontology, this is performed by defining a set of questions the ontology should be able to answer. The question in this case is called a *Competency Question (CQ)*. The importance of CQ lies in its usage to evaluate the coverage of the ontology (where we use coverage to mean the modeled concepts and relationships are adequate to answer the intended questions).

In the third step of the process, we investigate whether any already available ontologies can be used. Reusing ontologies enforces one of the major objectives of the Semantic Web which is sharing knowledge among different software systems. The fourth and fifth steps are to capture concepts and their relationships, respectively. This is achieved by analyzing the domain while taking into consideration the scope of the ontology and CQs. There are a number of methods that can be used to capture the concepts, namely: Top-Down, Bottom-Up, and Middle-Out (Fernndez-Lpez & Gmez-Prez, 2002). With the Top-Down method, the most abstract concepts are obtained first, then one moves to more concrete concepts. By contrast, the Bottom-Up method moves from identifying the most concrete concepts to the most abstract. The third method, Middle-Out, works by obtaining relevant concepts, then moving towards concrete and abstract concepts. In our methodology, the Middle-Out method is used (which is also recommended by Methontology).

The sixth step is to evaluate the ontology. As stated earlier, CQs are used in this evaluation by examining the coverage of the ontology. If the ontology fails to answer any of the questions in CQs, the ontology is edited again and the process goes back to identifying new concepts. Once the ontology passes the evaluation, it is formalized using one of the ontology languages such as RDFS and OWL.

3.2 F-DOS Design

F-DOS is designed in such a way that it can easily integrate further domains. This is performed by dividing the ontologies into domain ontologies and an upper ontology. In this section, the design of F-DOS is discussed more deeply.

3.2.1 The Upper (Core) Ontology

The Upper Ontology represents a wide range of the concepts modeled in the system. It is the main ontology and we will refer to it as the Core Ontology. Since the modeled data will be used to conduct forensic investigations, all data objects that are extracted from a smartphone are interpreted as evidence. Part of the Core Ontology is depicted in Figure 3, which shows the relation between three main concepts (also called classes), namely *DataObject*, *EvidenceElement*, and *DataSource*. A *DataObject* is used to represent a native entity which is linked to a sequence of bytes, such as a document file, image file, or a database. The *EvidenceElement*, on the other hand, is used to represent a set of conceptual objects that are derived from a *DataObject*. A *DataObject* may have one or more *EvidenceElement*. For instance, given a database of calendar events, values in a table that correspond to a specific calendar event, is interpreted as an instance of the class *DataObject*, while each DataObject in turn is linked to a *EvidenceElement* of type *CalendarEvent* (*CalendarEvent* is a class modeled in the Calendar domain ontology). Each *EvidenceElement* is eventually mapped to a concept modeled in a domain ontology (and in this case is *CalendarEvent*). Therefore, the *DataObject* class is linked directly to physical objects, while the *EvidenceElement* class is linked to conceptual objects.

Structuring the data based on *DataObject* and *EvidenceElement* classes allows for

a more flexible interpretation between physical and conceptual objects. This means that (depending on the level of granularity of the data) a document file can be directly mapped to the class *DocumentFile*, or the content of the document can be further processed so that more data is extracted, such as persons, organizations, and locations, and then linked to their corresponding conceptual classes.

The Core Ontology also encodes the source of where an instance of the *DataObject* class is located through the *DataSource* class. The absolute path of a file can be used to indicate the source of a *DataObject* instance. However, in case of a deeper indexing technique, the offset of the *DataObject* instance's value can be used to indicate its source location. Such indexing techniques are not detailed in this paper. This information is essential to ensure the traceability of data when needed by the investigator. Looking at the raw data, for instance, is a common practice by many forensic analysts to look for any possible deleted data. Another role of the Core Ontology is to ensure that the domain ontologies are interconnected. This assembles the entire picture of the smartphone model. Thus, relations between classes from different domain ontologies are created at the Core Ontology level.

3.2.2 Domain Ontologies

A domain ontology is used to model a specific domain of discourse. All domain ontologies are linked to the Core Ontology via the *EvidenceElement* class. The following is a description of some of the domain ontologies:

- **The Contact Ontology**: Contact information plays an important role in finding links between people, which is one of the major tasks in any forensic analysis.

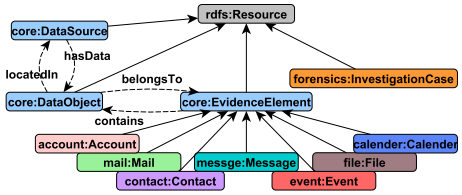


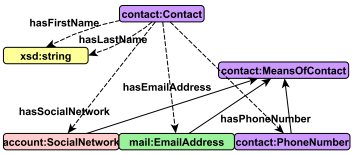
Figure 3: The Core Ontology which is the main ontology in F-DOS. The Resource class is the mother of all classes. The "subClassOf" relationship is shown as a solid line, while other relationships are shown as dashed lines. Classes are color coded based on the domain ontology that they belong to.

As depicted in Figure 4a, the Contact Ontology has the *ContactResource* as the mother class. It has two subclasses which are namely the *Entity* class and *MeansOfContact* class. The *Entity* class models two concepts that are either a person or an organization through the *Person* and *Organizations* classes respectively. In this ontology, a person is distinguished by the presence of contact information which forms two subclasses: the *ContactablePerson* class and *UncontactablePerson* class.

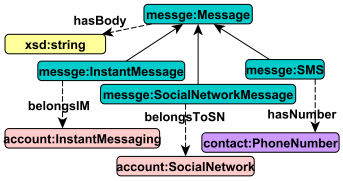
The *MeansOfContact* class models ways to communicate with a person. The five main communication means are: phone number, email address, social network account, instant messaging account, and websites such blogs. Each instance of the class *ContactablePerson* should have at least one relation with one of the subclasses of the *MeansOfContact* class.

As indicated in Section 3.1, CQs are used to evaluate the ontology. Examples of CQs for the Contact Ontology are as follows ('X' indicates an input):

- What communication means a contact with the first name 'X'



(a) The Contact Ontology



(b) The Message Ontology

Figure 4: Illustration of part of the Contact and Message Ontologies. Some labels are shortened due to space limit.

has?

– What contacts have a phone number AND social network account AND Email Address?

- **The Message Ontology**: this models the message concept as illustrated in Figure 4b. In this context, a message is a unit of communication between one or more parties. The ontology models three types of messages which are: SMS, social network message, and instant message. The following are examples of CQs for the Message Ontology:
  - What messages were exchanged between two contacts that have phone number 'X' and 'Y'?
  - What messages belong to the social network 'Z'?
- **The Investigation Case Ontology**: it is responsible for maintaining information about the case under investigation. It records information such as the investigator's name, case name, and the date of investigation. Each EvidenceElement (such as a file, email, phone number, etc.) instance is linked to the InvestigationCase class's instance. Another important piece of information encoded by the Investigation Case Ontology is the hash of the knowledge base. This provides an integrity checking mechanism to detect any attempted tampering of the knowledge base.
  - What is the name of the investigator who investigated the case named 'X'?
- **Other Domain Ontologies**: other domain ontologies are also used to model other aspects of a smartphone. This includes the Email Ontology, Calendar Ontology, Location Ontology, File Ontology, and Event Ontology. For instance, the Event Ontology models objects that are associated with date and time such as calendar events, call logs, message sending and receiving time and so on.

4. QUERYING THE KNOWLEDGE BASE AND REASONING

The result of modeling the content extracted from smartphones based on F-DOS is a knowledge base. This knowledge base can be queried by an investigator to gather information related to the case under investigation. The complexity of these queries can vary from simple queries within a single domain ontology to more complex ones that involve more than one domain ontology. A standard query language called SPARQL is used to query the knowledge base. The following is an example of a SPARQL query to answer the first competency question of the Contact Ontology (refer to Section 3.2.2):

Listing 2: Example of a SPARQL query to answer a CQ from the Contact ontology.

```
SELECT ?contact ?x
WHERE {
        ?contact ?relation ?x.
        ?x a contact:MeansOfContact.
        ?contact contact:hasFirstName ?name.
        FILTER regex(?name, "X", "i")
}
```

Reasoning can also be applied in order to infer new information from existing facts in the knowledge base using a set of rules. We distinguish two types of rules. The first is called Environmental Rule (ER) which is used to infer new knowledge about the modeled environment (which in this case is the smartphone). The second type is called Forensic Analytic Rule (FAR) which is used to infer new knowledge that can assist the investigator in the forensic analysis. An example of an ER is to ensure that a Contact must have a means of contact, which can be represented as in Listing 3 (this is a human readable syntax of OWL):

Listing 3: Example of an ER applied to the Contact Ontology.

```
Person(?x)∧ MeansOfContact(?y) ∧
hasMeansOfContact(?x, ?y) ⇒ Contact(?x)
```

Listing 4 shows an example of a FAR which is used to mark all messages of a suspect contact as suspicious using the class *Suspicious*:

Listing 4: Example of a FAR to infer suspicious messages.

```
Contact(?x) ∧ Message(?y) ∧
hasSentMessage(?x, ?y) ∧ Suspicious(?x)
⇒ Suspicious(?y)
```

5. CONCLUSION

In this paper we proposed F-DOS, a set of ontologies that model smartphone content for the purpose of forensic analysis. Although F-DOS describes aspects related to the knowledge management only, it can play an essential role in forensic analysis tools. This role is characterized by its ability to encode the semantics of data using classes and relationships which model certain domains. The benefits of this encoding with respect to forensic analysis are (1) a unified representation of evidence, (2) the ability to explore elements of evidence and how they are interconnected, and (3) the ability to perform reasoning to infer new implicit knowledge from explicit ones.

REFERENCES

Ahmed, S. S. M. (2014). *Intrusion Alert Analysis Framework Using Semantic Correlation* (Unpublished doctoral dissertation). University of Victoria.

Cosic, J., Cosic, Z., & Baca, M. (2011). An Ontological Approach to Study and Manage Digital Chain of Custody of Digital Evidence. *Journal of Information and Organizational Sciences*, *35*(1), 1–13.

Dosis, S., Homem, I., & Popov, O. (2013). Semantic Representation and Integration of Digital Evidence. *Procedia Computer Science*, *22*, 1266–1275.

Fensel, D., Bussler, C., Ding, Y., Kartseva, V., Klein, M., Korotkiy, M., ... Siebes, R. (2002, June). Semantic Web Application Areas. In *the 7th International Workshop on Applications of Natural Language to Information Systems.* Stockholm, Sweden.

Fernndez-Lpez, M., & Gmez-Prez, A. (2002, June). Overview and Analysis of Methodologies for Building Ontologies. *The Knowledge Engineering Review*, *17*(2), 129–156.

Fernndez-Lpez, M., Gmez-Prez, A., & Juristo, N. (1996). Methontology: from ontological art towards ontological engineering. In *ECAI96 Workshop on Ontological Engineering* (pp. 41–51). Budapest.

Gruber, T. R. (1995, November). Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies*, *43*(56), 907–928.

Kota, V. K. (2012, December). An Ontological Approach for Digital Evidence Search. *International Journal of Scientific and Research Publications*, *2*(12), 1–5.

Luthfi, A. (2014). The Use of Ontology Framework for Automation Digital Forensics Investigation. *International Journal of Computer, Control, Quantum and Information Engineering*, *8*(3), 423–425.

Park, H., Cho, S., & Kwon, H.-C. (2009). Cyber Forensics Ontology for Cyber Criminal Investigation. In M. Sorell (Ed.), *Forensics in Telecommunications, Information and Multimedia* (pp. 160–165).

