Jun 12th, 8:45 AM

An Ontology-Based Forensic Analysis Tool

Mohammed Alzaabi
Khalifa University of Science, Technology and Research, United Arab Emirates,
mohammed.alzaabi@kustar.ac.ae

Andy Jones
Khalifa University of Science, Technology and Research, United Arab Emirates, Edith Cowan University, Perth, Australia, andrew.jones@kustar.ac.ae

Thomas A. Martin
Khalifa University of Science, Technology and Research, United Arab Emirates,
thomas.martin@kustar.ac.ae

AN ONTOLOGY-BASED FORENSIC ANALYSIS TOOL

Mohammed Alzaabi[1] (mohammed.alzaabi@kustar.ac.ae)
Andy Jones[1,2] (andrew.jones@kustar.ac.ae)
Thomas Anthony Martin[1] (thomas.martin@kustar.ac.ae)

[1]Khalifa University of Science, Technology and Research, United Arab Emirates
[2]Edith Cowan University, Perth, Australia
Phone: 9 716 597-8888
Fax: 9 716 561-1789

ABSTRACT

The analysis of forensic investigation results has generally been identified as the most complex phase of a digital forensic investigation. This phase becomes more complicated and time consuming as the storage capacity of digital devices is increasing, while at the same time the prices of those devices are decreasing. Although there are some tools and techniques that assist the investigator in the analysis of digital evidence, they do not adequately address some of the serious challenges, particularly with the time and effort required to conduct such tasks. In this paper, we consider the use of semantic web technologies and in particular the ontologies, to assist the investigator in analyzing digital evidence. A novel ontology-based framework is proposed for forensic analysis tools, which we believe has the potential to influence the development of such tools. The framework utilizes a set of ontologies to model the environment under investigation. The evidence extracted from the environment is initially annotated using the Resource Description Framework (RDF). The evidence is then merged from various sources to identify new and implicit information with the help of inference engines and classification mechanisms. In addition, we present the ongoing development of a forensic analysis tool to analyze content retrieved from Android smart phones. For this purpose, several ontologies have been created to model some concepts of the smart phone environment.

Keywords: digital forensic investigation, digital forensic analysis tool, semantic web, ontology, android

1. INTRODUCTION

As a result of the ongoing trends towards larger storage capacities for digital devices, the digital forensics domain is facing a number of serious challenges with the increased time and effort required to analyze data from these devices. Such trends have influenced the process of identifying relevant traces which is usually surrounded by a vast volume of irrelevant traces. Additionally, the complexity of data formats and their diversity have made the investigator spend much of the time in understanding the structure of the data rather than locating relevant evidence.

The existing forensic tools, which have been referred to as First Generation (FG) forensic tools by Daniel Ayers (Ayers, 2009), have shown a number of limitations in addressing the increasing complexity and volumes of data. Popular FG forensic tools such as EnCase and FTK tend to have a common architecture that is based on the same conceptual model for searching and presenting information.

Simson L. Garfinkel (Garfinkel, 2010) describes this type of model as a "Visibility, Filter, and Report" model. In this model, the retrieved traces are made *visible* to the investigator through a number of cascaded windows in the forensic tool. Derived traces to be analyzed are presented in a tree structure with a root object from which access to other data objects is attainable (Garfinkel, 2010). In the case of the analysis of a hard drive, the root of the tree structure may be a partition table of the drive which may lead to other directories and data objects that builds the file system stored in that drive. Individual

objects can also be viewed in a window. In order to reduce the number of displayed objects, the investigator can *filter* the results. Keyword search is currently dominating the techniques used to analyze the data (Louis and Engelbrecht, 2011). Finally, a *report* is auto-generated by the tool summarizing what was found. Instead of generating a comprehensive report of all of the collected data, most of the tools allow the examiner to choose what traces are to be included in the report.

FG forensic tools have a number of limitations as far as analyzing traces is concerned. Firstly, the investigator is usually presented with a familiar GUI interface that depicts a tree structure of the claimed files and directories from an image. However, since the volume of traces is increasing, the tree structure is also increasingly complex, which results in the investigator being overwhelmed with a large number of files and directories to be examined. Consequently, identifying relevant evidence becomes a more complex task. Cross analysis between different types of data sources is also becoming even more complicated as the sources to be analyzed are increasingly diverse.

Secondly, in order to identify potential evidence, examiners tend to use the advanced search techniques which are provided by most forensic tools. The examiner uses his/her experience and the available background information about the case to choose potentially suitable terms by which relevant traces can be revealed or more terms can be identified for further examination. Due to the dependence on the investigator's experience and the availability of background information about the case, the selected search terms will only be as good as these two factors. Where these two factors are not sufficient, searching for potential evidence can be inefficient.

With these limitations, the need for a more automated forensic analysis tool becomes apparent. In light of this, we believe that the use of an ontology can greatly assist the development of an automated analysis tool. This tool may allow the examiner to handle the problem of the large volumes and the complexity of data more efficiently. In this paper, a novel framework is proposed for a potential next generation forensic tool. The framework is based on semantic web technologies where ontologies are used to model the environment under examination. This model encompasses resources and their relations in a graph-based dataset by utilizing the Resource Description Framework (RDF). With this dataset, which acts as a network of data, a solid, interconnected knowledge base of the different evidence objects extracted from the device may be provided. However, to the extent that the semantic network and ontology act as assumptions, they may also lead the examination astray and the examiner using these methods must understand both their nature and their limitations in order to avoid errors and omissions in their results.

2. SEMANTIC WEB

Due to the openness notion of the World Wide Web (WWW), contributing to its web of information is relatively easy and effortless. This has led to a massive increase in the amount of web content and has introduced a number of serious shortcomings in the underlying Hypertext paradigm of the Web. This paradigm simply defines the structure of the Web and allows users to access, connect, and share information over the internet. One of these shortcomings is the failure in organizing the vast amount of web content in a logical manner which can make searching and locating specific information a difficult task.

The idea of the semantic web was mainly introduced to limit this problem and to have more organized, integrated, and consistent web content (Fensel et al., 2002). Based on the underlying notion of giving a well-defined meaning for each information item (such as, text, picture, or video) in the Web, these information items cannot only be understood by humans, but also by machines. For instance, once the machine can understand what a person, place, and event is, it can help the user to store that event automatically to the calendar. In the semantic web, things that exist in the world or can be described using the associated methodology are known as Resources (Allemang and Hendler, 2011), and the relationships between these resources are known as Relations. These two concepts form the basis of

any semantic web document. The strength of the semantic web lies in its ability to explicate the relationship between multiple resources, even when the resources come from more than one source, allowing them to be easily integrated. For instance, in a library management system, the librarian may need to integrate data from multiple publishers. Presenting any authorship relation between the author (i.e. the writer) and the book using the relation *hasAuthor* will allow the system to interpret this information for all publishers dataset, even without human intervention.

However, just introducing resources and relations to the content of the web will not solve the problem and may result in a more chaotic dataset, specifically if they have not been properly modeled. Modeling the environment is a key step towards a semantically well-structured document. Models, as described in (Allemang and Hendler, 2011), permit people to collaborate and organize the information that they would want to share and provide a consolidated understanding of the environment. The semantic web standards have introduced some modeling languages that have become the de facto standards for creating semantic web documents, such as RDFS and OWL. A model in the semantic web is known as an Ontology. Pioneers in this field have gone through many attempts in order to define what an ontology is. Probably, the most well-known definition is by Gruber (Gruber 1993) who defines an ontology as "an explicit specification of a conceptualization".

Authors such as Stevens et al. (2000), Mika (2005), and Guizzardi (2007) have explained the term conceptualization from Gruber's definition as identifying relevant concepts (or resources), the relations between them, and any constraints that they hold between them in a domain of discourse. In other words, a typical ontology consists of a finite number of terms (or vocabulary) and the relations between them. These terms denote some concept of a particular domain in the world. For instance, and considering the library management system example, the domain that is 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. Subclasses can also be derived from main classes to form a hierarchy of classes. For example, Science Book and Philosophy Book are two possible subclasses of the class Book.

Relations in an ontology represent associations between classes. Robert Stevens et al. (Stevens et al., 2000) have categorized ontology relations into two broad types: (1) Taxonomies, which allows for building a tree structure of classes/concepts in the ontology. A great example of this type is the inheritance relation, also known as "is a kind of" relation. For instance, a Science Book is a kind of Book. (2) Associative, which relates classes/concepts across the entire ontology. Examples of such a relation are many; however, one is the Nominative relation that describes the name of a concept - Publisher *hasName* publisherName.

One of the core technologies used in the semantic web is the Resource Description Framework (RDF). RDF is a World Wide Web Consortium (W3C) standard framework used to represent data on the Web (Klyne et al., 2004). It provides a formal method to encode information about Web resources 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. A triple describes a binary relation and can be represented by two nodes (subject and object), connected through an edge (predicate). For instance, we can describe the relation between one book and its author as the following triple:

```
PracticalRDF hasAuthor "Shelley Powers"
```

where PracticalRDF is the subject, Shelley Powers is the Object, and they are connected by the predicate *hasAuthor*. It is important to mention that RDF is utilizing the Uniform Resource Identifier (URI) for the purpose of naming the three elements of the triple. For example, the predicate *hasAuthor* is denoted as *http://digitalLibrary.com/hasAuthor*. This will facilitate the ability to uniquely identify the resources, so no two different resources can have the same name. Therefore, integrating information from non-local resources will not lead to naming conflicts (Horrocks ,2008). Since using

URIs may result in a longer name, it is always preferable to use prefixes to shorten the name of resources. Given the triple above, it can be represented as follows:

```
@prefix library: <http://digitalLibrary.com>
library: PracticalRDF
library: hasAuthor  "Shelley Powers".
```

Generally, RDF triples can be serialized in multiple formats. The previous triple and the ones used in this paper are represented in the Turtle serialization. Other serializations are XML, N-Triples, and Notation 3.

3. THE SEMANTIC WEB AND DIGITAL FORENSICS

Semantic web technologies have not only influenced the way that web applications are developed, but have also paved the way for new contributions in various sectors. One of these sectors, and perhaps the most well-known, is the biomedical sector. Doctors and researchers in this field are constantly required to take well-informed decisions regarding different diseases and symptoms. As such, heterogeneous data sources are required to integrate data from various topics such as cells, drugs, and proteins. Other sectors are knowledge management in large enterprises (Fensel et al., 2002) and software engineering (Happel and Seedorf, 2006).

Semantic web technologies have also been seen in the digital forensics sector; however, published works in this particular area are still sparse.

DIALOG is an ontology which has been developed by Damir and Tahar (Kahvedžić and Kechadi 2009), for the purpose of describing forensic investigation results. It defines a vocabulary of the main concepts involved in the forensic investigation domain. The ontology models four main dimensions related to forensic investigation; which are Crime Case, Evidence Location, Information, and Forensic Resource. Each of these dimensions is responsible for modeling a particular concept in a forensic investigation. For instance, the Crime Case dimension models concepts which are related to the type of crime. The taxonomy of the Crime Case defines the Cyber Crime Case and the Non Cyber Crime Case as the most general concepts to differentiate between types of crime cases. Cyber Crime Case in turn is also divided to involve high-tech crimes such as fraud and software piracy.

Turner (Turner, 2005) has introduced a new concept for data acquisition and representation called Digital Evidence Bag (DEB). DEB is an abstracted model that permits investigators to acquire data from multiple sources of evidence. Evidence sources are organized in a conceptual structure called *Evidence Bag*. Each evidence bag contains a set of files where information about the case (such as investigator information, evidence acquisition process, list of evidence contained) and the evidence source (such as list of files and directories of the image and their metadata) are stored. In addition, DEB emphasizes the issue of data integrity by including a hash signature for the evidence bag to maintain its provenance.

One year after proposing DEB, Schatz and Clark (Raghavan et al., 2009) extended this work and introduced the Sealed Digital Evidence Bag (SDEB). SDEB has provided a new representation approach where metadata and evidence information are integrated using a pre-defined domain ontology of the context of the case. SDEB uses RDF to annotate evidence related metadata and uses ontology to describe their vocabulary. For instance, an image file is annotated by imposing its metadata and is linked to the Image File concept in the ontology. This process is done recursively for all evidence resources and the obtained information is stored in the evidence bag.

Damir and Tahar (Kahvedžić and Kechadi, 2011) also proposed an ontology-based approach for representing evidence from forensic investigations. The authors refer to this approach as an evidence management methodology for encoding the semantic information of evidence. This approach aims to

assist the investigator in report writing, evidence communication, and to relieve the investigator from the task of manually describing the evidence. Part of the methodology is to involve the investigator in annotating the evidence. To illustrate this, the authors gave an example of a file that contains some text and an image. The text is referring to an individual and the image is also displaying an individual at a certain location. From the ontology point of view, the file, text, image, individual, and the location are all resources. Therefore, in this case, the investigator can link the text to the individual resource it describes, and link the image to the individual resource as well as to the location resource it captures.

It can be seen that all of the above research clearly has the emphasis on representing and managing results obtained from a digital forensic investigation. For instance, the taxonomy of the DIALOG ontology is developed to hold information about the committed crime cases (through the Crime Case dimension) and the forensic resources that were used to conduct the investigation (through the Forensic Resource dimension). Such information is crucial for evidence management and reporting tasks. The emphasis of our ongoing research, however, is towards the analysis of digital forensic evidence. Hence, utilizing the DIALOG ontology will not satisfy our needs.

The works proposed in (Raghavan et al., 2009) and (Kahvedžić and Kechadi, 2011) are similar to the annotation process performed in our framework. Nevertheless, the same reason why DIALOG cannot be applied in our framework is also valid for both proposals. They model the evidence for management purposes, but not for analysis purposes, which is the aim of our research. The framework presented in this paper not only annotates data, but also tries to correlate evidence from various sources and identify new, implicit information. This can be achieved by the use of inference engines and classification mechanisms.

4. FFRAMEWORK FOR ONTOLOGY-BASED FORENSIC ANALYSIS TOOLS

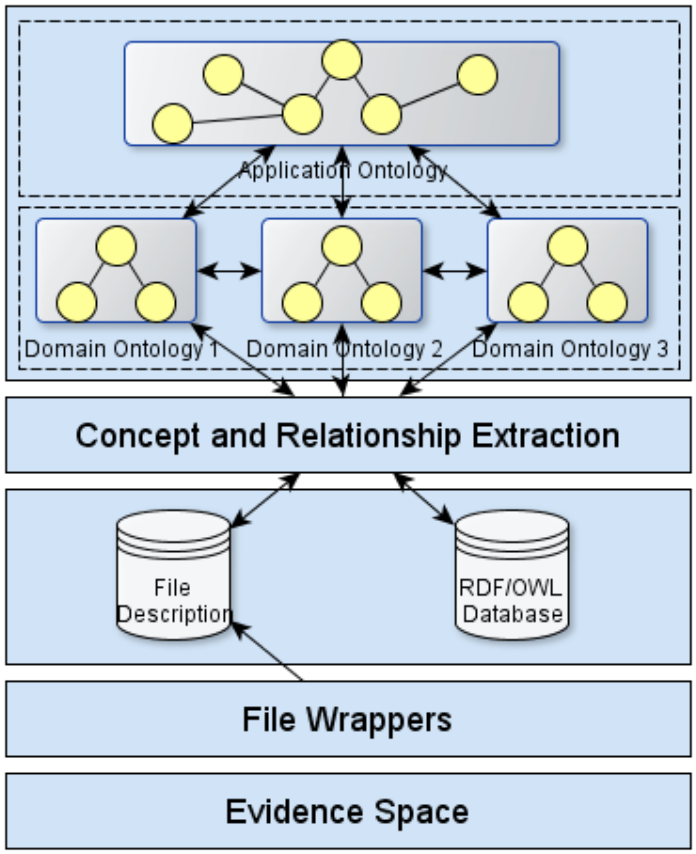

Figure 1 Ontology-based framework for forensic analysis tools.

The proposed framework, shown in Figure 1, is a layered ontology-based framework that follows the principle of superimposing the metadata of existing evidence resources. This metadata is used by the Concept and Relationship Extraction layer where various concepts and the interrelationship between them are extracted and maintained. This extraction is conducted with the help of a structured representation of particular domains of interest which are referred to as Domain Ontologies. The encoded knowledge (which is represented in the concepts and their relationships) is maintained in an RDF/OWL database allowing forensic tools to interact with a centralized database that facilitates the organization, access, and reuse of potential evidence objects. The framework consists of five main layers, namely the Evidence Space, File Wrappers, File Description and RDF/OWL databases, Concept and Relationship Extraction, and Domain and Application Ontologies.

4.2 Evidence Space

The evidence space is where potential evidence objects such as files purported to be documents, images, videos, and databases are located. In forensic terms, this space can be a forensic image of a device under examination. It may contain potential evidence objects that are structured such as databases (which conform to relational models), semi-structured such as XML files (which may not conform to any rational model but are organized by tags), or less structured such as videos and images.

4.3 File Wrappers

The file wrapper is simply a program that extracts descriptive information about various types of files from the evidence space. A main source of this information is the metadata of a file which may

provide property-value pairs that may improve the process of relevant content retrieval. Examples of such pairs are: file size, date, and MIME type. In this framework, these file wrappers are tightly-coupled with the structure of the file system of the device. Thus, different devices may require different file wrappers to work with. This layer is the only layer that contains device-dependent components. Another primary purpose of the wrappers is to ensure a consolidated form for retaining the extracted data. This is particularly useful for managing and accessing the information.

4.4 File Description and RDF/OWL Database

This layer mainly holds two core databases; the File Description and RDF/OWL databases. The former is used by the file wrappers in order to store the extracted descriptive information of files from the Evidence Space. The latter, which is the RDF/OWL database, is used by the Concept and Relationship Extraction layer to store the entire knowledge base of the potential evidence under examination. The structure of this knowledge base is governed by the upper ontology layers. The format used in these databases is the Resource Description Framework (RDF).

4.5 Concept and Relationship Extraction

The major purpose of this layer is firstly to extract concepts from the File Description database and determine to which class this concept belongs to based on the upper ontology layers. Examples of such concepts that can be obtained from a smart phone are: a contact which belongs to the Contact class, an image which belongs to the Media class, and a Word document file which belongs to the Document class. By utilizing the metadata of these concepts, relationships with other concepts (and hence classes) can be maintained. For instance, each message instance may hold the sender information such as the name and the phone number in its metadata. This information can be linked to a contact instance from the Contact class given that the two instances have the same phone number.

4.6 Domain and Application Ontologies

These two layers collaboratively form an ontological model for a particular environment. This model consists of concepts (or classes) and the relationship among them. A domain ontology formally models the concepts and their relations in a particular domain, rather than modeling generic subjects of the world. For instance, in a smart phone environment, Message, Person, Email, and Event can be considered as individual domain ontologies.

Dividing the whole ontology into more domain ontologies strengthens the structure of the model in various aspects. Firstly, since the domain ontology concentrates on a specific concept, designing the ontology should become less complex and result in a more homogenous concepts and relations. Secondly, any modification to the domain ontology will be more manageable and less time consuming. Thirdly, adding a new concept to the application ontology becomes a straightforward process as domain ontologies can be inserted as “add-ons”.

Defining domain ontologies on its own is not sufficient to model the entire environment of something such as a smart phone; therefore, the Application Ontology is used to assemble the entire picture of the model. The Application Ontology introduces a new layer of interconnected domain ontologies. This is done by adding high level relations between domain ontologies. For instance, a relation called *hasSent* can be used between the Contact class and the Message class (which means that a contact has sent a message) as a high level relation. This relation forms an abstract view of the relations between the Contact class and the Message class. More detailed relations can be derived from this abstract relation such as: *hasSentSMS*, *hasSentImMessage*, or *hasSentEmail*.

5. SEMANTIC ORGANIZATION OF THE EVIDENCE SPACE

5.1 Annotation

By considering the evidence space as the primary source for classes instance data, the evidence space should be processed in a way that allows machines to understand its content. This is directly linked to the purpose of the semantic annotation. The annotation process, which is performed by the File Wrapper, is typically achieved by annotating the resources from the evidence space in a machine-understandable manner using metadata. In other words, the annotator creates or enhances the metadata of a file allowing that file to be processed more effectively by machines. This step facilitates a formal description as well as a new access method to the resources available from the evidence space. For instance, considering a trace assumed to be an email, the trace's metadata generated by the automated process and associated with this trace may be the email address of the sender, recipients, title, and the received date.

In the proposed framework, the annotation process is taking place in the File Wrapper. The extracted resources from the File Wrapper will be annotated by imposing their metadata. The annotated data will then be stored in the File Description database. This database stores the information in the RDF format.

5.2 Association

Resources from the evidence space may have relations created between them by the automated process, representing some asserted association between the different entities they represent. The process of asserting these associations has many potential benefits to an investigator, as they may help the investigator to understand the overall picture of the environment under examination. The links established between resources may lead to a more probable source of evidence. Thus, the automatic discovery of such associations becomes a key feature for the next generation forensic tools allowing the investigator's time to be organized more efficiently.

By providing a meaningful understanding of the association between resources, the investigator will not only be able to place these resources in the context of investigation, but will also be exposed to other evidence objects that were not immediately obvious. For instance, in a situation where the investigator would like to explore more about a particular resource of type Person, the association with other resources that have relations with it can be identified. Figure 2 depicts these relations with other resources of the same type (i.e., Person).

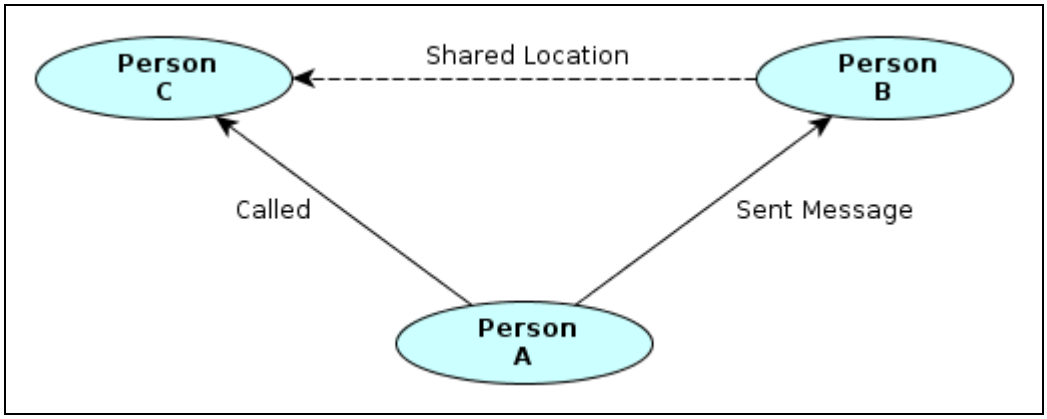

Figure 2 Association representation between resources

If Person A is being investigated, two associations with Person B and Person C may be asserted by the examination process. The asserted association between Person A and Person B is represented by a relation where Person A may have sent a message to Person B, while the asserted association between

Person A and Person C is represented by a relation where Person A may have called Person C. Although these two asserted associations can be sufficient to answer the investigator enquiries, further evidence may also be considered which may lead to more potential evidence. In this example, asserting that Person B has shared a location with Person C may open further questions to the investigation.

Two levels of asserted association are involved in this framework. The first level maintains asserted relationships among the resources themselves, a so-called Core Knowledge Association. On the other hand, the second level is called Indexing Association. This asserted derived association level maintains asserted derived relationships between the physical files from the evidence space and the identified resources.

5.2.1 Core Knowledge Association

In this level of asserted association, the asserted relationship between resources (or concepts in the ontology) is asserted. These associations represent the core knowledge of the environment under examination by defining how concepts are turned into interconnection in the ontology. Therefore, three types of asserted association may be distinguished in this level; asserted association among classes of a single domain ontology, asserted association among more than one domain ontology, and asserted association of the application ontology. Figure 3 shows these types of asserted associations.

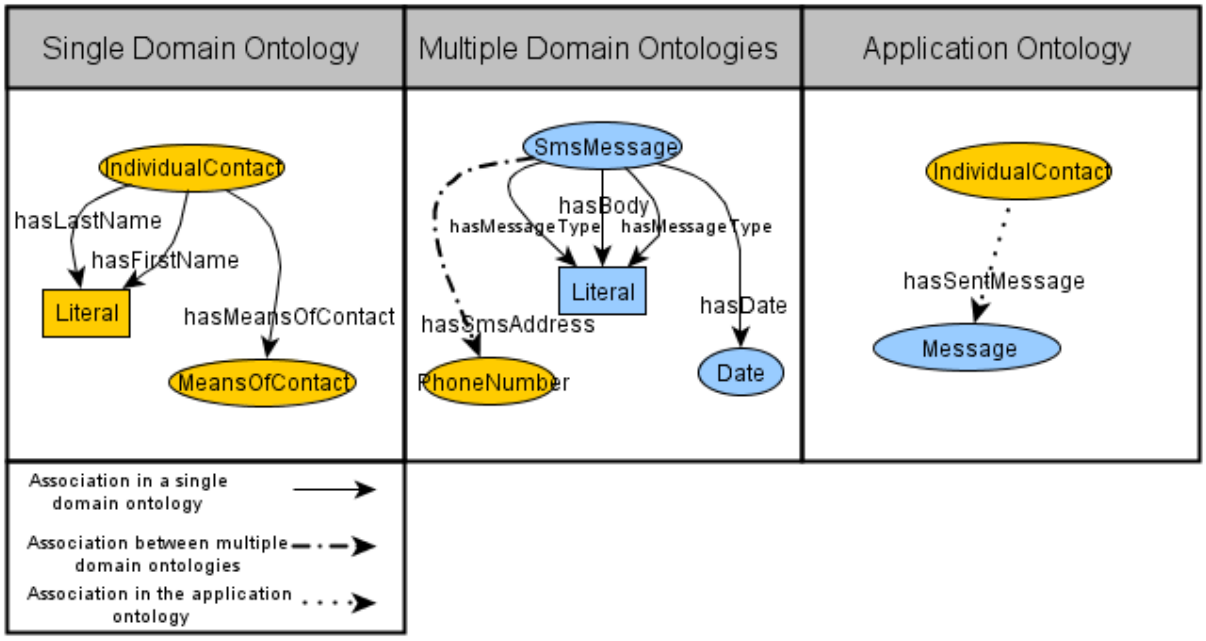

Figure 3 Illustration of the different types of Core Knowledge asserted associations based on how classes are connected between different ontology levels

The previous classification is mainly based on how classes are asserted to be connected between different ontology levels. However, another way to classify the Core Knowledge asserted associations is based on the nature of the asserted relationship that each asserted association type represents. These types are: Instantiation, Inheritance, and Property.

The instantiation is used to show an asserted instance of a class. This relation is usually represented by the RDF property *type*. For instance, <#Report1, rdf:type, #Document> indicates that the resource Report1 is an instance of the Document class. The inheritance, on the other hand, facilitates a relation between two resources where one of them is a subtype of the other. This relation becomes handy

particularly when a subclass of a main class is to be illustrated. For instance, the inheritance relation would be used to show that a class called SmsMessage is a subclass of the Message class. The property relation is defined by the user to express an environment-specific asserted relationship. For example, the asserted relation *SentMessage* from Figure 2 expresses a specific relation where Person A has sent a message to Person B based on the assertions of the relationships.

5.2.2 Indexing Association

Being able to express the asserted knowledge retrieved from the evidence space, through the previously discussed asserted relations, is almost certainly not adequate to perform a full examination. In some circumstances, the examiner may wish to associate the physical location of a particular asserted file to its corresponding resource in the knowledge base. Thus, the second level of asserted association relates each discovered resource to its location in the evidence space. For example, if the examiner comes across an asserted author of an asserted Word document and would like to explore and read more about that document, the indexing association will link the asserted author (which is a resource) to that asserted document file. In other words, the asserted indexing association indexes the evidence space to allow easy and fast retrieval of asserted files.

In the proposed framework, setting up the asserted indexing association actually means linking resources from the File Description database to the corresponding ontological concepts. Since the File Description database retains only the extracted asserted metadata of files from the evidence space, the asserted indexing association will be restricted to that information. Nevertheless, the level to which the resources could be extracted may vary depending on what information there is to examine. For instance, the asserted file's content may be utilized to retrieve more resources. This process may be automated through techniques that are commonly seen in Artificial Intelligence (AI). Some techniques that may be utilized here are:

Named Entity Recognition (NER). NER is part of a multiple Information Extraction (IE) processes. It automatically recognizes named entities from texts. These named entities could include people names, organization names, location names, time, and date. Early NER systems were utilizing rule-based approaches in identifying named entities; however, recent systems are making use of machine learning approaches. Nadeau and Sekine (Nadeau and Satoshi 2007) presented a detailed survey about NER systems and their techniques. The benefit that NER techniques can add to the proposed framework is that of another source where more resources (such as people and organization names) can be indentified from file contents. The newly discovered entities can then be mapped to the proper ontological concept.

Hyperlink Extraction. Hyperlink extraction techniques can be used to identify and extract hyperlinks from file content such as from textual and XML files. Each hyperlink would be an instance of the domain ontology URL which creates a Core Knowledge asserted association. For instance, if we have an XML file with a URL in its content, there would be an asserted association between the concept XML File from the File domain ontology and the URL domain ontology that may be represented by the relation *containsUrl*. Indeed, there might be also an indexing association between that URL resource and the XML file.

Terms and Theme Extraction. To gain a wider view of the semantic content of a document or a collection of documents, Terms Extraction techniques can be used to gather words from these documents that may be used to describe their content. Terms Extraction can be also extended to classify a collection of documents based on the underlying asserted semantic context of their content; which is known as Theme Extraction. This allows for further classification of documents into the ontology.

6. IMPLEMENTATION

In this paper, we have limited the scope of discussion to the encoding of potential evidence acquired from smart phone devices. The current architecture of the most popular smart phone operating systems (such as Android and iOS) allows the sharing of data to be done internally between applications and externally between multiple platforms. For instance, the contact manager of a smart phone may fuse contacts gathered from more than one source. Contacts from social networking accounts and instant messaging accounts may all be merged and presented in one central place. Such a feature allows for more relations to be discovered from various information resources. For this paper, we have simulated a demonstrative tool using TopBraid[1] – an ontology editing tool that facilitates the creation, editing and querying of ontologies and knowledge bases. The Android operating system has been chosen to apply the proposed framework to.

6.1 Ontology for Smart Phones

Since we are dealing with smart phones, a set of ontologies has been developed to model this environment. Once again, for demonstration purposes, only four simplified ontologies are implemented. The proposed framework divides ontologies into two types, namely Domain Ontology and Application Ontology. Domain ontologies are used to model a particular concept in the environment. The domain ontologies that have been developed for this demonstration are: Contact, SMS Message, and Call Log ontologies. These three ontologies form the basis of the application ontology (the forth ontology), which interconnects between domain ontologies through a set of asserted relations. At this stage, a set of terms has been identified to form classes and subclasses of the ontologies as well as the asserted relations between them. The language used to model these ontologies is RDF Schema (RDFS) and Figure 4 illustrates the application and domain ontologies that model a small element of the smart phone environment.

[1] Available at: http://www.topquadrant.com/products/TB_Composer.html

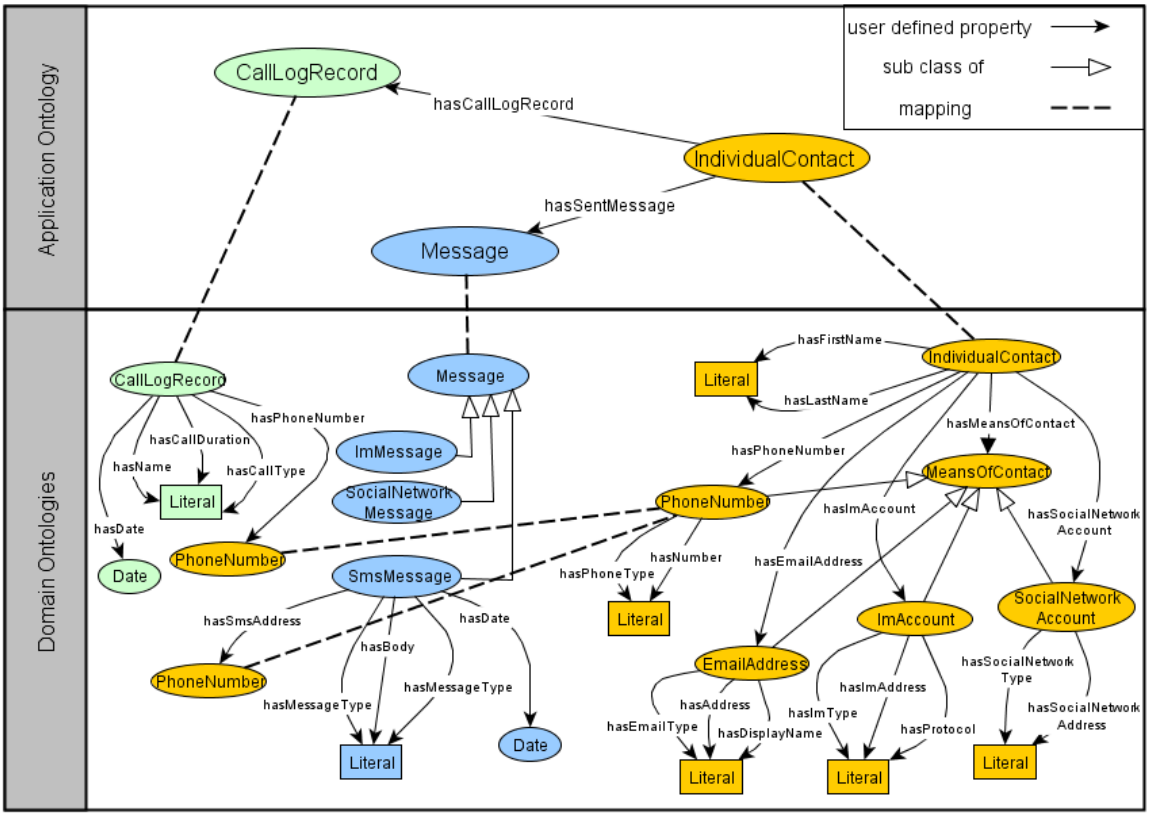

Figure 4 Representation of domain and application ontologies of some concepts in the smart phone environment

6.2 Annotation of Data

The main layer in the proposed framework that is responsible for annotating data from the evidence space is the File Wrappers layer. As discussed in the File Wrappers Section, this layer is device-dependent; which means that different devices require different file wrappers. This is directly linked to the structure in which the data is stored in. Since we are focusing on Android smart phones, most of the relevant data to our analysis is typically stored in SQLite databases. Therefore, a set of file wrappers has been implemented to interact with data that corresponds to the developed ontologies (i.e. for Contact, SMS Message, and Call Log ontologies). These file wrappers superimpose the asserted metadata of resources found in these databases. For instance, contacts information is typically disseminated among multiple tables in the contact database. In this case, the purpose of the file wrapper is to gather contacts information along with its asserted metadata and convert it to the RDF format. The annotated traces are then stored in the File Description database. A portion of the annotated trace for a single contact resource is shown in Listing 1 (represented in the Turtle format).

```
contact:MohammedAlzaabi
      a       contact:IndividualContact ;
      contact:hasEmailAddress
              <http://malzaabi.com/domainOnt/contact#malzaabi@m.com> ;
      contact:hasFirstName
              "Mohammed"^^xsd:string ;
      contact:hasLastName
              "Alzaabi"^^xsd:string ;
      contact:hasPhoneNumber
              <http://malzaabi.com/domainOnt/contact#+971500000456> .
```

Listing 1 Annotated data for a single contact resource

6.3 Association between Data

The type of asserted association represented for this demonstration is limited to the Core Knowledge Association. This type of asserted association finds asserted relations between resources identified by the annotator and stored in the File Description database as RDF triples. This is performed by the Concept and Relationship Extraction layer which takes the vocabulary defined by the ontologies as an input. The layer will try to map between the ontologies' vocabulary and the RDF triples; and hence, populate the ontologies.

Reasoning engines take an essential role in finding asserted associations which are not explicitly defined in the knowledge space (i.e., in the RDF triples). Such asserted associations allow the examiner's view to be explored to potential evidence that cannot be easily interpreted by human minds. It also facilitates asserted linkages between data from multiple domain ontologies. This, of course, requires further rules to be added to the reasoning engine. As this is an ongoing research, simple reasoning rules have been used. One of these rules is permitting a more abstracted view of the evidence. For instance, since we did not explicitly declare that all instances of the SMS Message class are also instances of the Message class (although SMS Message is normally a subclass of the Message class), querying the knowledge without the reasoning engine will not return any instances from the Message class. In this case, using a reasoning engine will allow such inference to be carried out. This becomes particularly handy as the Message class will combine all asserted messages from its subclasses; which permits the examiner to see asserted messages from various resources.

6.4 Enquire of the Knowledge

Once the ontologies are populated and stored in the RDF/OWL database, the knowledge base becomes ready to be interrogated. The query language used is SPARQL Protocol and RDF Query Language (SPARQL). SPARQL was developed to query graph-based datasets. Like SQL language, SPARQL allows the user to retrieve data that satisfies a given selection statement. Although for this demonstration we use SPARQL to directly query the knowledge base, for a practical analysis tool, it can be extended to be used with a graphical user interface where the examiner does not need to have a prior knowledge about SPARQL.

Enquiring of the knowledge base can be performed in three levels; within a single domain ontology, between two or more domain ontologies, and within the application ontology. Table 1 explains these three levels with an example for each one.

Table 1 Examples of SPARQL queries for different levels of the ontology

Level	Targeted Ontology	Example	Comment

Single Domain Ontology	Contact ontology	**SELECT** ?name ?phoneNumber
		WHERE {
		?contactResource contact:hasFirstName ?name; contact:hasPhoneNumber ?phoneNumberResource .
		?phoneNumberResource contact:hasNumber ?phoneNumber.
		}
			This query retrieves all contact first names and phone numbers.

Multiple Domain Ontologies	Contact and Message ontologies	**SELECT** ?phoneNumber ?body
		WHERE {
		?phoneNumResource contact:hasNumber ?phoneNumber.
		?msgResource message:hasSmsAddress ?phoneNumber .
		?msgResource message:hasBody ?body .
		FILTER regex(?phoneNumber, "X")
		}
			This query retrieves all SMS messages that have been sent by a contact with the phone number X.

Application Domain Ontology	Application ontology	**SELECT** *
		WHERE {
		?contactReource app:hasCallLogRecord ?callLogRecordReource.
		}
			This query retrieves all call log records for all contacts.

7. CONCLUSION

In this paper, we have presented the ongoing development of a potential next generation of forensic analysis tools. The framework described in this paper is based on the use of semantic web technologies which aims to provide a semantic-rich environment to facilitate evidence analysis. Ontologies are used to model the environment under examination. This model encompasses resources and their asserted relations in a graph-based dataset by utilizing the RDF language. We believe that this tool can reduce some of the problems associated with the large volumes and complexity of data under analysis. In addition, to validate the effectiveness of the proposed framework, a simulation program has been developed to analyze contents retrieved from Android smart phones. Several ontologies have also been created to model some concepts involved in the smart phone environment. However, several limitations should be acknowledged in this proposal, such as the assumptions made throughout the development of the ontologies. To some extent, the ontology represents the interpretation of the developer (or a group of developers) towards a domain of discourse in the real world. Such assumptions may mislead the examiner and result in a wrong interpretation of the discovered traces. Therefore, the examiner must consider both the nature of the methods used and their limitations to avoid errors or omissions in their results. Another limitation lies in presenting too many traces to the investigator which may lead him/her astray. As such, previous knowledge about the case may help the investigator to shed the light on some key traces.

REFERENCES

Allemang, D., and Hendler, J. (2011). Semantic Web for the working ontologist: Effective modeling. *RDFS and OWL*, Morgan Kaufmann, USA.

Ayers, D. (2009). A Second generation computer forensic analysis system. *Digital Investigation*, *6*(supplement), S34–S42.

Fensel, D., Bussler, C., Ding, Y., Kartseva, V., Klein, M., Korotkiy, M., Omelayenko, B., and Siebes, R. (2002). Semantic Web Application areas. *The 7th International Workshop on Applications of Natural Language to Information Systems*, June 27–28, 2002. Stockholm, Sweden.

Garfinkel, S. L. (2010). Digital Forensics research: The next 10 years. *Digital Investigation*, *7* (supplement), S64–S73.

Gruber, T. R. (1993). A Translation approach to portable ontology specifications. *Knowledge Acquisition*, *5*(2), 199–220.

Guizzardi, G. (2007). On Ontology, ontologies, conceptualizations, modeling languages, and (meta) models. *Frontiers in Artificial Intelligence and Applications*, 155, 18–39.

Happel, H., and Seedorf, S. (2006). Applications of ontologies in software engineering. *Workshop on Sematic Web Enabled Software Engineering*, November 5–9. Athens, GA, U.S.A.

Horrocks, I. (2008). Ontologies and the Semantic Web. *Communications of the ACM*, *51*(12), 58–67.

Kahvedžić, D., and Kechadi, T. (2009). DIALOG: A Framework for modeling, analysis and reuse of digital forensic knowledge. *Digital Investigation*, *6*(supplement), S23–S33.

Kahvedžić, D., Kechadi, T. and Baggili, I. (2011). Semantic Modelling of digital forensic evidence. *Digital Forensics and Cyber Crime*, Ibrahim Baggili (ed), 149–156. Springer Berlin Heidelberg.

Klyne, G., Carroll J. J., and McBride, B. (2013). Resource Description Framework (RDF): Concepts and abstract syntax. Retrieved from http://www.w3.org/TR/rdf-concepts/ on February 10, 2013.

Louis, A. L., and Engelbrecht, A. P. (2011). Unsupervised discovery of relations for analysis of textual data. *Digital Investigation*, *7*(3–4), 154–171.

Mika, P. (2005). Ontologies Are Us: A unified model of social networks and semantics. *International Semantic Web Conference*, November 6–10. Galway, Ireland.

Nadeau, D., and Satoshi, S. (2007). A survey of named entity recognition and classification. *Linguisticae Investigationes*, *30*(1), 3–26.

Raghavan, S., Clark, A., and Mohay, G. (2009). FIA: An open forensic integration architecture for composing digital evidence. *Forensics in Telecommunications, Information and Multimedia*, ed. Matthew Sorell, 83–94. Springer Berlin Heidelberg.

Stevens, R., Goble, C. A., and Bechhofer, S. (2000). Ontology-based Knowledge representation for bioinformatics. *Briefings in Bioinformatics*, *1*(4), 398–414.

Turner, P. (2005). Unification of Digital evidence from disparate sources (digital evidence bags). *Digital Investigation*, *2*(3), 223–228.