An open architecture for digital evidence integration

Bradley Schatz and Andrew Clark
Information Security Institute
Queensland University of Technology
{b.schatz, a.clark}@qut.edu.au

Abstract

Recently the need for "digital evidence bags" – a common storage format for digital evidence – has been identified as a key requirement for enabling inter-organisational sharing of digital evidence, and interoperability between forensic analysis tools. Recent work has described an ontology based approach to correlation of event log based evidence, using semantic web technologies for describing and representing event log based digital evidence. In this paper we apply the representational approach to the integration of metadata related to digital evidence, and propose a globally unique identification scheme for digital evidence and related metadata. We relate the representational approach to the digital evidence bags concept identifying a number of shortcomings. We propose an alternative architecture for digital evidence bags, which we call the sealed digital evidence bags architecture. This approach treats bags as immutable objects, and facilitates the building of a corpus of digital evidence by composition and referencing between evidence bags. This architecture facilitates modular forensic tool development and interoperability between forensics tools.

1 INTRODUCTION

The rapid pace of innovation in digital technologies presents substantial challenges to digital forensics. New memory and storage devices and refinements in existing ones provide constant challenges for the acquisition of digital evidence. The proliferation of competing file formats and communications protocols challenges one's ability to extract meaning from the arrangement of ones and zeros within. Overarching these challenges are the concerns of maintaining the integrity of any evidence found, and reliably explaining any conclusions drawn.

Researchers and practitioners in the field of digital forensics have responded to these challenges by producing tools for acquisition and analysis of evidence. To date, these efforts have resulted in a variety of ad-hoc and proprietary formats for storing evidence content, analysis results, and evidence metadata, such as integrity and provenance information. Conversion between the evidence formats utilized and produced by the current generation of forensic tools is complicated. The process is time consuming and manual in nature, and there exists the potential that it may produce incorrect evidence data, or lose metadata (DFRWS 2005). Validation of the results produced is hindered by this lack of format standardisation.

It is with these concerns in mind that calls have been made for a universal container for the capture of digital evidence. Recently, the term "Digital evidence bags" was proposed to refer to a container for digital evidence, evidence metadata, integrity information, and access and usage audit records (Turner 2005). Subsequently, the Digital Forensics Research Workshop (DFRWS) recently formed a working group with a goal of defining a standardised Common Digital Evidence Storage Format (CDESF) for storing digital evidence and associated metadata (DFRWS, 2005).

Another source of complexity related to the ad-hoc nature of forensic tools is the absence of a common representational format for evidence metadata. This is not a trivial problem due

to the nature of the forensics domain, which deals with massive conceptual complexity within multiple layers of abstraction. The challenge here is to identify a means that decouples the models of evidence and evidence metadata used by forensics tools from the implementation logic of these tools. Furthermore, this needs to be accomplished in a manner that facilitates the establishment of provenance and maintains integrity.

This representational problem is not simply limited to the challenge of tool interoperability. In outlining the "Big Computer Forensic Challenges", Spafford observes that practitioners and researchers in the field of digital forensics do not use standard terminology (Palmer, 2001). It is not surprising then that we find limited attention paid to the formal definition of taxonomies or ontologies describing this domain.

We propose the use of ontologies as a solution to these terminological and representational problems. We have produced a number of basic ontologies modelling the domain of digital evidence acquisition, computer hardware, and networks, and described these ontologies using the Web Ontology Language (OWL) (McGuiness, 2001). A subsumer of taxonomies, ontologies are defined by Gruber (1993) as an explicit specification of a conceptualization. More descriptively, an ontology is a means of conceptualising a domain of discourse, in terms of concepts, properties, and relationships of entities. Ontology languages, the purpose of which are to describe ontologies, hold the promise of empowering machines to have greater ability to reason over and analyse information, by nature of sharing a common understanding of the information at hand (Undercoffer et. al., 2004). Furthermore, ontologies encourage knowledge sharing and reuse within a domain, which has the potential to lead towards a convergence of vocabulary in the forensics domain.

In this paper we propose an open architecture for integrating digital evidence by applying an ontology based approach to Turner's digital evidence bags concept. We enumerate the representational requirements for the metadata component of an open common digital evidence storage format, and formalise the domain by describing it with an ontology. We demonstrate an architecture for digital evidence bags which facilitates modular composition of forensic tools by way of an open metadata format. Further, we propose a novel means of identifying digital evidence, and digital evidence bags, which supports arbitrary referencing of information within and between digital evidence bags. Additionally, we propose an alternative approach to Turner's design, based on a sealed bag metaphor.

2 RELATED WORK

In this section we firstly describe current approaches to digital evidence storage containers. We then review current architectures. In order to put the digital evidence containers and metadata in the context of the forensic tool landscape, we then outline the state of the art in theory of operation of digital forensics tools. To provide background to our metadata approach, we describe the basics of the World Wide Web Consortium's (W3C's) semantic web technology stack, the ontology language OWL and the data model called Resource Description Framework (RDF). A survey of the use of ontology in the computer security field is then presented.

1.1 Digital evidence storage formats

The Advanced Forensics Format (AFF) has recently been proposed as a disk image storage format, It includes storing of acquisition related metadata in the same container as the disk image. Garfinkel et al (2006) describe the AFF and summarise the key characteristics of nine different forensic file formats, and outline the desirable characteristics for an image storage container. They conclude that the AFF is the only publicly disclosed forensic format which supports storage of arbitrary metadata. However, the metadata storage mechanism in

the AFF is limited to name/value pairs and makes no provision for attaching semantics to the name.

The de-facto standard in commercial forensics software, Encase[1], uses a monolithic case file for storing case related metadata and stores filesystem images in separate and potentially segmented files. The format of the case file is proprietary.

1.1.1 Digital evidence bags

Turner introduces the concept of a Digital Evidence Bag (DEB) by attempting to replicate the key features of the physical evidence bags used for traditional evidence capture. The key structural components of a physical evidence bag are the bag itself, a means of bag identification (potentially a serial number), an area for recording evidence related information (which we refer to as a tag), and optionally, a tamper evident security seal. We categorise the key features of physical evidence bags as follows:

Evidence Metadata Records : Standard evidence metadata includes a description of the evidence, the location, date and time of the acquisition of the evidence.
Provenance Records : Includes chain of custody information, as well as information pertaining to the collector of the evidence.
Identification Records :Identification information includes a unique serial number (or seal number) which uniquely identifies the bag, and other case related information such as the case number, item number, collecting organisation, suspect and victim.
Integrity Device : Pieces of evidence collected at an investigation scene are placed in evidence bags and sealed on the spot, potentially with a tamper evident tape closure seal. This seal, and the construction characteristics of the bag itself, help to ensure Integrity of the evidence by indicating tampering.
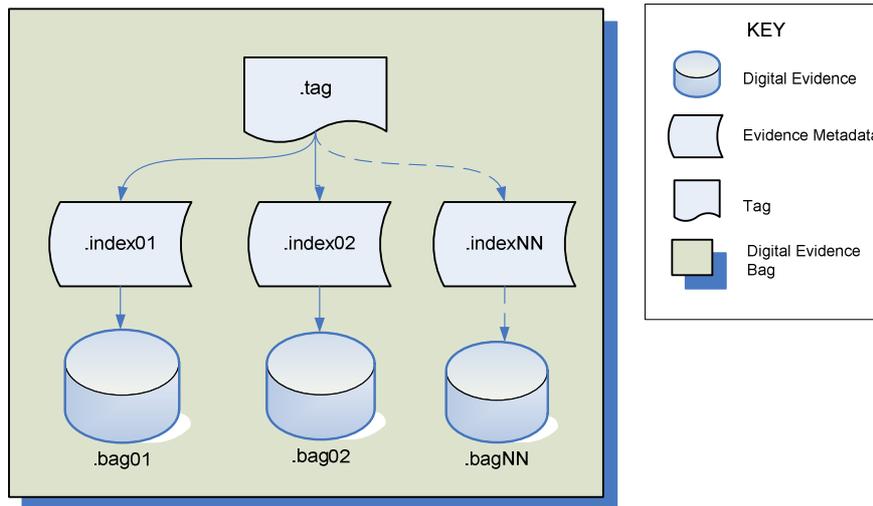**Evidence Container** : The inside of the bag.

It is worth noting here that the use of the features listed above varies dependent on jurisdiction.

Turner's proposal for a digital evidence bag addresses a number of aspects of the above features. A file archive structure is proposed which defines a specific naming scheme for files containing digital evidence, separate files containing evidence metadata, and a singular file which contains evidence integrity, provenance and identification information. Figure 1 depicts the structure of Turner's digital evidence bag.

A DEB is a collection of the Digital Evidence files, Index Files and a single Tag File. Turner does not detail the implementation of the container grouping these evidence files, however we assume that in practise, a DEB would be an archive (tar, zip, etc) within which these files are contained.

---

[1] http://www.guidancesoftware.com/

.tag

.index01 .index02 .indexNN

.bag01 .bag02 .bagNN

KEY

Digital Evidence

Evidence Metadata

Tag

Digital Evidence Bag

Figure 1: Turner's Digital Evidence Bag

Individual elements of digital evidence collected, such as filesystem images, network traces, or the contents of image files are stored in Digital Evidence Files, which are identified by a file extension *.bagNN*. The NN refers to a unique number. Correspondingly, Evidence Metadata, such as file last access time is stored in similarly named files with an extension *.indexNN*. The pairing of a single Digital Evidence File with its corresponding Evidence Metadata file is refered to by Turner as an Evidence Unit. Turner does not describe naming of the files other than the extensions defined. It is unclear as to whether of not multiple pieces of content are stored in a *.bagNN* file.

Integrity, provenance and identification information are stored as unstructured text within the Tag File, which is identified by the file extension .tag. The tag file also enumerates the names of all of the Evidence Units.

The architecture of Turner's Digital Evidence Bags appears to be oriented more towards a single monolithic digital evidence bag being used in a case, as a container for all digital evidence acquired. Secondary evidence (evidence derived from the analysis of earlier acquired evidence, such as files extracted from a filesystem image) would appear in this scheme to be added to the same digital evidence bag as the original image. This involves modification to the tag file and the addition of new files to the evidence bag. Provenance is assured by the onion like use of hashing of the contents of the Tag File.

A potentially confusing aspect of Turner's DEB proposal is that modification of the Tag file, and the addition of new files to the DEB may lead the layman to the conclusion that the monolithic bag is in fact never sealed, thus raising doubts as to the integrity of the evidence. While this may be seen more as an impedance mismatch in translating the evidence bag metaphor, we suggest an alternate architecture for digital evidence bags. The architecture we present favours treating evidence bags as immutable objects. Addition of information is done outside the bag, in much the same way that information is added to the tag of a physical evidence bag without breaking the seal of the bag.

Turner's structure does not define a scheme for referencing of evidence and metadata between digital evidence bags. Therefore the ability to compose multiple evidence bags into a corpus does not appear feasible. The format and vocabulary of the Evidence Metadata, Identification Information and Provenance Information is syntax free, and unspecified, geared only towards human interpretation.

1.2 Ontology and knowledge representation languages

Current approaches to representing digital evidence and digital evidence related metadata may be characterised as predominately ad-hoc (in the case of open source tools) or closed and monolithic (in the case of proprietary tools). In response to the absence of an interoperable evidence representation we find inspiration in the recent efforts of the WWW Consortium (W3C) towards the vision of a Semantic Web, a web of machine and human interpretable data (Berners-Lee et. al., 2001). While the goal of the semantic web is a web of information published on the internet, our proposal parallels this in the small by proposing the architecture for a semantic web of evidence. In this section we present a brief introduction to standards which comprise the current semantic web technology stack.

Currently, the semantic web stack comprises two logical layers: a data layer, and an ontology layer (presented below in Figure 2).
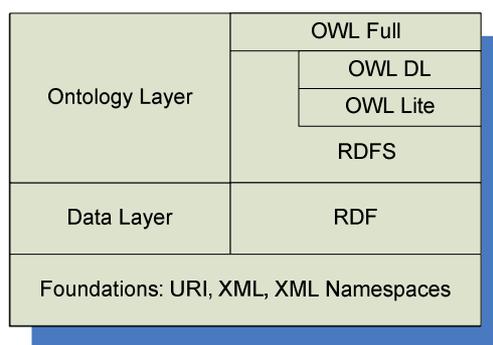
Figure 2: Current Semantic Web Standards

The data layer is comprised of the Resource Description Framework (RDF) language, which is based on a simple graph based data model. RDF is used to provide a common data representation. The basic unit of data in an RDF graph is the triple, which consists of a Subject, a Predicate, and an Object. The triple is represented using nodes and a single arc[2] as in Figure 3. The meaning of a RDF triple, when asserted, is that a relationship, named by the Predicate exists between the Subject and the Object.

Figure 3: An RDF Triple

The Subject and Predicate are both named using a Uniform Resource Identifier (URI) (Berners-Lee et. al. 1998) Often, a URL, a more specific type of URI is used. The Object may either be a URI or a Literal, which is a string that may be optionally given a data type. The use of URI's to name the nodes and arcs in RDF triple enables building complex graphs of assertions.

The RDF data model is defined in terms of nodes and arcs, as such, it is an abstract model. A textual serialization of RDF graphs is a necessary precursor to interoperability. For this the RDF standard defines an XML based serialization called XML/RDF (Klyne & Carrol, 2004).

While the semi-structured data model of RDF provides an extensible vocabulary for naming elements of data, it only provides rudimentary features for categorising these data elements

---

[2] We note that the arrow in the diagram indicates the ordering relationship, not flow.

within the graph. For example, a popular model for categorising "things" is the Frame-based one originally proposed by Minsky (1975), where a frame represents an object or concept, to which are attached attributes (or properties, or slots) that represent component parts of the concept or object. The Object-Oriented paradigm underlying Java and C++ may be seen as the application of frame based theory to the structuring of software (Lassila, et al., 2001).

Recent evolution of the Frame based-approach to knowledge representation has led to ontologies becoming widely used as a means for specifying and describing concepts and their relationships (McGuiness, 2001). Several ontology languages have been developed in recent times, including the Web Ontology Language (OWL), and DAML+OIL (Harmelen et al., 2004). OWL, which has recently been standardised by the W3C, provides a mechanism for describing an ontology in terms of collections of descriptions of concepts within a domain of discourse (classes), properties of classes, and restrictions on properties. OWL hooks into the RDF language in a manner that attaches meaning to the vocabulary used within the RDF.

Both OWL and DAML+OIL are based on a branch of logics called Description Logics (DL). These logics are a subset of First Order Logic (FOL) that are well suited to expressing terminology and instance information, with efficient and decidable classification oriented inference characteristics. The OWL language, itself a dialect of RDF, provides support for merging of ontologies, through the use of language features which enable importing other ontologies and enable expression of conceptual equivalence and disjunction (Smith et al., 2004). This facilitates separate ontology development, refinement and re-use.

1.3   Ontology in computer security & forensics

There is little to no published research specifying formal ontologies for computer forensics or computer-related crime. Schatz et. al. (2004), however use an ontology to describe categorisations and abstraction hierarchies modelling the concepts embodied in event logs. Abstraction aware rules are used to correlate events to higher level event abstractions for the purpose of forensics. This work publishes no ontology, and does not address issues of evidence identification, integration, provenance or integrity.

We have identified limited literature focusing on modelling of the domain of cases. Bogen and Dampier (2005) apply case domain modelling as a structured approach for analysing case facts, identifying relevant case concepts, and documenting this information. Their focus is on modelling as a conceptual tool informing methodology, rather than as a means of fixing the semantics to evidence material.

A number of applications of ontologies have been observed in the computer security field especially relating to intrusion detection. Raskin et al. (2001) argue for the adoption of ontology as a powerful means for organising and unifying the terminology and nomenclature of the information security field. They observe that the use of ontology in the information security field will increase the systematics, allow for modularity and could make new phenomena predictable within the security domain. Schumacher (2003) focuses on systematic approaches to improving software security, by using *Security Patterns*, the application of the design patterns approach to security. Ontologies are used as a means to model both the security concepts referred to by the patterns, as well as the patterns themselves.

A "Target Centric Ontology for Intrusion Detection" describing model of computer attack was produced by Undercoffer et al. (2004). Their ontology is based on the following basic classes : Host, System Component, Attack, Input, Means, and Consequence. They use this

ontology as the model for a rule based distributed IDS. The "Network Entity Relationship Diagram (NERD)" (Goldman et al. 2001) ontology was defined as a component of an IDS alert fusion prototype, SCYLLARUS. This ontology was defined in the early description logic environment CLASSIC (Borgida et al., 1989). Only the ontology, which contains concepts focused around network and host was published.

3 DEFINITIONS

Our concerns involve representation and terminology. In order to avoid confusion, we define the following terms we use throughout the paper, and in our digital evidence ontology. As we are talking about Digital Evidence, we omit the use of the word Digital in our definitions.

Evidence Content: Contiguous bytes of computer data. Typically data which is stored in a file, or a stream of a file, or in raw storage, such as the ordered sectors of a disk.
Evidence Content File: A file containing evidence content.
Evidence Metadata: Contextual information which is related to Evidence Content. For example, commonly gathered Evidence Metadata related to a JPEG image might be the file name, the path which it was stored in, and the last modification, last access and creation times of the file.
Provenance Metadata: Information which relates to the provenance of the evidence. For example, information about who captured the evidence, where it was stored, what tools were used, and integrity controls fall into this category.
Integrity Metadata: Metadata which is used to detect the modification of evidence content or metadata.
Digital Evidence: Refers to a related set of Evidence Content (or analysed evidence content), Evidence Metadata.
Secondary Evidence : Digital evidence produced as a product of an analysis tool.
Image : A contiguous sequence of bytes, which is a copy of a digital artefact.

4 AN OPEN ARCHITECTURE FOR DIGITAL EVIDENCE BAGS

The primary aim of our work is to identify a general solution which meets the representational needs for storing metadata in digital evidence bags. We seek to do this in a manner that allows evolution, separate definition of, and interoperability between the abstractions which are used in forensic tools, in a manner that is not dependant on the management of a single entity or governing body.

We look to the near future, where analysis cases may involve digital evidence from sources orders of magnitude more numerous than the current norm. In fact we see the beginnings of this challenge as investigations of P2P networks involve multiple terabyte sized images, sourced from numerous locations and computers. We expect that the monolithic approaches to digital evidence containers will not scale to this future, for reasons such as evidence bag size, concurrent access, and IO efficiency. For example, consider the a case where two multi-terabyte images must be acquired. The use of a single monolithic DEB for containing both images would imply serialising access to the DEB, and prohibit acquiring the images in parallel. With current IO speeds, this would add tens if not hundreds of hours to the acquisition time.

In order to address these scaling issues we propose a compositional rather than monolithic approach to assembling of a corpus of digital evidence. We do this by defining an identification scheme that is independent of location and global in nature. This architecture facilitates the building of a corpus of evidence by recursively embedding digital evidence bags within digital evidence bags, as well as by reference, which we depict in Figure 4. We

call our architecture Sealed Digital Evidence Bags (SBEB's) in reference to Turner's proposal.

Figure 4: Referencing nested and external digital evidence bags

For example, in the case of the multi-terabyte imaging case discussed above, both imaging processes could happen in parallel, producing two digital evidence bags. A further digital evidence bag, which references both these images could then be used for adding provenance metadata such as the examiner's name and case number.

1.4   Evidence and metadata identification

Recalling that in RDF, Subjects, Predicates and Objects are named using a URI, we use a special category of URI called a Uniform Resource Name (URN) (Moates, 1997) for identifying digital evidence bags and metadata instances. URN's are intended to serve as a persistent, location-independent resource identifier. Following work done in the life sciences area in uniquely identifying proteins in distributed databases, which has resulted in the definition of the Life Sciences Identifier (LSID) standard (Seneger, 2004), we propose a digital evidence specific URN scheme. This scheme, which we call Digital Evidence Identifiers (DEID) is based on the organisation of the tool user, and employs message digest algorithms as a globally unique identifier. The format of a Digital Evidence Identifier is as follows:

urn:deid:organisation:digestalgorithm:digest:discriminator

For example, we identify a particular image taken of a file in our example further below using the following URN:

urn:deid:isi.qut.edu.au:sha1:dc04e8f06b2a32e7d673c380c4d2c8a1d5ea17d4:image

The string "deid" is used to provide a unique namespace for digital evidence identifiers. We provide scoping information in the organisation field which would potentially enable one to resolve a URN back to set of information or an evidence bag as has been done in the LSID work. The *digestalgorithm* field refers to the message digest algorithm used to generate text in the following field. The *descriminator* field is provided for further addition of naming terms. It should be noted that we rely on the collision free nature of message digest algorithms in order to assure globally unique names. Given the current state of uncertainty with regard to the possibility of collisions using MD5 or SHA1, our proposal provides for the use of other digest algorithms.

Of course these identifiers are long and unwieldy and not suited for use as names for the evidence we are concerned with. Evidence may be given more human friendly, case specific names by asserting further RDF triples which have the identifier as the subject.

1.5 Digital evidence bag structure

Sealable digital evidence bags follow a similar structure to Turner's bags. However, in order to facilitate an interoperable representation, we use RDF for the Tag and Evidence Metadata. The Tag File of any digital evidence bag is called *Tag.rdf*. The naming of the Evidence Metadata files is tool or user determined, however the extension is *.rdf* to signify that the format of the file is RDF.

The XML/RDF format does not support recursive definition of RDF/XML content within the content of another RDF/XML content block, and makes no provision for arbitrary text outside the syntax of the XML syntax. This leads us to maintain integrity information regarding the content of the Tag in a file external to the Tag, unlike the DEB proposal. Turner's DEB uses an onion like approach where a hash of the contents of the Tag is recursively appended to the Tag. We instead define a Tag Integrity File, called *Tag.rdf.sig*, which contains integrity information pertaining to the Tag. Sealable digital evidence bags are designed to be created and populated with evidence and metadata, then sealed exactly once. The Tag of an SDEB is immutable after the Tag Integrity File has been added to the SDEB. Before that the bags are unsealed and mutable.
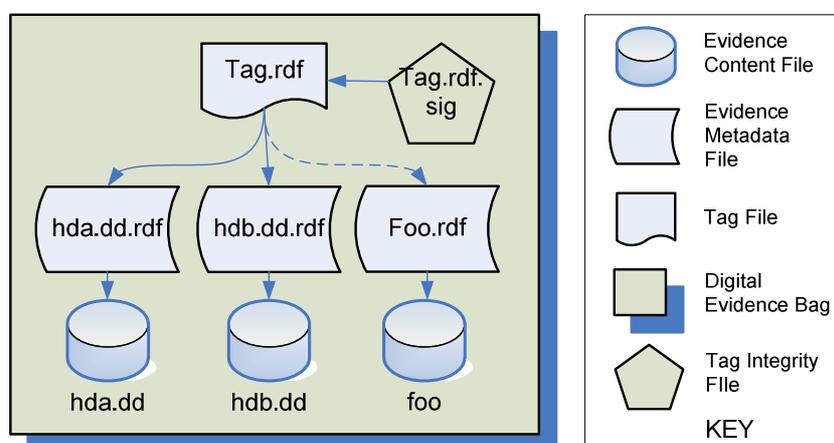
Figure 5: Proposed digital evidence bag structure

In order to demonstrate the SDEB architecture in context, we have developed a prototype online acquisition tool for creating a digital evidence bag containing images of the Internet Explorer cache and history index files (these are also referred to as web browser logs). These files are typically located in a number of subfolders of the *\Local Settings\Temporary Internet Files* path under the user's profile directory on a windows host. The files in question are all named *index.dat*.

We present the contents of the digital evidence bag produced by the prototype tool called *acquireIELogs.py* in Table 1. The tool creates images of the browser log files according to a programmatic naming scheme based on their original filename, in combination with the user name, the type of file (cache or history), and the specific history file set. For reasons discussed below in the section titled "Integrity" we do not define the integrity mechanism for the Tag File. As such the tool does not produce the Tag Integrity file.

```
jbloggs.history.MSHist012006010420060105.index.dat.rdf
jbloggs.history.MSHist012006010420060105.index.dat
jbloggs.history.MSHist012006010320060104.index.dat.rdf
jbloggs.history.MSHist012006010320060104.index.dat
jbloggs.history.MSHist012005121220051219.index.dat.rdf
jbloggs.history.MSHist012005121220051219.index.dat
jbloggs.history.MSHist012005121920051226.index.dat.rdf
jbloggs.history.MSHist012005121920051226.index.dat
jbloggs.cache.index.dat.rdf
jbloggs.cache.index.dat
```

jbloggs.history.index.dat.rdf
jbloggs.history.index.dat
Tag.rdf

Table 1: The contents of a browser log SDEB

1.6 Metadata Model

The Evidence Metadata Files produced by the prototype tool all contain metadata of a similar format to that presented in Table 2 in an abridged form. Figure 5 presents a node arc graph of the portions of the same data. In this case, we discuss the Evidence Metadata File named *jbloggs.cache.index.dat.rdf*.

```
<de:FileImage rdf:about="urn:deid:isi.qut.edu.au:sha1:4056e4786fc460d9adbe98a0bc19b29a2104c476:image">
        <de:imageContainer rdf:resource="file:///./jbloggs.cache.index.dat"/>
        <de:imageOf
rdf:resource="urn:deid:isi.qut.edu.au:sha1:4056e4786fc460d9adbe98a0bc19b29a2104c476:original"/>
        <de:acquisitionTool>
                <de:OnlineAcquisitionTool rdf:about="http://www.isi.qut.edu.au/2005/acquireIELogs.py">
                        <de:name>acquireIELogs.py</de:name>
                        <de:version>0.1</de:version>
                </de:OnlineAcquisitionTool>
        </de:acquisitionTool>
</de:FileImage>

<wb:BrowserCacheFile rdf:about="urn:deid:isi.qut.edu.au:sha1:4056e4786fc460d9adbe98a0bc19b29a2104c476:original">
        <fs:filePath>D:\Documents and Settings\jbloggs\Local Settings\Temporary Internet
Files\Content.IE5\\index.dat</fs:filePath>
        <de:messageDigest
rdf:datatype="http://www.w3.org/2000/09/xmldsig#sha1">4056e4786fc460d9adbe98a0bc19b29a2104c476</de:messagedige
st>
</wb:BrowserCacheFile>
```

Table 2: XML/RDF content of Evidence Metadata File named jbloggs.cache.index.dat.rdf

This file contains RDF instance data which asserts two top level instances. The instances describe the relationship between the Evidence Content (the content of a Evidence Content File in the digital evidence bag) and the original piece of evidence, which is an image of a Web Browser Cache File, located on a particular host. The contents of these two files are, from the digital perspective, identical. This results in a DEID URN with the same message digest value. However their locations are substantially different. We discriminate between the two instances by using the labels "image" and "original" in the discriminator field of the DEID URN. This distinguishs between the *FileImage* and the *BrowserCacheFile* respectively.

The tool generates Provenance Metadata identifying itself by name, location, and version, relating itself to the *FileImage* by use of the *acquisitionTool* property. Provenance information identifying the examiner running the tool would be added to a separate evidence bag, which refers to this sealed one. We do this in order to simplify the acquisition tool, preferring that more complex data entry and annotation tasks are performed using a task specific tool, such as an analogue of Turner's Tag editor application.

The property and class names used in the vocabulary above are defined in ontologies specific to the domains of discourse that we are dealing with. The prefix *de* is an alias for an ontology stored in the document located at *http://isi.qut.edu.au/2005/digitalevidence*, which describes the digital evidence domain. Hence, *de:FileImage* refers to a specific concept (a class) defined in this ontology. Similarly we define an ontology for filesystem related concepts aliased as *fs* (*http://isi.qut.edu.au/2005/filesystem*) and web browser related aliased as *wb* (*http://isi.qut.edu.au/2005/webbrowser*).
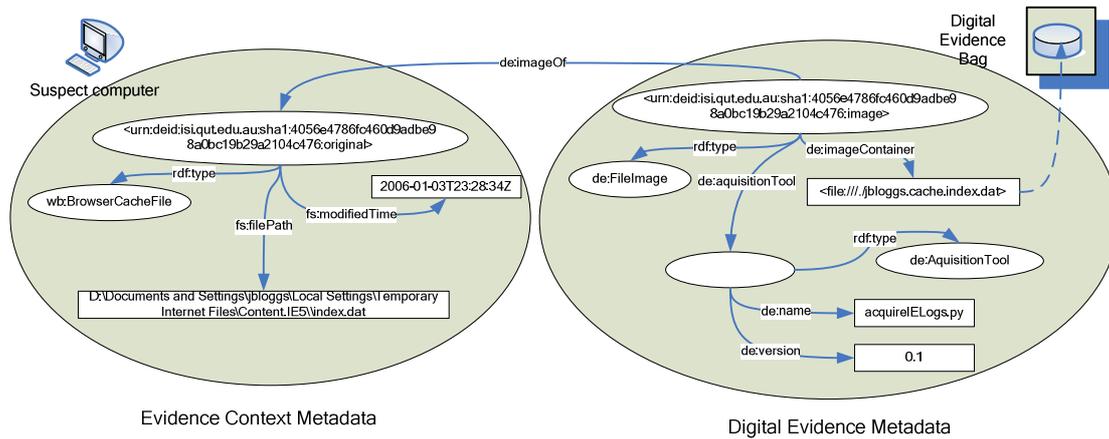
Figure 6: RDF Graph relating Evidence Context and Digital Evidence Metadata

Figure 6 depicts a portion of the RDF graph implied by the content of the Evidence Metadata File discussed above and presented in Table 2. We discriminate here between the Evidence Context Metadata, and Digital Evidence Metadata.

1.7 Tag file

The tag file contains the RDF data representing the digital evidence bag and its contents. The DEID of the *deb:DigitalEvidenceBag* instance is based on the hash of the content of the Evidence Metadata Files, in the order in which they are defined in Table 3. The *deb:bagContents* property is an ordered list which refer to instances of digital evidence metadata contained in the digital evidence metadata files.

```
<deb:DigitalEvidenceBag rdf:about="urn:deid:isi.qut.edu.au:sha1:44bc23235f5e797aae992e5de09524e9071fd8c6">
  <deb:bagContents>
   <rdf:Seq>
    <rdf:li rdf:resource="urn:deid:isi.qut.edu.au:sha1:dc04e8f06b2a32e7d673c380c4d2c8a1d5ea17d4:image"/>
    <rdf:li rdf:resource="urn:deid:isi.qut.edu.au:sha1:4a03ed30ebdf919004d4b40222b721c4771adee9:image"/>
    <rdf:li rdf:resource="urn:deid:isi.qut.edu.au:sha1:c117652d98a4f612979c19f5701d278e025749fa:image"/>
    <rdf:li rdf:resource="urn:deid:isi.qut.edu.au:sha1:05de1243f67753150334968a2effcc4f8114ef45:image"/>
    <rdf:li rdf:resource="urn:deid:isi.qut.edu.au:sha1:f3a9fd3fcc017d822f10bc4466b6d19ddbdd5042:image"/>
    <rdf:li rdf:resource="urn:deid:isi.qut.edu.au:sha1:4056e4786fc460d9adbe98a0bc19b29a2104c476:image"/>
   </rdf:Seq>
  <deb:bagContents>
</deb:DigitalEvidenceBag>
```

Table 3: Digital Evidence Bag instance data stored in the Tag File

1.8 Integrity

Current best practise for ensuring the integrity of digital evidence involves the use of collision resistant message digest functions. Typically a message digest is taken of the original evidence, and recorded in a manner that asserts the time of the digest being taken (often via contemporaneous notes or printouts). The integrity of subsequent images made, or copies of images made may then be ensured by taking the message digest of the image or copy, and comparing with the original message digest.

In this proposal, integrity of evidence and evidence metadata is ensured by the use of chained message digests. Besides using the message digest of each piece of Evidence Content as a component of a unique identifier for both the Evidence Content Metadata instance and the Digital Evidence Metadata instance, we also define a property within the class *de:EvidenceContext* class called *de:messageDigest*. This property is presented in context in Table 4.

```
<wb:IEBrowserCacheFile rdf:about="urn:deid:isi.qut.edu.au:sha1:4056e4786fc460d9adbe98a0bc19b29a2104c476:original">
        <de:messageDigest  rdf:datatype=http://www.w3.org/2000/09/xmldsig#sha1
       >4056e4786fc460d9adbe98a0bc19b29a2104c476</de:messagedigest>
```

```
</wb:IEBrowserCacheFile>
```
Table 4: Evidence Content message digest property

The value of the *de:messageDigest* property is the hash of the Digital Evidence Content obtained from the file. Work in the xml signature area has already defined a datatype representing a sha1 message digest, and defined a URI representing this datatype, we use the URL *http://www.w3.org/2000/09/xmldsig#sha1* to specify the datatype of this property.

Integrity of the Evidence Metadata Files is maintained within the Tag File, by definition of separate *de:EvidenceMetadataContainer* instances per Evidence Metadata File, as presented in Table 5. Integrity of the content is assured by the inclusion of a message digest of the Evidence Metadata File, using the *de:messageDigest* property.

```
<deb:EvidenceMetadataContainer>
  <deb:contains rdf:resource="urn:deid:isi.qut.edu.au:sha1:dc04e8f06b2a32e7d673c380c4d2c8a1d5ea17d4:image" />
  <de:messageDigest rdf:datatype=http://www.w3.org/2000/09/xmldsig#sha1
        >731251ae7216b935cccf51a4018a00d8d89a89cd</de:messagedigest>
  <fs:filePath>file:///./jbloggs.history.index.dat.rdf</fs:filePath>
</deb:EvidenceMetadataContainer>
```
Table 5: Evidence Metadata Container Metadata stored in the Tag File.

As the focus of this paper is not the mechanics of integrity maintenance, we do not specify the format or contents of the Tag Integrity File. We expect that the contents of the file may be formatted according to the XML Signatures standard (Bartel et. al., 2002), or some other standard. We do not consider here what kind of archive is used as the bag medium.

1.9 Evidence provenance

We provide no construct that directly translates to the audit oriented functions of the Tag Continuity Blocks of the DEB proposal, as we expect that further application of tools to sealed bags will result in new digital evidence bags being produced. The Provenance Metadata within these new bags would refer back to the original bag, thus serving this role.

1.10 Clarifications

It appears that the DEB allows a number of pieces of evidence to be stored in a single Evidence Content File. We restrict the definition of the Evidence Content File to refer to a container with exactly one piece of evidence content.

5   USAGE SCENARIO – IMAGING AND ANNOTATION

We demonstrate the modular manner in which forensic tools may interoperate with evidence bags built using the sealed digital evidence bags approach by way of the following example.

In this case, the examiner uses a DEB enabled hard drive imaging application for acquiring the evidence image. This tool is scripted together from a variant of the UNIX *dd*[3] tool, and the Linux *hdparm* utility[4]. The examiner acquires the hard drive using this utility, resulting in a digital evidence bag containing an Evidence Content File, called *hda.dd*, an Evidence Metadata File, called *hda.dd.rdf*, as well as *Tag.rfd*. The imaging application is designed to be as simple as possible, and produce a sealed digital evidence bag. It automatically generates a message digest of the *Tag.rdf* file and stores it in the Tag Integrity File,

---

[3] A low level block oriented copying tool found on most UNIX variants.

[4] A utility which queries information such as serial numbers, size, and addressing information from hard disks.

Tag.rdf.sig. At this point the evidence bag is sealed, and considered immutable, depending on the underlying scheme of implementation of the Tag Signature.

However, the examiner has further data associated with this digital evidence bag, namely the Job ID, a case specific name, the examiner's name and identifying details, and perhaps the serial number printed on the drive. An evidence annotation program is used by the examiner to create a new, unsealed digital evidence bag, and the original digital evidence bag embedded within it. A new Tag File is created within this the new bag by the annotation application. The additional data is entered using the annotation user interface, and added to the Tag File. In this case the annotation editor eschews creating a new Evidence Metadata File, as no new evidence has been acquired.

There are two distinct activities involved in the above scenario: evidence acquisition and evidence annotation. By the former, we refer to the process of making an exact copy of a piece of digital evidence, for example a hard disk. The latter refers to the act of recording details relevant to the acquisition process and the evidence source. By modularizing these two tasks, individual tool complexity is reduced, which has the potential to increase reliability and enable testing at a more granular level. Bugs in the consuming forensic tool (the annotation tool), are more likely not to jeopardize the integrity of the product of the evidence acquisition task.

The annotation tool annotates the information in the original sealed digital evidence bag by asserting new properties and their values, related to the DEID of the particular piece of information from the subject bag, as new RDF triples. These triples are stored in the Tag File of the new unsealed DEB. In reference to the above example, the new data is related to the instance representing the Hard Disk by means of its unique identifier. A depiction of a portion of the RDF graph formed from the new information as well as the original evidence metadata is presented in Figure 7.
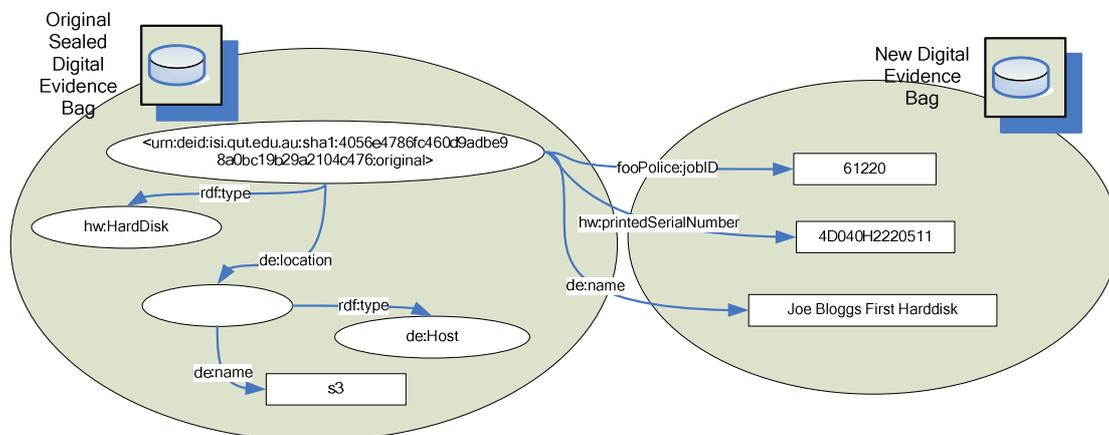
Figure 7: RDF graph resulting from addition of new metadata to embedded DEB

Modularity is not only facilitated in terms of interoperability between forensics tools, but also facilitates modular composition of ontologies. In this way an organisation could create its own specific ontology (say for the purpose of adding an organisation specific identifier) which would seamlessly integrate with the existing RDF graph and ontology. We allude to a further ontology (*fooPolice*) which defines the *fooPolice:jobID* property, in Figure 7.

6   CONCLUSION AND FUTURE WORK

In this paper we have proposed an approach to addressing the representational challenges in building modular, and interoperable forensics tools. We propose the use of Resource Description Framework (RDF) as a common data representation layer for digital evidence

related metadata, and the use of ontologies for describing the vocabulary related to this data. Furthermore, we propose a globally unique identification scheme for identifying digital evidence and related metadata. We have applied the approach to Turner's Digital Evidence Bags container proposal, and identified a number of shortcomings or omissions. An alternative structure for the digital evidence bag was proposed and a novel architecture for digital evidence integration which we call the Sealed Digital Evidence Bag approach was identified. Proof of concept was demonstrated by way of describing the operation of a prototype online acquisition application.

We have focused on validating the approach to representation in this work, have built a simple set of digital evidence related ontologies, and a prototype acquisition tool which are published at http://www.isi.qut.edu.au/2005/sdeb/. This ontology is however a proof of concept, and we believe that the field of digital forensics would benefit from a standardised ontology describing its domain.

We have not considered media layer or container of the digital evidence bag, and expect that work may be needed in reconciling the composable nature of our digital evidence bags, with concerns of IO and storage efficiency.

7   ACKNOWLEDGEMENTS

8   BIBLIOGRAPHY

Bartel, M., Boyer, J., Fox, B., LaMaccia, B., Simon, E., (2002), 'XML-Signature Syntax and Processing', http://www.w3.org/TR/xmldsig-core/, Accessed 9 January 2006.

Berners-Lee, T., Fielding, R., Masinter, L., (1998), Uniform Resource Identifiers (URI): Generic Syntax, http://www.ietf.org/rfc/rfc2396.txt, Accessed 9 January 2006.

Berners-Lee, T., Hendler, J., Lasilla, O., (2001), 'The Semantic Web', *Scientific American*, May 2001.

Bogen, A.C., Dampier, D.A., (2005), 'Preparing for Large-Scale Investigations with Case Domain Modeling', *In 2005 Digital Forensics Research Workshop (DFRWS),* New Orleans, LA, 17–19 August 2005.

Borgida, A., Brachman, R. J., McGuinness, D. L. and Resnick, L. A. (1989) 'CLASSIC: A Structural Data Model for Objects', In *ACM SIGMOD International Conference on Management of Data,* Portland, Oregon, pp. 58-67.

DFRWS, (2004), Common Digital Evidence Storage Format, http://www.dfrws.org/CDESF/index.html, Accessed 21 December 2005.

Garfinkel, S.L., Malan, D.J., Dubec, K., Stevens, C.C., Pham, C., 'Disk Imaging with the Advanced Forensics Format, Library and Tools', http://www.simson.net/ref/2006/ifip119_aff.pdf, Accessed 9 Mar 2006.

Goldman, R., Heimerdinger, W., Harp, S., Geib, C., Thomas, V. and Carter, R. (2001) 'Information Modeling for Intrusion Report Aggregation', In *DARPA Information Survivability Conference and Exposition II,* IEEE, Anaheim, CA.

Gruber, TF., (1993), 'A translation approach to portable ontologies' *Knowledge Acquisition* **5**(2) 199–220.

Harmelen, F. v., Patel-Schneider, P.F. and Horrocks, I., (2004), 'Reference description of the DAML+OIL (March 2001) ontology markup language', http://www.daml.org/2001/03/reference.html, Accessed 20 July, 2004.

Klyne, G., & Carrol, J. (eds), (2004), Resource Description Framework (RDF) : Concepts and Abstract Syntax, http://www.w3.org/TR/rdf-concepts/, Accessed 21 December 2005.

Lassila, O. and McGuinness, D. (2001), The Role of Frame-Based Representation on the Semantic Web, *Linköping Electronic Articles in Computer and Information Science*, 6(5).

McGuinness, D. L. (2001), 'Description Logics Emerge from Ivory Towers', In *International Workshop on Description Logics,* Stanford, CA.

Minsky, M., 'A Framework for Representing Knowledge', in *The Psychology of Computer Vision*, McGraw-Hill, New York, 1975.

Moates, R., (1997), URN Syntax, http://www.ietf.org/rfc/rfc2141.txt, Accessed 6 January 2006.

Palmer, G (ed). 'A Road Map for Digital Forensics Research', In *First Digital Forensic Research Workshop*, Ucita, New York, August 7-8 2001.

Raskin, V., Hempelmann, C. F., Triezenberg, K. E. and Nirenburg, S. (2001) 'Ontology in information security: a useful theoretical foundation and methodological tool', In *Workshop on New Security Paradigms,* Cloudcroft, New Mexico.

B Schatz, G Mohay, A Clark, (2004) "*Generalising Event Forensics Across Multiple Domains*" presented at the 2004 Australian Computer Network and Information Forensics Conference (ACNIFC).

Schumacher, M. (2003) 'Security Engineering with Patterns', *Lecture Notes in Computer Science,* vol. 2754.

Seneger, M., (2004), Life Sciences Identifiers LSID Response, http://www.omg.org/cgi-bin/doc?lifesci/2003-12-02 , Accessed 6 January 2006.

Smith, M. K., Welty, C. and McGuinness, D. L., (2004), OWL Web Ontology Language Guide,
http://www.w3.org/TR/owl-guide/, Accessed 20 July 2004.

Turner, P., (2005), Unification of Digital Evidence from Disparate Sources (Digital Evidence Bags), *In 2005 Digital Forensics Research Workshop (DFRWS),* New Orleans, LA, 17–19 August 2005.

Undercoffer, J. L., Joshi, A., Finin, T., and Pinkston, J., 'A Target Centric Ontology for Intrusion Detection: Using DAML+OIL to Classify Intrusive Behaviors' *Knowledge Engineering Review*, January 2004.