Cyber Forensics: Representing and (Im)Proving the Chain of Custody Using the Semantic web

Tamer Fares Gayed, Hakim Lounis
Dépt. d'Informatique
Université du Québec à Montréal
Case postale 8888, succursale Centre-ville, Montréal
QC H3C 3P8, Montréal, Canada
gayed.tamer@courrier.uqam.ca
lounis.hakim@uqam.ca

Moncef Bari
Dépt. d'Éducation et Pédagogie
Université du Québec à Montréal
Case postale 8888, succursale Centre-ville, Montréal
QC H3C 3P8, Montréal, Canada
bari.moncef@uqam.ca

Abstract - Computer/Digital forensic is still in its infancy, but it is a very growing field. It involves extracting evidences from digital device in order to analyze and present them in a court of law to prosecute it. Digital evidences can be easily altered if proper precautions are not taken. A chain of custody (CoC) document is used to demonstrate the road map of how evidences have been copied, transported, and stored throughout the investigation process. With the advent of the digital age, the tangible CoC document needs to undergone a radical transformation from paper to electronic data (*e*-CoC), readable and consumed by machines, and applications. Semantic web is a flexible solution to represent different information, because it provides semantic markup languages for knowledge representation, supported by different vocabularies for provenance information. These features can be exploited to represent the tangible COC document to ensure its trustworthiness and its integrity. Moreover, querying mechanisms can be also incorporated over this represented knowledge to answer different forensic and provenance questions asked by juries concerning the case in hand. Thus, this paper proposes the construction of a framework solution based on the semantic web to represent and consume the forensic and provenance knowledge related to the tangible COC document.

Keywords - Knowledge Representation; Chain of Custody; Provenance Vocabularies; Semantic Web; Resource Description Framework.

I. INTRODUCTION

Computer/Digital forensic is a growing field. It combines computer science concepts including computer architecture, operating systems, file systems, software engineering, and computer networking, as well as legal procedures. At the most basic level, the digital forensic process has three major phases; Extraction, Analysis, and Presentation. Extraction (acquisition) phase saves the state of the digital source (ex: laptop and desktop, computers, mobile phones) and creates an image by saving all digital values so it can be later analyzed. Analysis phase takes the acquired data (ex: file and directory contents and recovering deleted contents) and examines it to identify pieces of evidence, and draws conclusions based on the evidences that were found. During presentation phase, the audience is typically the judges; in this phase, the conclusion and corresponding evidence from the investigation analysis are presented to them [1].

Nevertheless, there exists others forensic process models, each of them relies upon reaching a consensus about how to describe digital forensics and evidences [2][17].

Like any physical evidence, digital evidence needs to be validated for the legal aspects (admissibility) in the court of law. In order for the evidence to be accepted by the court as valid; chain of custody for digital evidence must be kept, or it must be known who exactly, when, and where came into contact with the evidence at each stage of the investigation [3].

The role of players (first responders, investigators, expert witnesses, prosecutors, police officers) concerning CoC is to (im)prove that the evidence has not been altered through all phases of the forensic process. CoC must include documentation containing answers to these questions:

- Who came into contact, handled, and discovered the digital evidence?
- What procedures were performed on the evidence?
- When the digital evidence is discovered, accessed, examined, or transferred?
- Where was digital evidence discovered, collected, handled, stored, and examined?
- Why the evidence was collected?
- How was the digital evidence collected, used, and stored?

Once such questions ("i.e., known as 5Ws and the 1H") are answered for each phase in the forensic process, and players will have a reliable CoC which can be then admitted by the judges' court.

This paper proposes the creation of electronic chain of custody (*e*-CoC) using a semantic web based framework that represent and (im)prove the classical/traditional paper-based CoC during the cyber forensics investigation.

The Knowledge representation concept has been persistent at the centre of the field of Artificial Intelligence (AI) since its founding conference in the mid 50's. This concept described by Davis & al. by five distinct roles [28]. The most important is the definition of knowledge representation as a surrogate for things. This paper suggests the construction of electronic chain of custody (*e*-CoC) using semantic web as a surrogate of the tangible one.

 ISBN: 978-1-61208-218-9

Semantic web will be a flexible solution for this task because it provides semantic markup languages such as Resource Description Framework (RDF), RDF Scheme (RDFS), and Web Ontology Language (OWL) that are used to represent different knowledge.

In addition, the semantic web is rich with different provenance vocabularies [10], such as Dublin Core (DC), Friend of a Friend (FOAF), and Proof Markup Language (PML) that can be used to (im)prove the CoC by answering the 5Ws and the 1H questions.

The remainder of this document is organized as follows: section 2 presents the problem statement encounters the tangible CoC, the related works is presented in section 3, section 4 provides a brief background about the semantic web, the proposed solution is presented in section 5, and finally, conclusion in the last section.

II. PROLBLEM STATEMENT

The continuous growing of devices and software in the field of computing and information technology creates challenges for the cyber forensics science in the volume of data ("i.e., evidences") being investigated. It also increases the need to manage process and present the CoC in order to minimize and facilitate its documentation.

The second issue is related to the interoperability between digital evidence and its CoC documentation. Last works concentrated mainly on the representation and correlation of the digital evidences [24][25] and as an indirect consequence, the improving of the CoC by attempting to replicate the key features of physical evidence bags into Digital Evidence Bags (DEB) [5]. However, the documentation of CoC for digital evidences remains an exhausted task. Knowledge communication between the digital evidence and the information documentation about the evidence, apart from natural language, can create some automation and minimize human's intervention.

The third issue concerns the CoC documents. They must be affixed securely when they are transported from one place to another. This is achieved using a very classical way: seal them in plastic bags ("i.e., together with physical evidence if it exists, such as hard disk, USB, cables, etc."), label them, and sign them into a locked evidence room with the evidences themselves to ensure their integrity.

The fourth issue is about the judges' awareness and understandings are not enough to evaluate, understand, and take the proper decision on the digital evidences related to the case in hand. One solution is to organize a training program to educate the juries the field of Information and Communication Technology (ICT) [6]. From the point of view of the authors, this will not be an easy task to teach juries the ICT concepts. The other solution is to provide a descriptive *e*-CoC using forensic and provenance metadata that the juries can query to find the answers to their questions through these metadata.

The last issue is that the problem is not only to represent the knowledge of the tangible CoC in order to solve the issues mentioned above, but also to express information about where the CoC information came from. Juries can find the answers to their questions on the CoC, but they need also to know the provenance and origins of those answers. Provenance of information is crucial to guarantee the trustworthiness and confidence of the information provided. This paper distinguishes between forensic information and provenance information. Forensic information is responsible to answer the 5Ws and 1H questions related to the case in hand, while provenance information is responsible to answer questions about the origin of answers ("i.e., what information sources were used, when they were updated, how reliable the source was").

III. RELATED WORK

Works related to this paper can be summarized over three dimensions.

The first dimension is the works on improving the CoC. In [22], a conceptual Digital Evidence Management Framework (DEMF) was proposed to implement secure and reliable digital evidence CoC. This framework answered the who, what, why, when, where, and how questions. The 'what' is answered using a fingerprint of evidences. The 'how' is answered using the hash similarity to changes control. The 'who' is answered using the biometric identification and authentication for digital signing. The 'when' is answered using the automatic and trusted time stamping. Finally, the 'where' is answered using the GPS and RFID for geo location.

Another work in [23], discusses the integrity of CoC through the adaptation of hashing algorithm for signing digital evidence put into consideration identity, date, and time of access of digital evidence. The authors provided a valid time stamping provided by a secure third party to sign digital evidence in all stages of the investigation process.

Other published work to improve the CoC is based on a hardware solution. SYPRUS Company provides the Hydra PC solution. It is a PC device that provides an entire securely protected, self contained, and portable device ("i.e., connected to the USB Port") that provides high-assurance cryptographic products to protect the confidentiality, integrity, and non-repudiation of digital evidence with highest-strength cryptographic technology [15]. This solution is considered as an indirect improving of the CoC as it preserves the digital evidences from modification and violation.

Recently, a work for managing and understanding CoC has been provided using an ontological approach. This approach can be used to share common understanding of the structure of the digital forensic domain among different players, among software agents, and between players and software. This approach can also be used to enable the reuse of knowledge in digital investigation process [29].

The second dimension concerns knowledge representation. An attempt was performed to represent the knowledge discovered during the identification and analysis phase of the investigation process [26]. This attempt uses the Universal Modeling Language (UML) for representing knowledge. It is extended to a unified modeling methodology framework (UMMF) to describe and think about planning, performing, and documenting forensics tasks.

 ISBN: 978-1-61208-218-9

The third dimension is about the forensic formats. Over the last few years, different forensic formats were provided.

In 2006, Digital Forensics Research Workshop (DRWS) formed a working group called Common Digital Evidence Storage Format (CDEF) working group for storing digital evidence and associated metadata [12]. CDEF surveyed the following disk image main formats: Advanced Forensics Format (AFF), Encase Expert Witness Format (EWF), Digital Evidence Bag (DEB), gfzip, ProDiscover, and SMART.

Most of these formats can store limited number of metadata, like case name, evidence name, examiner name, date, place, and hash code to assure data integrity [12]. The most commonly used formats are described here.

AFF is defined by Garfinkel et al. in [27] as a disk image container which supports storing arbitrary metadata in single archive, like sector size or device serial number. The EWF format is produced by EnCase's imaging tools. It contains checksums, a hash for verifying the integrity of the contained image, and error information describing bad sectors on the source media.

Later, Tuner's digital evidence bags (DEB) proposed a container for digital crime scene artifacts, metadata, information integrity, access, and usage audit records [5]. However, such format is limited to name/value pairs and makes no provision for attaching semantics to the name. It attempts to replicate key features of physical evidence bags, which are used for traditional evidence capture.

In 2009, Cohen et al. in [4] have observed problems to be corrected in the first version of AFF. They released the AFF4 user specific metadata functionalities. They described the use of distributed evidence management systems AFF4 based on an imaginary company that have offices in two different countries. AFF4 extends the AFF to support multiple data sources, logical evidence, and several others improvements such the support of forensic workflow and the storing of arbitrary metadata. Such work explained that the Resource Description Framework (RDF) [7] resources can be exploited with AFF4 in order to (im)prove the forensics process model.

IV. SEMANTIC WEB

Semantic web is an extension of the current web, designed to represent information in a machine readable format by introducing knowledge representation languages based in XML. The semantic markup language such as Resource Description Framework (RDF), RDF Schema (RDFS) and the web ontology language (OWL) are the languages of the semantic web that are used for knowledge representation.

According to the W3C recommendation [7], RDF is a foundation for encoding, exchange, and reuse of structured metadata. RDF supports the use of standards mechanisms to facilitate the interoperability by integrating separate metadata elements (vocabularies) defined by different resource description communities ("e.g., Dublin Core").

RDF consists of three slots: resource, property, and object. Resources are entities retrieved from the web ("e.g., persons, places, web documents, picture, abstract concepts, etc."). RDF, resources are represented by uniform resource identifier (URI) of which URLs are a subset. Resources have properties (attributes) that admit a certain range of value or attached to be another resource. The object can be literal value or resources.

The main aim of the semantic web is to publish data on the web in a standard structure and manageable format [8]. Tim Berners Lee outlined the principles of publishing data on the web. These principles known as Linked Data Principles ("i.e., LD principles"):

- Use URI as names for things.
- Use HTTP URIs so that people can look up those names.
- When someone looks up a URI, provide useful information using the standards (RDF, SPARQL).
- Include RDF statements that link to other URIs so that they can discover related things.

The Linking Open Data (LOD) project is the most visible project using this technology stack (URLs, HTTP, and RDF) and converting existing open license data on the web, into RDF according to the LOD principles [9]. The LOD project created a shift in the semantic web community. Instead of being concentrated on the ontologies for their own sake and their semantic (languages to represent them, logics for reasoning with them, methods and tools to construct them), it becomes on the web aspects ("i.e., how data is published and consumed on the web").

Semantic web provides provenance vocabularies that enable providers of web data to publish provenance related metadata about their data. Provenance information about a data item is information about the history of the item, starting from its creation, and including information about its origins. Provenance information about the data on the web must comprise the aspects of publishing and accessing the data on the Web. Providing provenance information as linked data requires vocabularies that can be used to describe the different aspects of provenance [11][10][13][14].

V. SOLTUION FRAMEWORK

The solution framework is about the use of the semantic web to represent the CoC using RDF and improve its integrity through different built in provenance vocabularies. Thus, the CoC forensic information and its provenance metadata will be published and consumed on the web.

There exist various vocabularies to describe provenance information with RDF data. The popular standard metadata that can be used in different contexts is the Dublin core metadata terms defined in the RDFS schema [19]. The main goal of consuming this provenance metadata is to assess the trustworthiness and quality of the represented knowledge.

The W3C Provenance Incubator Group detected the needs for provenance in different context. Provenance Interchange Language (PIL) has been considered by the Provenance Interchange Working Group (PIWG) to publish and access provenance using that language. Heterogeneous systems and agents can export and import their provenance information into such a core language and reason over it [30].

 ISBN: 978-1-61208-218-9

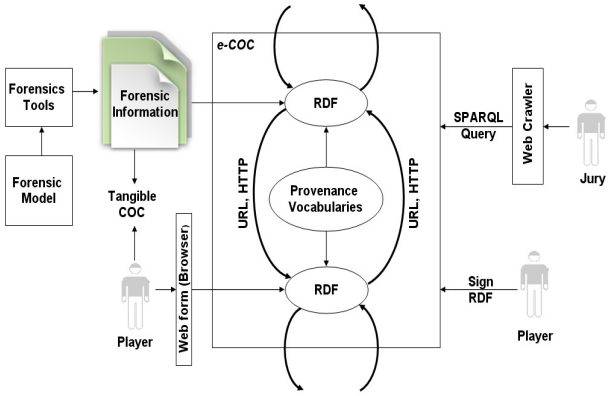

Figure1. Framework for representing and improving CoC using semantic web

As mentioned in Section 2, digital evidence can be stored in open or proprietary formats ("e.g., CDEF, AFF, EWF, DEB, gfzip, ProDiscover, and SMART"). These formats store forensic metadata ("e.g., the sector size and device serial number"). The most advanced format for representing the digital evidence is the AFF4 which is an extension of the AFF to accommodate multiple data sources, logical evidence, arbitrary information, and forensic workflow [16].

The framework proposed in Figure-1 shows that the tangible CoC can be created manually from the output of forensic tools ("e.g., AFF4 or any other format"). AFF4 can be modeled into RDF. Human creates the CoC according to a predefined form determined by the governmental /commercial institution and fill the forms from the forensic information synthesized from the forensic tools. Experience can be used, if necessary, to prune or add some forensic metadata not provided on the current output format.

This framework is generalized to all phases of the forensic investigation. As we have different forensic models with different phases, the framework can be adapted for different phases of different models. A summary of different process models can be found in [2][17]. Each phase for specific forensic process model has its own information: forensic metadata, forensic algorithm, player who came into contact, etc.

The AFF4 can be directly represented using the RDF. Researchers have proposed several solutions on the use of AFF4 and RDF resources to improve digital forensics process model or software. The tangible CoC associated to each phase/digital evidences can be also represented in RDF. Players of each phase can enter the necessary information through a web form interface which is then transformed to RDF triple using web service tool ("e.g., triplify") [18]. The RDF data are supported by different build in provenance vocabularies like DC [19], FOAF [20], and Proof Markup Language (PML) [21]. For example, the provenance terms used by the DC are: dcterms:contributor, dcterms:creator, dctemrs:created, dcterms:modified, dcterms:provenance, which can give the juries information about who contributed, created, modified, the information provided to the court.

FOAF provides classes and properties to describe entities such as persons, organizations, groups, and software agents. The Proof Markup Language describes justifications for results of an answering engine or an inference process.

CoC representation for each phase in RDF data can be linked with another, using the same principle of the LOD project ("i.e., RDF graph/statement can be linked and be navigated using the semantic technology stack: URLs, HTTP, and RDF"). Digital evidence may be also integrated with its CoC information. After representing all information related to the digital evidence and its associated CoC, the player who comes into work in this phase can finally sign his RDF data. Finally, we will have an interlinked RDF based on the LOD principle which represents the whole *e*-CoC of the case in hand.

Juries can use application based on the same idea of the web crawler; they can not only navigate over the interlinked RDF graph/statement, but also, run query through a web application over the represented knowledge using SPARQL query language, and find the necessary semantic answers about their forensic and provenance questions.

The proposed framework can provide solutions to the issues mentioned in Section 2. Representation of the tangible CoC knowledge to RDF facilitates the management and processing because it is a machine readable form (first issue). It is also interoperable; digital evidence representation and its CoC description can be unified together under the same framework ("i.e., RDF"). Also, each player comes into role can secure ("i.e., using cryptographic approaches") and sign his RDF data ("i.e., using digital signature") to ensure the integrity and identity, respectively (second issue and third issue). On the other hand, juries can consume and navigate over the interlinked RDF data which present the whole and detailed information about the history of evidence from its collection to its presentation in the court (fourth issue). Provenance vocabularies can also be used to provide extra and descriptive metadata beyond the forensic metadata provided by the forensic tools (last issue).

VI. CONCLUSION

This paper proposes the construction of a semantic web based framework to represent and (im)prove tangible chain of custody using RDF and provenance vocabularies. After the definition and analysis of all related information (metadata) for each phase in a selected forensic process ("i.e., source will be the human experience and forensic tools output"), we will focus on the conversion and representation of tangible COCs information into interlinked RDF (*e*-COCs). This representation will contain forensic and provenance metadata (built-in/custom) related to the case in hand. The last phase will be the construction of a web interface that let the juries consume and query these interlinked RDF data in order to answer all questions related to the COCs of evidences and their provenances.

 ISBN: 978-1-61208-218-9

REFERENCES

[1] Erin Kenneally. Gatekeeping Out Of The Box: Open Source Software As A Mechanism To Assess Reliability For Digital Evidence. Virginia Journal of Law and Technology. Vol 6, Issue 3, Fall 2001.

[2] Michael W. Andrew "Defining a Process Model for Forensic Analysis of Digital Devices and Storage Media" Proceedings of the 2nd International Workshop on Systematic Approaches to Digital Forensic Engineering SADFE 2007

[3] Ćosić, J., Bača, M. Do we have a full control over integrity in digital evidence life cycle, Proceedings of ITI 2010, 32nd International Conference on Information Technology Interfaces, Dubrovnik/Cavtat, pp. 429-434, 2010

[4] Cohen, M.; Garfinkel, S.; Schatz, B. Extending the advanced forensic format to accommodate multiple data sources, logical evidence, arbitrary information and forensic workflow. Digital Investigation, 2009. S57-S.

[5] Turner, P. Unification of Digital Evidence from Disparate Sources (Digital Evidence Bags). In 5th DFRW. 2004. New Orleans

[6] Judges' Awareness, Understanding and Application of Digital Evidence, Phd Thesis in computer technology in Education, Graduate school of computer and information sciences, Nova Southeastern University, 2010

[7] RDF: Model and Syntax Specification. W3C recommendation, 22 Februrary 1999, www.w3.org/TR/REC-rdf-syntax-19990222/1999

[8] The semantic web, Linked and Open Data, A Briefing paper By Lorna M. Campbell and Sheilla MacNeill, June 2010, JISC CETIS

[9] Christian Bizer, Tom Heath, Tim Berners-Lee: Linked Data - The Story So Far. Int. J. Semantic Web Inf. Syst. 5(3): 1-22 (2009)

[10] Olaf Hartig: Provenance Information in the Web of Data. In Proceedings of the Linked Data on the Web (LDOW) Workshop at the World Wide Web Conference (WWW), Madrid, Spain, Apr. 2009

[11] Olaf Hartig and Jun Zhao: Publishing and Consuming Provenance Metadata on the Web of Linked Data. In Proceedings of the 3rd International Provenance and Annotation Workshop (IPAW), Troy, New York, USA, June 2010

[12] CDESF. Common Digital Evidence Format. 2004 [Viewed 21 December 2005]; Available from: http://www.dfrws.org/CDESF/index.html

[13] [Olaf Hartig and Jun Zhao: Using Web Data Provenance for Quality Assessment. In Proceedings of the 1st Int. Workshop on the Role of Semantic Web in Provenance Management (SWPM) at ISWC, Washington, DC, USA, October 2009 Download PDF

[14] Olaf Hartig, Jun Zhao, and Hannes Mühleisen: Automatic Integration of Metadata into the Web of Linked Data (Demonstration Proposal). In Proceedings of the 2nd Workshop on Trust and Privacy on the Social and Semantic Web (SPOT) at ESWC, Heraklion, Greece, May 2010

[15] Solvinghe digital Chain of Custody Problem, SPYRUS, Trusted Mobility Solutions, © Copyright 2010

[16] M. I. Cohen, Simson Garfinkel and Bradley Schatz, Extending the Advanced Forensic Format to accommodate Multiple Data Sources, Logical Evidence, Arbitrary Information and Forensic Workflow, DFRWS 2009, Montreal, Canada

[17] MD Köhn, JHP Eloff and MS Olivier, "UML Modeling of Digital Forensic Process Models (DFPMs)," in HS Venter, MM Eloff, JHP Eloff and L Labuschagne (eds), Proceedings of the ISSA 2008 Innovative Minds Conference, Johannesburg, South Africa, July 2008 (Published electronically)

[18] http://triplify.org/Overview

[19] http://dublincore.org/

[20] http://www.foaf-project.org/

[21] P. P. da Silva, D. L. McGuinness, and R. Fikes. A Proof Markup Language for Semantic Web Services. Information Systems, 31(4-5):381–395, June 2006

[22] Ćosić, J., Bača, M. (2010) A Framework to (Im)Prove Chain of Custody in Digital Investigation Process, Proceedings of the 21st Central European Conference on Information and Intelligent Systems, pp. 435-438, Varaždin, Croatia

[23] Ćosić, J., Bača, M. (2010) (Im)proving chain of custody and digital evidence integrity with timestamp, MIPRO, 33. međunarodni skup za informacijsku i komunikacijsku tehnologiju, elektroniku i mikroelektroniku, Opatija, 171-175

[24] Schatz, B., Mohay, G. And Clark, A. (2004) 'Rich Event Representation for computer Forensics', Proceedings of the 2004 Asia Pacific Industrial Engineering and Management System

[25] Schatz,B., Mohay, G. and Clark, A.(2004) 'Generalising Event Forensics Across Multiple domains' Proceedings of the 2004 Australian Computer Network and Information Forensics Conference (ACNIFC 2004), Perth, Australia.

[26] Bogen, A. and D.Dampier. Knoweldge discovery and experince modeling in computer forensics media analysis.In international Symposium on Information and Communication Technologies.2004: Trinity College Dublin

[27] Garfinkel, S.L., D.J. Malan, K.-A. Dubec, C.C. Stevens, and C. Pham, Disk Imaging with the Advanced Forensics Format, library and Tools. Advances in Digital Forensics (2nd Annual IFIP WG 11.9 International Conference on Digital Forensics), 2006

[28] Davis, R.,H. Shrobe, and P. Szolovits. What is a knowledge representation? AI Magazine, 1993. 14(1): p.17-3

[29] Jasmin Ćosić, Zoran Ćosić, Miroslav Bača, An Ontological Approach to Study and Manage Digital Chain of Custody of Digital Evidence, Journal of Information and Organizational Sciences (JIOS) e-ISSN: 1846-9418, Vol. 35, No.1 (2011), PP.1-13

[30] Provenance Interchange Working group http://www.w3.org/2011/01/prov-wg-charter