Cyber Forensics: Representing and Managing Tangible Chain of Custody Using the Linked Data Principles

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

Moncef Bari
Dépt. de Didactique
Université du Québec à Montréal
Succursale Centre-ville, H3C 3P8,
Montréal, Canada
bari.moncef@uqam.ca

Rafek Nicolas
Service Technique Global
IBM
1360 René Lévesque, H3G 2W6
Montréal, Canada
rnicolas@ca.ibm.com

***Abstract*—Tangible Chain of Custody (*CoC*) in cyber forensics (*CF*) is a document accompanying digital evidences. It records all information related to the evidences at each phase of the forensics investigation process in order to improve and prosecute them in a court of law. Because a digital evidence can be easily altered and loses its value, the *CoC* plays a vital role in the digital investigation by demonstrating the road map of Who exactly, When, Where, Why, What and How came into contact with the digital evidence. With the advent of the digital age, the tangible *CoC* document needs to undergo a radical transformation from paper to electronic data (*e-CoC*). This *e-CoC* will be readable, and consumed by computers. The semantic web is a fertile land to represent and manage the tangible *CoC* because it uses web principles known as Linked Data Principles (LDP), which provide useful information in Resource Description Framework (RDF) upon Unified Resource Identifier (URI) resolution. These principles are used to publish data publicly on the web and provide a standard framework that allows such data to be shared, and consumed in a machine readable format. This paper provides a framework explaining how these principles are applied to represent the chain of custodies and used only by actors in each forensics process, in order to be consumed at the end by the jury in a court of law. This paper also illustrates this idea by giving an example of the authentication phase imported from the Kruse forensics process.**

Keywords-Chain of Custody; Knowledge Representation; Provenance Vocabularies; Forensic Models; Semantic Web; Linked Data Principles; Public Key Infrastructure.

I. INTRODUCTION

Digital forensic is a technique for acquiring, preserving, examining, analyzing and presenting digital evidence in accordance with evidentiary rules and legal standards. One of the most essential parts of the digital investigation process is the chain of custody (*CoC*) [18]. *CoC* is a chronological document that accompanies all digital evidence in order to avoid later allegations of tampering with such evidences. *CoC* provides useful information about the digital evidences studied using a certain forensic process by answering 5 W and 1 H questions. The 5 W are the When, Who, Where, Why, What and the 1 H is the How. Because cyber forensic is a daily growing field and requires the accommodation on the continuous changes of digital technologies (i.e., concurrency with the knowledge management), the tangible *CoC* information also needs to undergo a radical transformation from paper to electronic data (*e-CoC*), readable and consumed by the computers. This transformation will be achieved through the support of different technologies used by the semantic web [3][7[8].

Today, the semantic web is the web of data, which is not just concentrated for the interrelation between web documents but also between the raw data within these documents. This data interrelation is based on four aspects known as the linked data principles (LDP). In 2006, Berners Lee outlined a set of rules [3][10] for publishing data on the web using these principles. They are used to apply general architecture of the World Wide Web [11] and explain that the data (content/resources) should be related to one another just as documents are already:

- Use Unified Resource Identifier (URI) as names for things and they are used as globally unique identification mechanism [12].
- Use Hyper Text Transfer Protocol (HTTP) as universal access mechanism so that people can look up those names [13].
- When someone looks up a URI, provide useful information using the standards (Resource Description Framework, SPARQL).
- Include RDF statements that link to other URIs so that they can discover related things (i.e., people locations, or abstract concepts).

Publishing data in a structured way can facilitate the consumption of such data and help its consumer to take the proper decision.

This paper resumes the task provided in [1][2]. These works provided a novel framework that uses the LDP to represent the tangible *CoC* in order to be consumed by the juries in the court of law. The framework provided in these works was abstract. This paper elaborates it into a set of layers and explains in detail the performed task in each of them by giving an example of the authentication phase imported from Kruse model [17]. This is the first

 ISBN: 978-1-61208-273-8

work combining in the same framework the following disciplines: cyber forensics, semantic web, provenance of information, and security. We present how the semantic web and its technologies is a fertile land to represent the tangible *CoC* knowledge using the principles of the linked data and how this data is controlled/managed, and consumed only by the role player at each forensic phase and the jury of the court, respectively.

This work expands the framework provided in [1][2] with a security approach such as Public Key Infrastructure (PKI)[15][58] to ensure the identity and the authentication of each role player participating in the investigation process. Thus, the security approach arises in this context to protect and foster the published information related to the case in hand from unauthorized access.

This work also argues against the solution proposed in [65] concerning the judges' awareness and understanding of the digital evidence. This solution seeks to educate the juries about the field of Information and Communication Technology (ICT). However, the aim of this paper is the construction of an assistant system, offering the ability to juries to navigate, discover (dereference) and execute different queries on the represented information. This idea is underlined using code examples describing different aspects related to the representation of the chain of custody using LDP (e.g., Figure 2,3 and 4 are generated from RDF/XML codes using [66]). However, the PKI approach provided in this paper is theoretically presented and will be implemented in later publications. All concepts and components of this future system are discussed through the solution framework in Section 5.

The organization of this paper is as follows: the next section discuses the state of the art of the semantic web and the web of data. Section 3 outlines the reasons why the authors used LDP for representing the *CoC*. Section 4 provides a quick view about the forensics models and describes the tasks that are performed in the authentication phase and how forensics terms are specified in order to be later represented using LDP. Section 5 explains the solution framework in detail from the representation of data to its consumption by juries and explains how such data is controlled by only the authorized actors (i.e., role players and juries who participated in the court case). Finally, the last section concludes and summarizes this work, and presents the future extensions of the proposed framework. The related works in this paper are not presented in a separate section. However, they are mentioned in detail in [1][2], and through different references along the paper, especially in the explanation for each layer.

II. STATE OF THE ART: SEMANTIC WEB AND THE WEB OF DATA

Semantic web is an extension of the current web (i.e., from document to data) [7][8], designed to represent information in a machine readable format by introducing RDF model [16] to describe the meaning of data and allows them to be shared on the web in a flexible way. The classical way for publishing documents on the web is just naming these documents using URI and hypertext links. This fact allows the consumer to navigate over the information on the web using a web browser application and querying the information by typing keywords in a search engine that is working using the support of HTTP protocol. This is called the web of documents.

With the same analogy, entities and contents (data) within documents can be linked between each others using typed linked and with the same principles used by the web (i.e., web aspects). This is called the web of data. The Linking Open Data (LOD) project is the most visible project using this technology stack (URLs, HTTP, and RDF) and converts existing open license data on the web into RDF according to the LDP [3][10]. The LOD project created a shift in the semantic web community. Instead the concern was on the ontologies for their own sake and semantic, it becomes on the web aspects (how to publish and consume data on the web). Ontologies are used then to foster and serve the semantic interoperability between parts that want to exchange such data. There are known as lightweight ontologies [23] that use the full advantages of semantic web technologies, minimum OWL constructs, and reuse existing RDF vocabularies wherever possible.

According to the W3C recommendation [16], RDF is a foundation for encoding, exchange, and reuse of structured metadata. It can be serialized using different languages (e.g., RDF/XML [40], Turtle [41], RDFa [42], N-Triples [43], N3 [44]). RDF consists of three slots called triples: resource, property, and object. Also, resources are entities retrieved from the web (e.g., persons, places, web documents, pictures, abstract concepts, etc,). RDF resources are represented by uniform resource identifiers (URIs), of which URLs are a subset. Resources have properties (attributes) that admit a certain range of values or that are attached to another resource. The object can be a literal value or a resource.

While RDF provides the model and syntax for describing resources, it does not define the meaning of those resources. That is where other technologies such as RDF Schema (RDFS) come in. RDFS specifies extensions to RDF that are used to define the common vocabularies in RDF metadata statement and enables specification of schema knowledge. It develops classes for both resources and properties. However, RDFS is limited to a subclass hierarchy and a property hierarchy with domain and range definitions of these properties. RDFS limitations are: range restrictions, disability of expressing disjoitness between classes, combination between classes, cardinality restriction, and characteristics of properties [22].

The work presented in this paper is a framework based on the RDF model and its related vocabularies, managed by different web aspects (LDP), for representing and managing the tangible chains of custody of digital

 ISBN: 978-1-61208-273-8

investigations. Next section provides the reasons why LDPs are suitable and useful to represent the digital investigation *CoCs*.

III. ADVANTAGES OF USING LDP FOR REPRESENTING *CoC*

Knowledge representation has been persistent at the centre of the field of Artificial Intelligence (AI) since its founding conference in the mid 50's. This concept is described by Davis et al. [33] through five distinct roles. The most important is the definition of knowledge representation as a surrogate for things. Thus, before providing the solution framework, we decided to underline why linked data is selected to represent the tangible *CoC* in cyber forensics. Thus, this section lists all the advantages and the common features of using linked data to represent the *CoC* for cyber forensics:

1. *CoC* and LDP are metaphors for each others. The nature of *CoC* is characterized by interrelation/dependency of information between different phases of the forensics process. Each phase can lead to another one. This interrelation fact is the basic idea over which the linked data is published, discoverable, and significantly navigated using RDF links. RDF links in LDP will not be used only to relate the different forensic phase together, but it can also assert connection between the entities described in each forensic phase. Also, RDF typed links enable the data publisher (role player) to state explicitly the nature of connection between different entities in different and also same phases, which is not the case with the un-typed hyperlinks used in HTML.
2. Linked data enables links to be set between items/entities in different data sources using common data model (RDF) and web standards (HTTP, URI, and URL). As well, if the *CoC* is represented using the LDP, the items/entities in different phases can be also linked together in forensics process. This will generate a space over which different generic applications can be implemented:
   - *Browsing applications*: enable juries to view data from one phase and then follow RDF links within the data to other phases in the forensics process.
   - *Search engines*: juries can crawl the different phases of the forensics process and provide sophisticated queries.
3. Linked data applications that are planned to be used by juries, will be able to translate any data even it is represented with unknown vocabulary. This can be realized using two methodologies. First, by making the URIs that identify vocabulary terms dereferenceable (i.e., it means that HTTP clients can look up the URI using the HTTP protocol and retrieve a description of the resource that is identified by the URI) so that the client applications can look up the terms, which are defined using RDFS and OWL. Secondly, by publishing mappings between terms from different vocabularies in the form of RDF links. So, for any new term definition, the consumption applications are able to provide and retrieve for the juries extra information describing the provided data.
4. Nowadays, RDFS [34] and OWL [35] are partially adopted on the web of data. Both are used to provide vocabularies for describing conceptual models in terms of classes and their properties (definition of proprietary terms). RDFS vocabularies consist of class *rdfs:class* and property *rdf:property* definitions, which allow the subsumption relationships between terms. This option is useful for juries to infer more information from the data in hand using different reasoning engines. For example, RDFS uses a set of relational primitives (e.g., *rdfs:subclassof*, *rdfs:subpropertyof*, *rdfs:domain,* and *rdfs:range* that can be used to define rules that allow additional information to be inferred from RDF graphs). Also, OWL extends the expressivity of RDFS with additional modeling primitives that provide mapping between property terms and class terms, at the level of equivalency or inversion (e.g., *owl: equivalentProperty*, *owl:equivalentClass*, *owl:inverseof*).
   RDFS and OWL are not yet fully adopted on LDP, but soon the full adaptation will be achieved. This will be a great advantage to add more property and class terms to the semantic dimension of the linked data, and therefore, provide useful and descriptive information [4] [5].
5. Representing *CoC* data using LDP will be enriched with different vocabularies such as Dublin Core (DC) [30], Friend of a Friend (FOAF) [31], and Semantic Web Publishing (SWP). Also, vocabulary links is one type of RDF links that can be used to point from data to the definitions of the vocabulary terms, which are used to represent the data, as well as from these definitions of related terms into other vocabularies. This mixture is called schema in the linked data; it is a mixture of distinct terms from different vocabularies to publish the data in question. This mixture may include terms from widely used vocabularies as well as proprietary terms. Thus, we can have several vocabulary terms to represent the forensics data and make it self descriptive (using the 2 methodologies mentioned in point 3) and enable linked data applications to integrate the data across vocabularies and enrich the data being published.
6. Juries need to avoid heterogeneity and contradictions about the information, which are provided to them in the court in order to take the proper decision. Linked data try to avoid heterogeneity by advocating the reuse of terms from widely deployed vocabularies

(same agreement of ontology). LDP is then useful to represent this type of information.

7. As mentioned at point 1, a forensics process contains several phases which are dependent and related to each others. Each entity is identified by a URI namespace to which it belongs. An entity appearing in a phase may be the same entity in another phase. The result is multiple URIs identifying the same entity. These URIs are called URI aliases. In this case, linked data rely on setting RDF links between URI aliases using the *owl#sameas* that connect these URIs to refer to the same entity. The advantages of this option in *CoC* representation are:
   - *Social function*: investigation process is a common task between different players. The descriptions of the same resource provided by different players allow different views and opinions to be expressed.
   - *Traceability*: using different URIs for the same entity allows juries that use the *CoC* published data to know what a particular player in the investigation process has to say about a specific entity of the case in hand.

   Same thing occurs not only at the level of URI but also at the level of terms. Players of the forensics process may discover at a later point that a property vocabulary contains the same term as the built in one. Players could relate both terms, stating that both terms actually refer to the same concept using the OWL (*owl: equivalentClass*, *owl: equivalentProperty)* and RDFS vocabularies (*rdfs: subclassOf*, *rdfs:subProperyOf*).
8. Provenance metadata can also be published and consumed on the web of data [6]. Such metadata provide also an answer to six questions, but at the level of the data origin (i.e., Who published/created the data, Where this data is initially published/created, What is the published data, When/Why the data is published, and How the data is published). These vocabularies can be used concurrently with the forensics data, to describe their provenance and complement the missing answers related to the forensic investigation.

All these advantages are motivations for using LDP to represent the tangible *CoC* in cyber forensics. Next section explains the first step of representing a forensic phase using LDP. This is illustrated through the authentication phase imported from Kruse [17] forensics model.

IV. DIGITAL FORENSICS PROCESS MODELS

Different Digital Forensics Process Models (DFPM) has been proposed since 2000 (e.g., Kruse [17], the United State Department of Justice (USDOJ) [18], Casey [19], Digital Forensics Research Workshop (DFRW) [25], and Ciarhuin [21]) to assist the players of investigations process reaching conclusions upon completion of the investigation.

TABLE I. THE *CoC* OF THE AUTHENTICATION PHASE

Questions	Subject	Terms (Custom / Built in)
Who	The Role Player	Investigator
What	Verify the integrity of the acquired data	Evidence
Why	Ensure the completeness and integrity of data	Hash
Where	The place that this task took place	Location
When	The time that this task took place	Date
How	Procedures utilized to perform the checking task	Algorithm

Investigation models are numerous. Many works were provided to explain and compare such models [18][19][21][26][29]. Nevertheless, all works provided in the forensics process globalize the 5 W and 1 H questions once over the whole forensics process. However, these questions must be posed over each phase of the forensics process, separately, since each question in a forensic phase is not the same for another phase (i.e., 'What' question, of the collection phase is not the same as the 'What' for the identification phase). For example, the Kruse model has 3 forensics phases, thus, it should have 3 different *CoCs*.

Furthermore, some phases from different forensics models may have unique technical requirement but they differ only on their names [24]. The work presented by Yussof et al. [26] underlines 46 phases from 15 selected investigation models that have been produced throughout 1995 to 2010, and then identifies the commonly shared processes between these models.

The first step of representing *CoC* for a phase in a forensics process is to identify the essential terms that can be used to describe this phase. The identification of terms is achieved through the descriptions of different processes and tasks performed within this phase. For instance, the essential task of Authentication in the Kruse model, is to verify the integrity of acquired/extracted data. The verification of integrity is to ensure that the information presented is complete and has not been altered in any authorized/unauthorized manner, since the time it was extracted, transmitted, and stored by an authorized source [27].

The role player of this task is called the investigator. He is responsible (Who) to check the integrity of the extracted evidence (What), by comparing the checksum generated from hashing algorithm (How) (e.g., CRC-Cyclic Redundancy, Cryptographic Hash function such as MD5 and SHA1), in order to ensure the completeness and the integrity of data (Why), by comparing the hash/checksum of the original data with the hash/checksum of the copied data [28]. If the checksum of the original data is not the same as the checksum of the copied one, the data is then altered. If not, it keeps always its integrity. The *CoC* should also record the date/time (When) and the location/machine (Where) this task took

 ISBN: 978-1-61208-273-8

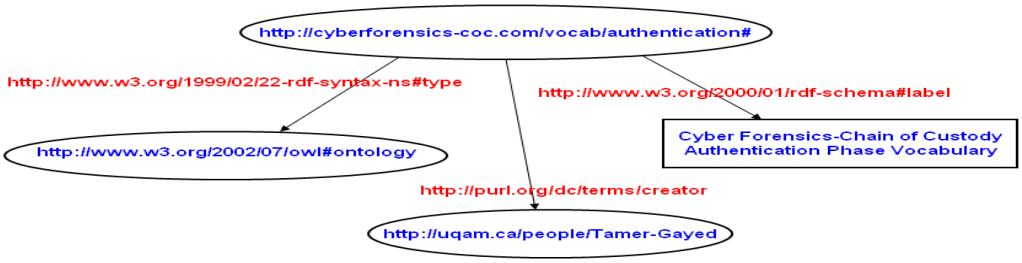

Figure 1. Definition of Authentication Ontology

place. Table I defines the essential terms used to describe the authentication phase. Definition of such terms is explained in next section. Figure 4 provides an example of this chain of custody.

V. CYBER FORENSICS *CoC* FRAMEWORK

CF-CoC framework provided in this paper (see figure 2) explains how tangible *CoCs* are represented.

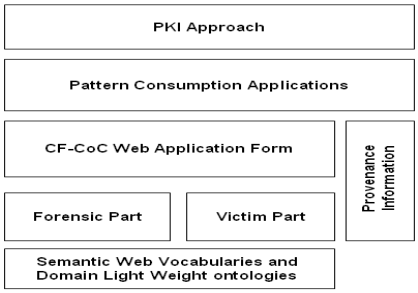

Figure 2. CF-*CoC* Framework

*CoC*s are described using RDF model, which integrate well defined vocabularies in the semantic web. *CoCs* contains mainly forensics information that needs to be described with new proprietary terms.

Creating the vocabularies and terms of each phase is performed through the construction of lightweight ontology using RDFS and the Web Ontology Language OWL (see Figure 3). These terms are used to represent and describe different information related to a victim of any cyber crime (e.g., child pornography, pedophilia, prostitution, blackmail, extortion, harassment, defamation, forgery, spam, theft, etc), and the forensics part, who is responsible to investigate and provide the result of the investigation process. Such information together forms the *CoC*.

Each forensics phase has its own *CoC*. In each phase, the player role is responsible to prepare and create the *CoC* of the phase in which he worked in. Each player role constructs his *CoC* using a web form that allows the player to import different resources (i.e., from the victim and forensic parts) or create new triples using well predefined/custom vocabularies. The results will be a set of interrelated triples describing all phases in the forensics process. These triples are consumed by juries in a court of law, using different patterns of consumption applications on the semantic web. Along this scenario, provenance dimension (metadata/model) is also integrated with the forensics data to answer all questions related to the origins of this data. Published data and their consumption will be published and consumed by the authorized people who can to work on the current cyber crime case. PKI is used to ensure identities and authorization of each role player. Next sub-sections will describe each layer in detail.

A. *Semantic web vocabularies and domain light weight ontologies*

All the vocabularies of the semantic web can be divided into two main categories: built-in vocabularies and custom (property terms) vocabularies. The latter are created using the former (known as light weight ontology) upon the needs to describe particular domain (i.e., cyber forensics), when the former do not provide all terms that are needed to publish/describe the content of a data set. In this context, the data set is the chains of custodies. The custom vocabulary will be an ontology created for each phase in forensics models. Each ontology contains a set of terms (classes or properties) describing the forensic phase that this ontology represents. Player records all information that he encountered in the forensics phase through the support of these property terms and the terms of the built-in vocabulary.

Some examples of common well established vocabularies are RDFS, OWL, Dublin Core Metadata Initiative (DCMI) vocabulary [30], Friend-of-a-friend (FOAF) vocabulary [36], Semantically-interlinked Online Communities (SIOC) vocabulary [37], and Description of a Project (DOAP) Vocabulary [38]. For instance, Table I contains terms that should be defined (e.g., investigator term is defined by the investigated verb) and terms that are predefined (e.g., date, evidence).

The name space defined for this phase is given in [64]. The *cfcoc-auth:investigator* and *cfcoc-auth:investigated* terms are defined in this ontology (figure 3) using well defined terms (e.g., *foaf: person* and *owl:ObjectProperty* respectively). Some principles are provided in [14] describing how to select existing vocabularies and how to develop new terms. Some tools

 ISBN: 978-1-61208-273-8

Figure 3. Definition of investigator and investigated term (Class and Property)

can be used in the development of new terms such as Neologism [45], Protégé [46], TopBraid Composer [47].

B. Victim and Forensics Parts

This section describes the mechanism of how the resources of victim and forensics parts are represented on the web. The essential thing to publish data is to have a unique domain/namespace minted by unique URL owned by the publisher. URI HTPP is used to relate, and identify real-world objects and abstract concepts, thereby maximizing the discoverability of more data. Thus, URIs need to be dereferenceable to identify real objects (i.e., objects and documents should not be confused between each others). Therefore, a common practice called contents negotiating is used by an HTTP mechanism [13] that sends HTTP headers with each request to indicate what kind of documents they prefer. Servers can inspect then these headers and select an appropriate representation of resources (HTML document or RDF document). Content negotiation uses two different types of URIs:

- **303 URIs** (known as 303 redirect): server used to redirect the client request to see another URI of a web document, which describes the concept in question.
- **Hash URIs:** to avoid two http requests used by the 303 URIs. Its format contains the base part of the URI and a fragment identifier separated from the base by a hash symbol. When a client requests hash URI, the fragment part is stripped off before requesting the URI from the server. This means that the hash URI does not necessarily identify a web document and can be used to identify real-world objects.

Using first type of URI, victim or forensics part could publish on their servers the description of any concepts (e.g., real world object: persons) using two types of representations: HTML document containing a human-readable representation about a concept, and RDF document about the same concept. We will imagine here a victim company called Digital Test that wants to publish information about an investigator (i.e., we assume that the company has forensics department). This company can use 3 different patterns to describe the concept employee (the following can be applied to any resource):

- In [59], the URI identifying the person Jean-Pierre.
- In [60], the URI identifying the RDF/XML document describing Jean-Pierre.
- In [61], the URI identifying the HTML document describing Jean-Pierre.

Using the second type of URI, forensic or victim part can define different vocabulary terms in order to describe their profile in data published on the web. They may use also the Hash URI to serve an RDF/XML file containing the definitions of all these vocabulary terms. For example, Digital Test may assign the URL in [62] to the file, which contains a custom vocabulary describing different employee's concepts and appends fragment identifiers (using #), to the file's URI in order to identify the different vocabulary terms.

Furthermore, the forensic part will publish different resources and integrate forensic data resulted from the investigation process. This can be realized using the Advanced Forensic Format (AFF4). It is an open format for the storage and processing of digital evidences. Its design adopts a scheme of globally unique identifiers (URN) for identifying and referring to all evidences [32]. The great advantage of this format is representing different forensics metadata in the form of RDF triples (subject, predicate, and value), where the subject is the URN of the object the statement is made about and the predicate (e.g., datelogin, datelogout, evidenceid, affiliation, etc.) can be any arbitrary attribute, which can be used to store any object in the AFF4 universe. Thus, any information of victim and forensics part related to their profiles or forensics data can be easily represented and integrated together in a unified RDF model. Figure 4 shows an example of how the custom terms (e.g., investigator) are defined using lightweight ontology. Victim resources (e.g., who: Jean-Pierre defined by Digital Test), and forensics resources (e.g., What: evidence, Why: hash, Where: location defined in the AFF4), and terms from the DC vocabulary (e.g., When: date) are all integrated together in a unified framework answering the six questions of the authentication phase.

 ISBN: 978-1-61208-273-8

Figure 4. *CoC* for the authentication phase

The big difference between the terms defined in the AFF4 format and the customs terms is that the latter are URI resources that can be dereferenced while the former are a set of literals that are terminals. The higher the number of dereferenceable terms, the more the data provided to juries are descriptive.

C. CF-CoC Web Application Form

This section answers the question of how RDF triples are generated and connected to be later consumed and crawled by juries. In order to create manageable and electronic linked data *CoC*, the *CF-CoC* HTML web application form should be designed to:

- Import resources from the forensics parts (e.g., role players' profiles, results of forensic investigation).
- Import resources from the victim part (departments names, employees names, machines IDs, etc.).
- Create and describe resources by the support of, (i) existing terms imported from well established vocabularies, and (ii) new terms imported from custom vocabulary created to describe the *CoC* for each forensics phase.
- Add provenance metadata to forensics data. (i.e., provenance vocabularies are used to prove the origin of the imported data and improve their representation).

The work in this part is divided into two main tasks: entities identification, and the description of such entities with RDF using predefined or custom terms.

Entities identification is divided into two subtasks. Whether is to create/mint a new URL for a new entity or to import predefined URI for an existing entity from the victim or forensics part. *CF-CoC* will have its own domain name (cyberforensics-coc.com). This domain allows only each authorized player in the forensic process to use the namespace of this domain. After running the web server of this domain name at [63], players that are allowed to participate in the forensics process are free to mint URI in this namespace to use as names for things they want to describe. For example, if the investigator player of the authentication phase wishes to mint URIs to indentify different new entities in this forensics phase, he/she does extend the URI assigned to the domain URI in [63].

On the other hand, each player will create different types of triples that describe each resource of the forensics process with literals, linked to other resources, linked from other resources, and the resource itself using provenance metadata. The links that connect different phases to each others can be incoming or outcoming links. If a RDF triple links a resource A in a forensics phase to a resource B in another phase, the document describing B should include this triple. This triple is called incoming link to B and allows to navigate back to A even if the latter is not an object of any triples in the description of B. Also, the triple in the document describing A is called outcoming link.

Two languages can be used to generate RDF models from the data entered in the web form application: script languages or mapping ones. In both cases, the data is posted first in a relational database, and queried or mapped later. Script languages, for example PHP, can generate linked data in RDF/XML format by the help of the ARC library for working with RDF in PHP [48]. Mapping languages, for example D2RQ [49][50], map database contents into RDF vocabularies and OWL ontologies, and allow RDF data to be browsed and searched.

D. Pattern Consumption Application

Linked data is a style of publishing data that makes it easy to interlink, discover, and consume them on the semantic web. The idea explained in this section underlines how juries can consume (navigate and search) data at run-time, among different meaningful collection of triples published by role players.

As mentioned in the last section, the first way to publish linked data on the web is to make URIs that identifies data items dereferenceable into RDF descriptions. Three main patterns can be used by juries to consume this information of the *CoC* published by role

 ISBN: 978-1-61208-273-8

players: browsing, searching, and querying. Browsing is like traditional web browsers that allow users to navigate between HTML pages. Same idea is applied for linked data, but the browsing is performed through the navigation over different resources, by following RDF links and downloads them from a separate URL (e.g., RDF browsers such as Disco, Tabulator, or OpenLink) [51].

RDF crawlers are also developed to crawl linked data from the web by following RDF links. Crawling linked data is a search using a keyword related to the item in which juries are interested (e.g., SWSE and Swoogle). Juries can also perform extra search filtering using query agents. This type of searching is performed when SPARQL endpoints are installed, which allow expressive queries to be asked against the dataset. Furthermore, a void vocabulary (vocabulary of interlinked datasets) [39] contains a set of instructions that enables the discovery and usage of linked datasets through dereferenceable HTTP URIs (navigation) or SPARQL endpoints (searching), using SPARQL (*void:sparqlendpoint*) or URI protocol.

E. Provenance Metadata

Provenance of information is an essential ingredient of any tangible *CoC* quality. The ability to track the origin of data is a key component in building trustworthy, which is required for the admissibility of digital evidences. Classically, the provenance information about Who created and published the data and How the data is published, provides the means for quality assessment. Such information can be queried and consumed to identify also the outdated information. *CoC* data source should include provenance metadata together with the forensics data. Such metadata can be used to give juries data clarity about the provenance, completeness, and timeliness of forensics information and to strength the provenance dimension for the published data.

Provenance information can be integrated within the forensics data using three different methods. The first method is using the provenance vocabularies of the semantic web. The second one is to use open provenance model [31], and the last method is by exploiting named graphs for RDF triples, to add provenance metadata about each group of triples.

A widely deployed provenance vocabulary is Dublin Core [30]. For example, this vocabulary contains different predicates that can provide extra information related to the forensics data like the *dc:creator*, *dc:publisher,* and *dc:date*. The objects of these predicates can be represented by URI (e.g., deferenceable resources like the investigator Jean-Pierre) or literal/terminal (date) (see figure 4), identifying such objects. Another provenance vocabularies provided in [52][53], describe how provenance metadata can be created and accessed on the web of data. These vocabularies assess the quality and trustworthiness of the published data.

Open Provenance Model (OPM) provides an alternative and more expressive vocabulary that describes provenance in terms of agents, artifacts, and processes [54]. An extension of this work is the Open Provenance Model Vocabulary (OPMV), provided in [55]; it implements the OPM model using lightweight OWL. OPVM can be used also with other provenance vocabularies such as Dublin Core, FOAF, and the provenance vocabulary.

While many authors advocate the use of semantic web technologies (i.e., vocabularies, Light weight ontologies), Carroll et al. [56] take the opposite view and proposed named graphs as an entity denoting a collection of triples, which can be annotated with relevant provenance information. The idea of a named graph is to take a set of RDF triples, and consider them as one graph, and then assign to it a URI reference. Thus, RDF can be used to describe this graph using RDF triples, which describe the creator or the retrieval data of the graph. Linked data applications can use this description to access easily a particular graph (e.g., graph for the authentication phase) and back to the original source, if required.
The named graph is useful to juries to navigate and access provenance metadata related to a certain set of triples, and get more description about them (e.g., LDspider [57] allows crawled data to be stored in an RDF store using the named graphs data model). As the SPARQL is widely used for querying RDF data, it can also be used to query named graphs.

Recently, Omitola et al. [9] allows publishers to add provenance metadata to the elements of their datasets. This is presented through the extension of the void vocabulary into *voidp* vocabulary (light weight provenance extension for the void vocabulary). This vocabulary considers different properties, such as dataset signature, signature method, certification, and authority, in order to prove the origin of a dataset and its authentication.

F. PKI Approach

Provenance metadata are not sufficient to ensure that the published data belong to the right players. PKI approach allows juries to ensure from the identity of role players participated in the forensics investigation. PKI is a combination of softwares and procedures providing a way for managing keys and certificates, and using them efficiently. A certificate is a piece of information (like a passport) that provides a recognized proof of a person (or entity) identity. A very recent work provided by Rajabi et al. [58] explains how PKI is used to achieve the trustworthiness of linked data. In this work, PKI is used for trust management over the web linked data, where datasets are exchanged in a trusted way. PKI is adapted for a new application supporting juries to verify the

 ISBN: 978-1-61208-273-8

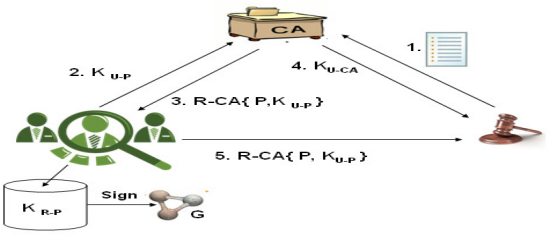

Figure 5. Application of PKI in *CoC* representation

identity of each role player who published the provided data through the investigation of their certificates. Each player in the forensics process should have his own certificate, which contains information about his identity and his digital signature.

A digital certificate alone can never be a proof of anyone's identity. A third trusted party is needed to confirm and sign the validity and authority of each player certificate. This party is then called certification authority (CA). Since a CA (e.g., VeriSign Inc.) relies on public trust, it will not put its reputation on the line by signing a certificate unless it is sure of its validity, the fact that makes them acceptable to the cyber security and cyber forensics fields. Any certificate contains pieces of information about the identity of the certificate owner (role player), such as distinguisher's name, and information about the CA (issuer of certification), such as CA's signature of that certificate, and general information about the expiration and the issue date of that certificate.

Generally, the scenario starts when the jury of the court sends a list of all players' cyber crime to the CA [15]. Each player sends a certificate request to the CA, to be signed, containing the requester's name, his PK, and his own signature. CA verifies the role player's signature with the public key in the digital signature to ensure that the private key used to generate the request matches with the public key in the certificate request. Figure 5 shows how the PKI certifications are applied in this context:

1. Juries send a list of players who are supposed to work on the current cyber crime case. Sending this list to the CA, controls the data access to only these players. This prevents the disclosure (keeps the confidentiality) of data to unauthorized people.
2. The role player generates a public-private key pair ($\{K_{U\text{-}P}, K_{R\text{-}P}\}$), where P is all information identifying the player, R is private, and U is public. The player stores the private key in a secure storage to keep its integrity and confidentiality, and then sends the public key $K_{U\text{-}P}$ to the CA.
3. The player's public key and its identifying information P are signed by the authority using its ($\{K_{R\text{-}CA}\}$) private key. The resulting data structure is back to the role player. R-CA $\{P, K_{U\text{-}P}\}$ is called the public key certificate of the role player, and the authority is called a public key certification authority (i.e., symbols outside brackets mean the signature of the data structure).
4. Juries obtain the authority's public key $\{K_{U\text{-}CA}\}$.
5. Each player creating a *CoC* must authenticate himself to juries by signing his RDF graph G using his private key R-P{G} (i.e., all triples describing a phase are assembled in one graph called G). Later, before the court session, each player sends the certification R-CA $\{P, K_{U\text{-}P}\}$ to juries accompanied with the signed graph R-P{G}.

The main idea behind this scenario is based on the PK cryptography, where senders (role players and CA) make signature using their private key, and the jury verifies these signatures using their public key.

VI. CONCLUSION AND FUTURE WORK

This paper explained how LDP can be applied to represent tangible *CoCs*. This paper provides several design options to construct the *CF-CoC* system. The best design combination is not on the scope of this paper. Along this dissertation, several contributions are provided:

1. New combination of several fields in the same framework, such as cyber forensics, semantic web, provenance vocabularies, PKI Approach, and LDP.
2. Underline that each phase in the forensics process should have its own *CoC* along any forensics model.
3. Provide a framework that leads to the creation of an assistance system for juries in a court of law.
4. Integrate provenance metadata to the victim/forensics data, in order to answer questions about the origin of information published by the role players during the forensics investigation.
5. Using the PKI approach to ensure the identities of each player participating in the forensics process.

In future work, the current framework will be extended by extra educational resources for aid purposes. These educational resources provide help to the role players and juries to respectively publish and consume the represented data.

REFERENCES

[1] T. F. Gayed, H. Lounis, and M. Bari, "Computer Forensics: Toward the Construction of Electronic Chain of Custody on the Semantic Web," SEKE 2012, pp. 406-411.

[2] T. F. Gayed, H. Lounis, and M. Bari, "Representing and (Im)proving the Chain of Custody Using the Semantic Web," The Fourth International Conference on Advanced Cognitive Technologies and Applications 2012, Nice, France, pp. 19-23.

[3] L. M. Campbell and S. MacNeill, "The semantic web, Linked and Open Data, A Briefing paper," June 2010, JISC CETIS.

[4] Provenance Requirements for the Next Version of RDF: A position paper based on the work of the W3C Provenance Incubator Group.

[5] B. Glimm, A. Hogan, M. Krötzsch, and A. Polleres, "OWL: Yet to arrive on the Web of Data?," April 2012, Lyon, France.

[6] O. Hartig and J. Zhao, "Publishing and Consuming Provenance Metadata on the Web of Linked Data," IPAW 2010, pp. 78-90.

[7] T. Berners-Lee, J. Hendler, and Ora Lassilia, "The semantic web" Scientific American, vol. 5, May 2001, pp. 34–44.

 ISBN: 978-1-61208-273-8

[8] C. Bizer, T. Heath, and T. Berners-Lee, “Linked data - the story so far,” Int. J. Semantic Web Inf. Syst., 2009, 5(3) pp. 1–22.

[9] T. Omitola, et al., “Tracing the Provenance of Linked Data using voiD,” In: The International Conference on Web Intelligence, Mining and Semantics (WIMS'11) May 2011, vol. 17.

[10] L. Tim Berners-Lee, “Design issues: Linked data,” from http://www.w3.org/DesignIssues/LinkedData.html [retrieved Feb 2013].

[11] I. Jacobs and N. Walsh, “Architecture of the world wide web,” vol. 1, 2004. http://www.w3.org/TR/webarch/. [Retrieved Feb 2013].

[12] T. Berners-Lee, R. Fielding, and L. Masinter, “RFC 2396 – Uniform Resource identifiers (URI): Generic Syntax,” http://www.isi.edu/in-notes/rfc2396. Aug 1998. [Retrieved Feb 2013].

[13] R. fielding, “Hypertext transfer protocol” – http/1.1 request for comments:http://www.w3.or/Protocols/rfc2616/rfc2616.html, 1999 [retrieved Mar 2013].

[14] Linked data: Evolving the web into a global data space, http://linkeddatabook.com/editiopns/1.0/. [Retrieved Feb 2013].

[15] R. Perlman, “An overview of PKI trust models, In IEEE network,” vol. 13, 1999, pp. 38-43.

[16] RDF Model and Syntax Specification. W3C recommendation, 22 Feb 1999, www.w3.org/TR/REC-rdf-syntax-19990222/1999. [Retrieved Jan 2013].

[17] W. Kurse and J. Heiser, “Computer Forensics: Incident Response Essentials book” Addison Wesley, 2002.

[18] Technical Working Group for Electronic Crime Scene Investigation, Electronic Crime Scene Investigation, “A Guide for first responders, United States Department of Justine,” 2001.

[19] E. Casey, “Digital Evidence and Computer Crime - Forensic Science,” Computers and the Internet, 3rd Edition. Academic Press 2011, pp. 1-807, ISBN: 978-0-12-374268-1,

[20] M. Reith, C. Carr, and G. Grunsch, “An examination of digital forensic models,” International Journal of Digital Evidence, vol.3, 2002.

[21] S.O. Ciardhuain, “An extended model of CC investigations,” International Journal of digital Evidence, vol. 3, 2004.

[22] G. Antoniou and F. V. Harmelen, “Web Ontology Language: OWL,” 2005, pp. 1-21.

[23] P. Hitzler and F. V. Harmelen, “A reasonable semantic web, Semantic Web,” vol. 1(1), 2010, pp. 39-44.

[24] B. D. Carrier, “A Hyptothesis-Based Approach to Digital Forensic Investigations,” Phd thesis, Center for Education and Research in Information Assurance and Security, Prudue University, West Lafayette, IN 47907-2086.

[25] G. Palmer, Technical Report, “A Road Map for Digital Forensic Research,” Digital Forensics Workshop (DFRWS), Utica, New York, 2001.

[26] Y. Yusoff, R. Ismail, and Z. Hassan, “Common Phases of Computer Forensics Investigation Models,” International Journal of computer science and information technology (IJCSIT), vol. 3, No 3, June 2011, pp. 17- 31.

[27] S. Vanstone, Oorschot, and A. Menezes, “Handbook of Applied Criptografy,” CRC Press, 1997.

[28] C. Brown, “Digital evidence: Collecting and Preservation”, 2006.

[29] M. Köhn, J. Eloff, and M. Olivier, "UML DFPMs," in Proceedings of the ISSA 2008 Innovative Minds Conference, Johannesburg, South Africa, July 2008.

[30] http://dublincore.org/documents/dcmi-terms/ [Retrieved Jan 2013].

[31] http://purl.org/net/opmv/ns. [Retrieved Dec 2012].

[32] M. Cohen, S. Garfinkel, and B. Schatz, “Extending the AFF to accommodate multiple data sources, arbitrary information and forensic workflow,” Digital Investigation, 2009, pp. 57-68.

[33] L. Davis, H. Shrobe, and P. Szolovits, “What is a knowledge representation?,” AI Magazine, 1993, vol. 14(1), pp. 17-3.

[34] D. Brickley and R. V. Guha. “RDF Schema,” W3C Recommendation http://www.w3.org/TR/rdf-schema/, 2004 [Retrieved Jan 2013].

[35] L. Deborah, McGuinness, and F. V. Harmelen, “OWL Web Ontology Language Overview,” W3C Recommendation. http://www.w3.org/TR/2004/REC-owl-features-20040210/, 2004 [retrieved Dec 2012].

[36] http://xmlns.com/foaf/spec/ [Retrieved Jan 2013].

[37] http://rdfs.org/sioc/spec/ [Retrieved Jan 2013].

[38] http://trac.usefulinc.com/doap [Retrieved Jan 2013].

[39] http://www.w3.org/TR/void/ [Retrieved Jan 2013].

[40] RDF/XML Specifications: W3C Recommendation 10 February 2004 http://www.w3c.org/TR/REC-rdf-syntax [Retrieved Jan 2013].

[41] Turtle – Terse RDF Triple Language: W3C Team Submission 14 January 2008, http://www.w3.org/TeamSubmission/turtle/.

[42] RDFa in XHTML: Syntax and Processing, W3C Recommendation 14 October 2008 http://www.w3.org/TR/rdfa-syntax/.

[43] N-Triples (in RDF Test cases: W3C Recommendation 10 February 2004) http://www.w3.org/TR/rdf-testcases/#ntriples.

[44] Notation3 (N3): A Readable RDF Syntax: W3C Team Submission 14 January 2008 http://www.w3.org/TeamSubmission/n3/.

[45] http://neologism.deri.ie/ [Retrieved Mar 2013].

[46] http://protege.stnaford.edu/ [Retrieved Mar 2013].

[47] http://www.topquafrant.com/products/TB_Composer.html [Retrieved Jan 2013].

[48] http://arc.semsol.org/ [Retrieved Mar 2013].

[49] http://sites.wiwiss.fu-berlin.de/sihl/bizer/d2r-server/index.html [Retrieved Feb 2013].

[50] http://d2rq.org/d2rq-language [Retrieved Feb 2013].

[51] D. Quan and D. R. Karger, “How to make a semantic web browser,” ACM, New York, USA. May 2004, pp. 255 – 265.

[52] O. Hartig and J. Zhao, “Publishing and Consuming Provenance Metadata on the Web of Linked Data,” IPAW 2010, pp. 78-90.

[53] O. Hartig, “Provenance Information in the Web of Data,” LDOW, April 20, 2009, Madrid, Spain.

[54] L. Moreau, et al. “The Open Provenance Model core specification,” (v1.1). Future Generation Comp. Syst., vol.27 (6), 2011, pp. 743-756.

[55] OPVM:http://open-biomed.sourceforge.net/opmv/opmvguide.html [retrieved Mar 2013].

[56] J. J. Carroll, C. Bizer, P. Hayes, and P. Stickler, “Named graphs, provenance and trust,” In Proceedings of the 14th international conference on WWW, NY, USA, 2005, ACM Press, pp. 613-622.

[57] R. Isele, A. Harth, J. Umbrich, and C. Bizer. “Ldspider: An open-source crawling framework for the web of linked data,” In ISWC 2010 Posters Collected Abstracts, vol. 658.

[58] E. Rajabi, M. Kahani, and M.-Angel Silicia, “Trustworthiness of Linked Data Using PKI”, April 2012, World Wide Web, Lyon, France.

[59] http://www.digitaltest/employee/Jean-Pierre [Retrieved Jan 2013].

[60] http://www.digitaltest/employee/Jean-Pierre.rdf [Retrieved Jan 2013].

[61] http://www.digitaltest/employees/Jean-Pierre.html [Retrieved Jan 2013].

[62] http://www.digitaltest.ca/employees [Retrieved Jan 2013].

[63] http://cyberforensics-coc.com/ [Retrieved Jan 2013].

[64] http://cyberforensicscoc.com/vocab/authentication# [Retrieved Jan 2013].

[65] G. C. Kessler, “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, Sep. 2010.

 ISBN: 978-1-61208-273-8