Digital Investigation 12 (2015) S102–S110

Contents lists available at ScienceDirect

Digital Investigation

journal homepage: www.elsevier.com/locate/diin

Leveraging CybOX™ to standardize representation and exchange of digital forensic information

Eoghan Casey*, Greg Back, Sean Barnum

The MITRE Corporation, 7525 Colshire Drive, McLean, VA 22102-7539, USA

Keywords:
Digital forensics
Standard representation
Digital forensic ontology
Digital forensic XML
CybOX
DFXML
DFAX

A B S T R A C T

With the growing number of digital forensic tools and the increasing use of digital forensics in various contexts, including incident response and cyber threat intelligence, there is a pressing need for a widely accepted standard for representing and exchanging digital forensic information. Such a standard representation can support correlation between different data sources, enabling more effective and efficient querying and analysis of digital evidence. This work summarizes the strengths and weaknesses of existing schemas, and proposes the open-source CybOX schema as a foundation for storing and sharing digital forensic information. The suitability of CybOX for representing objects and relationships that are common in forensic investigations is demonstrated with examples involving digital evidence. The capability to represent provenance by leveraging CybOX is also demonstrated, including specifics of the tool used to process digital evidence and the resulting output. An example is provided of an ongoing project that uses CybOX to record the state of a system before and after an event in order to capture cause and effect information that can be useful for digital forensics. An additional open-source schema and associated ontology called Digital Forensic Analysis eXpression (DFAX) is proposed that provides a layer of domain specific information overlaid on CybOX. DFAX extends the capability of CybOX to represent more abstract forensic-relevant actions, including actions performed by subjects and by forensic examiners, which can be useful for sharing knowledge and supporting more advanced forensic analysis. DFAX can be used in combination with other existing schemas for representing identity information (CIQ), and location information (KML). This work also introduces and leverages initial steps of a Unified Cyber Ontology (UCO) effort to abstract and express concepts/constructs that are common across the cyber domain.

© 2015 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Introduction

In the modern age, any type of investigation can have a digital dimension, ranging from computers as a source of information in homicides and terrorist attacks, to computers as instrumentalities of fraud and cyber-attacks. As a result, digital forensics supports decision makers in various domains, including law enforcement, incident response, malware analysis, cyber threat intelligence, and situational awareness. To combat crime effectively in the modern age, digital forensic information needs to be represented and shared in a form that is usable in any of these contexts.

When investigating a single incident, being able to combine the results from multiple tools that are used to extract information from the digital evidence supports forensic reconstruction, including timeline creation and link analysis. In addition, being able to automated the comparison of similar results from multiple tools facilitates dual-tool verification. When crime spans borders, sharing of information between investigative agencies in multiple jurisdictions

* Corresponding author. Reception Desk ext. 3-6004.
E-mail address: ecasey@mitre.org (E. Casey).

http://dx.doi.org/10.1016/j.diin.2015.01.014
1742-2876/© 2015 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

is crucial for a successful resolution. A fundamental requirement in digital forensics is to maintain information about evidence provenance as it is exchanged and processed, to help establish authenticity and trustworthiness.

Furthermore, without a standardized approach to representing and sharing digital forensic information, investigators in different jurisdictions may never know that they are investigating crimes committed by the same criminal. A similar challenge was recognized in traditional investigations of violent crime, and led to the development of the U.S. Federal Bureau of Investigation's Violent Criminal Apprehension Program (ViCAP) and Royal Canadian Mounted Police's Violent Crime Linkage System (ViCLAS). These programs collect distinctive details about unsolved violent crimes in disparate regions, and correlate this information to find links between related crimes.

Current efforts to manage and exchange digital forensic information are typically ad hoc, inconsistent, and limited in sophistication and expressivity. Combining results from different tools into a consistent format can be a laborious process that can result in errors or omissions. For example, using Excel to import and format data from various sources can result in items such as date-time stamps being altered, entries not being imported, and other problems that negatively impact forensic analysis.

Where standardized representations of digital forensic information are used, they are typically focused on an individual portion of the overall digital forensic process (Flaglien et al., 2011). Such focused efforts have benefits, supporting in-depth exploration of specialized domains such as file systems, but do not support broader representation and analysis. In addition, existing formalized representations of digital forensic information do not integrate well with each other, or lack coherent flexibility and semantic structure. Existing information sharing activities are often human-to-human exchanges of unstructured or semi-structured descriptions of digital forensic artifacts and analysis, and often require conversion from proprietary formats. For instance, individual forensic examiners document their findings on personal blogs, and share parsers in proprietary formats such as EnCase's EnScript.

To address these issues, this work aims to formalize and extend the management and direct machine-to-machine exchange of progressively more expressive sets of digital forensic information using fully-structured data. Specifically, this paper describes a community-driven solution to address this problem, which leverages the Cyber Observable eXpression (CybOX) language (http://cybox.mitre.org). CybOX is an open-source, community-driven effort to develop a standardized representation of digital observables led by the U.S. Department of Homeland Security (DHS) office of Cybersecurity and Communications. CybOX is designed to represent digital actions and objects along with their context, which can be leveraged in a wide variety of use cases, including incident response, intrusion detection, and digital forensics. Development of CybOX has occurred under the coordination of the DHS-funded and MITRE operated Systems Engineering and Development Institute (SEDI), a Federally Funded Research and Development Center (FFRDC). Thus, MITRE manages the CybOX website, supports community engagement, and oversees its discussion lists to enable open and public collaboration around CybOX with all stakeholders.

This paper proposes a new standard for representing and exchanging digital forensic information called Digital Forensic Analysis eXpression (DFAX) that leverages CybOX for representing the purely technical information. DFAX incorporates its own structures to represent the more procedural aspects of the digital forensic domain, including those for chain-of-custody, case management, and forensic processing. A related effort has already been accomplished in the development of the Structured Threat Information eXpression (STIX) language to represent cyber threat information (Barnum, 2012). STIX makes use of CybOX to represent technical cyber threat details, e.g., malicious IPs, domains, and file hashes, and adds other constructs to represent domain-specific information such as campaigns and threat actors.

The capture of general criminal justice related information has been considered in other efforts such as the National Information Exchange Model (www.niem.gov), EVIDENCE Project (2013), and FIDEX (NFSTC, 2010). However, there is a need in this space to accommodate more than just the criminal justice application, and as such, DFAX proposes general elements to cover all use cases.

The ontological view of DFAX is depicted in Fig. 1, showing where CybOX fits. As is clearly shown, at a high-level DFAX covers information about various roles involved in digital forensics, various actions these roles take, evidence records resulting from forensic actions, and domain specific concepts such as authorizations as well as various abstractions to lend context to roles and actions. Actions in particular play a significant role in DFAX. A `Forensic Action` is defined as any action performed on or resulting in an `Evidence Record`. DFAX also defines `Subject Action` and `Victim Action` that can describe associated digital traces.

This work also introduces and leverages initial steps of a Unified Cyber Ontology (UCO) effort to abstract and express constructs that are common across the cyber domain, and that can be leveraged for consistency and broad-scope interoperability by various domain specific languages, including DFAX and STIX. Two examples of these abstractions leveraged in DFAX are `Action Pattern` and `Action Lifecycle`. The `Action Pattern` construct enables the contextualization of a given `Action` instance as to what sort of behavior it may represent. The `Action Lifecycle` construct can be adapted to define phases of a forensic investigation (e.g., documentation, preservation, examination, analysis, presentation) and criminal activities such as a sexual predator's grooming of victims or a network intruder's method of operation (e.g., kill chain phases). This generalized approach can be used to classify each action in a case, which provides context to support further analysis.

This paper starts with an overview of existing work related to representing digital forensic information, and then describes how CybOX can be leveraged and extended to represent digital evidence, relationships between objects, and actions associated with digital forensic information. Use cases for structured digital forensic information are discussed, and examples are presented to demonstrate how DFAX provides a layer of domain specific information overlaid on CybOX. This paper includes links for community

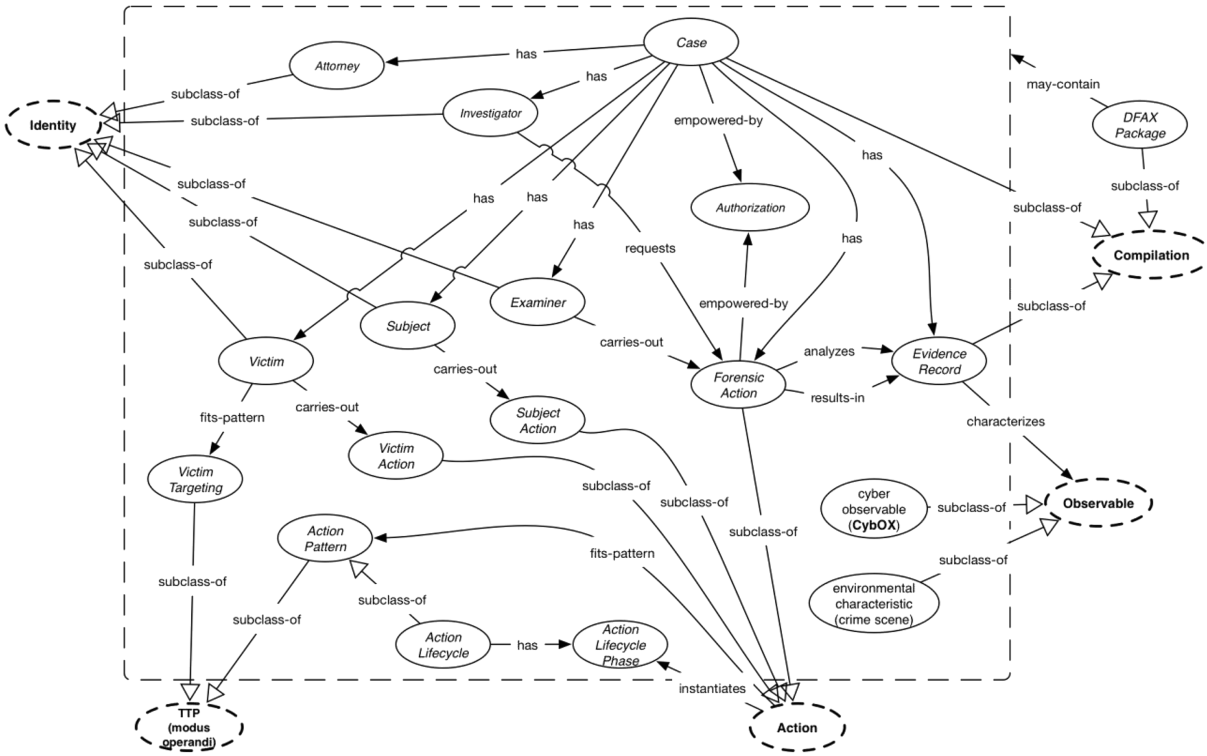

Fig. 1. DFAX high-level ontological view, leveraging CybOX. Items in bold, dotted ovals exist outside of DFAX, as concepts or schemas.

involvement to enable further development of this standardized representation.

Related work

There have been several schemas proposed in the past for representing digital forensic information, but these have not been widely adopted (Turner, 2005a, 2006; Eaglin and Craiger, 2005; Lee et al., 2008; Levine and Liberatore, 2009). One schema that is in use is Digital Forensics XML or DFXML (Garfinkel, 2009, 2012a). This schema was primarily developed to represent the output from tools used to analyze storage media, including file system parsers, file carvers, and hash set generators. DFXML is implemented in several digital forensic tools including Fiwalk (based on The SleuthKit), and as a Python library with bundled programs that read and write DFXML documents (Garfinkel, 2012a).

The XIRAF system was created by the Netherlands Forensic Institute (NFI) to support digital forensic analysis, and also uses an XML-based implementation to store its data (Alink et al., 2006; Bhoedjang et al., 2012). This schema was developed to represent the results of forensic examination, including information extracted from media and communications. The XIRAF schema adopts a parent–child approach to representing relationships between items, where each child item is extracted from its parent (NFI, 2013). For example, a computer is the parent of a hard drive, which, in turn, is the parent of a forensic duplicate; this duplicate can have a variety of associated children, including files, Windows Registry entries, Web browser history records, and e-mail. XIRAF treats both the forensic duplicate and its source (e.g., room, computer, evidence tag) as uniquely labeled "resources," and uses the term "item" to describe data that are extracted from a resource. Each item is uniquely identified and can be assigned a type (e.g., system, image, volume, file system, folder, allocated file, deleted file, unallocated content, account, email, attachment, phoneCall, textMessage) with associated properties such as metadata. However, the presumption that all digital forensic information must be organized in a hierarchical structure means that XIRAF lacks flexibility to represent other, non-hierarchical, relationships. The XIRAF schema is not widely used outside of NFI.

The Advanced Forensic Format (AFF4) implements another approach to representing digital forensic information (Schatz, 2007; Cohen et al., 2009) using the Resource Description Framework (RDF), a general purpose representational formalism from the field of knowledge representation. Although the majority of digital forensic tools do not support AAF4, Google Rapid Response (GRR) uses the AFF4 data model to store information in a MongoDB database (Cohen, 2013). The AFF4 data model is extremely flexible, but its use of RDF requires a supporting ontology to be agreed upon and clearly defined. Currently, there is no community consensus for such ontology to support consistent representation and exchange of digital forensic information. DFAX addresses this gap with an ontology that could be used as a basis for community consensus.

Table 1 summarizes the features of existing models to better understand the requirements for representing digital forensic information.

The current CybOX reference implementation is in XML and thus uses this format to represent objects and events related to digital forensics. The CybOX schemas are in active use and development, and are being incorporated into some information security and digital forensic tools. For instance, the digital forensic platform Autopsy, and the Volatility memory forensics framework can parse indicators in CybOX format using a publically available CybOX Python library (Levy, 2013). A similar library is being developed to support parsing and generating DFAX content in Python.

CybOX is also being used to support malware characterization (http://maec.mitre.org); incident response, threat indicator, and broad threat intelligence specification (http://stix.mitre.org); and attack pattern adornment (http://capec.mitre.org). CybOX was developed with extensibility in mind in order to represent a variety of digital objects, the relationships between them, and also the events associated with them. This schema currently covers common digital objects and associated characteristics, and new object types can be added to CybOX without altering the core schema. Objects that are currently represented within CybOX include those pertaining to devices, disk partitions and volumes, files and Windows Registry entries, system logs and network traffic. Proposed changes and updates to CybOX are handled through community collaboration and consensus both online and via in-person meetings.

XML advantages and limitations

The Extensible Markup Language (XML) is a common markup language for representing information in a structured manner. XML schemas are used to define a set of rules for encoding documents in a consistent manner that is both machine processable and human readable. An XML document stores information within units called *elements*. Each element is designated by a tag that defines the type of information it contains, and can have attributes that store related details.

A major advantage of XML is that it provides a mechanism to explicitly define and enforce the language structure for interoperability. The data stored in an XML document can be exchanged across different system platforms, software applications and programming languages. The forensic community can author the XML vocabulary needed to exchange data between tools. XML is flexible and the tag set can be extended to handle additional information relevant to the domain. This is an improvement over customized, vendor-specific formats where data may be lost in translation. Multiple views of the same content can be easily rendered using style sheets, and instance data can be validated against a defined specification. In addition, XML provides standard APIs to facilitate querying and manipulation of data.

One of the main limitations of XML is that it can be verbose, but this can be mitigated by other implementations such as EXI, which is compressed. In certain situations, it may be more effective to structure data using other formats. Choosing XML to define DFAX does not preclude the use of other implementations such as JSON, ASN.1, or Protocol Buffers. A limited XML vocabulary and schema can be translated into these or other serialization formats as needed.

Representing digital evidence using CybOX

When processing digital evidence, it is necessary to capture details about specific objects and their contextual properties such as manufacturers and serial numbers of storage media or mobile devices, and names of files stored on a device with their associated date-time stamps. In addition, for integrity and comparison purposes, it is necessary to record cryptographic hash values of digital objects.

The CybOX schema can represent many types of digital objects and their associated characteristics, including disks, devices, file systems, Windows Registry entries, memory, and network traffic, providing a solid foundation for representing digital evidence. Hash values, including MD5 and SHA256, can be captured for objects that contain data, such as files and memory contents. For example, Fig. 2 shows a CybOX representation of traces associated with the SDelete program, including the Registry entry corresponding to acceptance of the end-user-license agreement (EULA), and the repeated “Z” pattern used to

Table 1
Features of prior proposed schemas for digital forensic information.

	Digital evidence bags (DEB)	XIRAF	RDF	Digital evidence exchange (DEX)	Digital forensic XML (DFXML)	AFF4	DFAX/CybOX
	Turner, 2005a,b	Alink et al., 2006	Schatz, 2007.	Levine and Liberatore, 2009	Garfinkel, 2009.	Cohen et al., 2009	Open source (DHS/MITRE)
Open source	Y	N	N/A	Y	Y	Y	Y
Case information	Y	Y	N	N	N	Y	Y
Integrity assurance	Y	Y	Y	Y	Y	Y	Y
Chain of custody	Y	N	P	N	N	Y	Y
Evidence details	Y	Y	Y	N	N	Y	Y
Tool details	Y	Y	Y	Y	Y	Y	Y
Storage media contents	Y	Y	Y	N	Y	Y	Y
Mobile device contents	P	Y	N	N	N	N	Y
Assign object multiple types	Y	Y	Y	N	N	Y	Y
Parent–child relationships	Y	Y	Y	Y	Y	N	Y
Non-hierarchical relationships	N	N	Y	N	N	Y	Y
Actions	N	N	Y	N	N	N	Y
Action lifecycles	N	N	N	N	N	N	Y
Action patterns	N	N	N	N	N	N	Y

overwrite filenames. A globally unique identifier (GUID) for each observable and object is stored as an attribute of the Observable and Object elements.

Codifying and sharing information in a standardized manner enables digital investigators to search for similar patterns in their cases. Even when wiping tools other than SDelete are involved, the CybOX observable instance representation could be abstracted into a CybOX observable pattern to search for any executable that overwrites filenames with repeated letter. For instance, when PGP is used to wipe a file, it overwrites the filename with repeated "a" pattern. Finding similar patterns between cases can support reuse of previously effective solutions, such as forensic analysis methods for proving that wiping occurred and possibly recovering remnants of overwritten files, thus reducing duplication of effort and increasing consistency of forensic analysis (Casey, 2013). Furthermore, searching for specific patterns across cases can potentially reveal links between related crimes (Garfinkel, 2012b).

Although the CybOX schemas contain a large number of elements and attributes, nearly all of these elements are optional, so digital forensic applications can make use of what is needed and leave the remainder. In addition, CybOX is designed to be extensible so that schemas needed for digital forensics can be added, such as for capturing information associated with mobile devices. Furthermore, the development of CybOX is being actively driven and adopted by the broader cyber security community. Therefore, using CybOX as a foundation for DFAX will ensure alignment with other cyber security domains.

Provenance

There is general consensus in the forensic science community that it is important to establish the provenance of evidence, but few agree on what exactly provenance means. Turner discusses provenance in terms of maintaining chain of custody documentation, as well as documenting the context of data found on storage media, emphasizing the importance of temporal information such as timestamps (Turner, 2005a,b). Levine and Liberatore describe provenance in terms of "the set of tools and transformations that led from acquired raw data to the resulting product" (Levine and Liberatore, 2009).

For forensic purposes, to help establish the authenticity and reliability of evidence, it is important to capture where an item originated or was found (sometimes referred to as provenience in archeology), as well as how an item was handled after it was found.

In a legal context, the evidence authentication process uses information such as collection documentation, continuity of possession forms (chain of custody), audit logs from forensic acquisition tools, and integrity records, to help establish the trustworthiness of digital evidence.

In the context of forensic examination, provenance refers to the source and extraction method of specific items, such as the extraction of e-mail messages, attachments, and their associated metadata from a Microsoft Outlook PST file using a specific software application (Turner, 2005b).

Many aspects of provenance can be captured using CybOX, including the source (both human and electronic/tool) and timing of the originating evidential item and the processing of data using forensic tools. CybOX has the `cyboxCommon:Tool` element to record details about tools used to process digital evidence. However, CybOX does not support the concepts of evidence handling which DFAX provides. As depicted in Fig. 1, DFAX captures observables within `Evidence Records` which could also include environmental characteristics such as the details of a crime scene or where the evidence was physically located, and captures information about any `Forensic Action` associated with each `Evidence Record`, as well as tracking who performed each `Forensic Action` and when it was performed.

Complete technical representation of the physical location where evidence was obtained and the people associated with the evidence can be covered by existing schemas such as the Oasis Customer Information Quality Specification (www.oasis-open.org/committees/ciq/). Therefore, rather than recreating a new representation of such information, it may be more effective to leverage an existing schema for such data. DFAX and CybOX have been designed to accommodate such re-use – rather than include its own geolocation schema, it defines an extension point where an existing schema, such as KML, can be used.

Provenance also has relevance in forensic analysis, describing the evaluation of source such as determining whether a photo was taken using a given digital camera based on Photo Response Non Uniformity (PRNU) or First Step Total Variation (FSTV) comparison, or whether a video was recorded in a particular location based on Electrical Network Frequency (ENF) data. Analyzing the provenance of an item can also be used to ascertain whether it is forged or the genuine object. CybOX does not currently support these features but could be extended to capture them in a standardized manner.

Efficiency and cost of forensic processes

Recording the time taken to complete a `Forensic Action` in a standardized manner serves several purposes. In addition to being central to tracking provenance, this date-time information can be used to calculate which processes consume the most resources. Identifying bottlenecks in the overall forensic process creates opportunities to improve efficiency (Casey et al., 2013). In addition, this temporal information is useful for calculating damages associated with a crime.

Fully-structured data in DFAX

Capturing the relationships between items is important in digital forensics for provenance and investigative purposes. For example, a single smartphone can result in various such relationships, including:

- Android device contains SDCard
- SDCard contains incriminating photograph
- Photograph contains geolocation information

```
<?xml version="1.0" encoding="UTF-8"?>
<cybox:Observables [edited for length] cybox_major_version="2" cybox_minor_version="1">
  <cybox:Observable id="example:Observable-1">
    <cybox:Description>SecureDelete program from Microsoft's SysInternals suite.</cybox:Description>
    <cybox:Object id="example:Product-1">
      <cybox:Properties xsi:type="ProductObj:ProductObjectType">
        <ProductObj:Product>Secure Delete</ProductObj:Product>
        <ProductObj:Version>1.61</ProductObj:Version>      </cybox:Properties>
      <cybox:Related_Objects>
        <cybox:Related_Object idref="example:File-2">
          <cybox:Relationship>Characterizes</cybox:Relationship>
        </cybox:Related_Object>
        <cybox:Related_Object idref="example:RegistryKey-3">
          <cybox:Relationship>Created</cybox:Relationship>
        </cybox:Related_Object>
      </cybox:Related_Objects>
    </cybox:Object>
  </cybox:Observable>
  <cybox:Observable id="example:Observable-2">
    <cybox:Description>The actual executable file corresponding to the SDelete tool.</cybox:Description>
    <cybox:Object id="example:File-2">
      <cybox:Properties xsi:type="FileObj:FileObjectType">
        <FileObj:File_Name>sdelete.exe</FileObj:File_Name>
        <FileObj:File_Format>PE Windows Executable</FileObj:File_Name>
        <FileObj:Hashes>
          <cyboxCommon:Hash>
            <cyboxCommon:Type xsi:type="cyboxVocabs:HashNameVocab-1.0">SHA256</cyboxCommon:Type>
            <cyboxCommon:Simple_Hash_Value>97D27E1225B472A63C88AC9CFB813019B72
            598B9DD2D70FE93F324F7D034FB95</cyboxCommon:Simple_Hash_Value>
          </cyboxCommon:Hash>
        </FileObj:Hashes>
        <FileObj:Created_Time>2013-01-09T14:25:04-0800</FileObj:Created_Time>
      </cybox:Properties>
    </cybox:Object>
  </cybox:Observable>
  <cybox:Observable id="example:Observable-3">
    <cybox:Description>SDelete tool has been launched, and had it's EULA accepted.</cybox:Description>
    <cybox:Object id="example:RegistryKey-3">
      <cybox:Properties xsi:type="WinRegistryKeyObj:WindowsRegistryKeyObjectType">
        <WinRegistryKeyObj:Key>Software\Sysinternal\SDelete</WinRegistryKeyObj:Key>
        <WinRegistryKeyObj:Hive>HKCU</WinRegistryKeyObj:Hive>
        <WinRegistryKeyObj:Values>
          <WinRegistryKeyObj:Value>
            <WinRegistryKeyObj:Name>EulaAccepted</WinRegistryKeyObj:Name>
            <WinRegistryKeyObj:Data>1</WinRegistryKeyObj:Data>
          </WinRegistryKeyObj:Value>
        </WinRegistryKeyObj:Values>
      </cybox:Properties>
    </cybox:Object>
  </cybox:Observable>
  <cybox:Observable id="example:Observable-4">
    <cybox:Description>A file that has been overwritten using the SDelete tool.</cybox:Description>
    <cybox:Object id="example:File-4">
      <cybox:Properties xsi:type="FileObj:FileObjectType">
        <FileObj:File_Name condition="FitsPattern">Z+.Z+</FileObj:File_Name>
        <FileObj:Modified_Time condition="GreaterThanOrEqual">2015-01-03T11:49:44-
        0800</FileObj:Modified_Time>
      </cybox:Properties>
      <cybox:Related_Objects>
        <cybox:Related_Object idref="example:Product-1">
          <cybox:Relationship">Renamed_By</cybox:Relationship>
        </cybox:Related_Object>
      </cybox:Related_Objects>
    </cybox:Object>
  </cybox:Observable>
</cybox:Observables>
```

Fig. 2. CybOX representation of SDelete.

- Geolocation information is at suspect's home
- Photograph attached to e-mail
- Email sent from subject to victim
- SMS sent from subject to victim

A sample DFAX representation of this scenario is available at the community site (https://github.com/dfax/dfax).

DFAX leverages CybOX to represent associations between items, including parent–child relationships using the `Child_Of` and `Parent_Of` values in the ObjectRelationshipVocab-1.0 vocabulary, and through a wide range of more specific types of relationships, including `Installed`, `Created`, `Contains`, `Related_To`, and `Deleted`.

Being able to represent structure by defining relationships within the data enables search and analysis methods at a higher level of abstraction, including graph query and pattern matching. For instance, defined relationships between items as shown in Fig. 3 could be utilized to perform a graph search for all e-mail messages with a picture attachment from the subject to the victim.

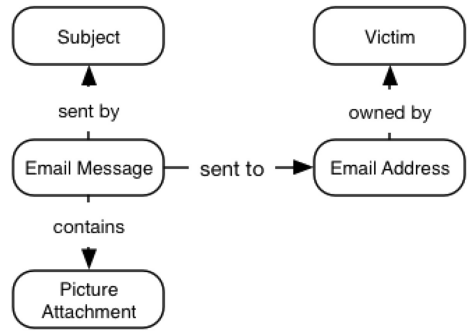

Fig. 3. Depiction of relationships (graph).

Representing actions using DFAX

DFAX extends the capability of CybOX to represent an action or multiple related actions, which can be useful for sharing knowledge and supporting more advanced forensic analysis. In addition to supporting provenance, `Forensic Actions` can give insight into which tools and methods are effective in particular circumstances. Of greater significance is the ability in DFAX to define actions associated with digital traces. This type of abstraction can provide higher-level, human understandable portrayals of activities for more efficient forensic analysis (Hargreaves and Patterson, 2012). For example, a high level action could be labeled `USB Device Connected`, which comprises multiple low level digital traces, the occurrence of which can be important in some digital investigations. A sample DFAX representation of traces associated with the action of a particular USB device being connected to a Windows system is available at the community site.

Rather than simply searching available data sources for the serial number of a specific USB device, by representing the information in a structured manner, DFAX enables more advanced forensic analysis. For example, the structured DFAX representation of `USB Device Connected` could be used to search multiple Windows systems for the combination of artifacts, including entries in the Registry and in the "setupapi" log file, providing enough information to determine that a particular USB device was plugged into other systems during a particular time frame. A search for a specific USB device being connected could be represented in simple pseudocode form as follows:

```
   FileObj:File_Path =
C:\Windows\inf\setupapi.dev.log
                    OR
 FileObj:File_Path = C:\Windows\setupapi.log
                  CONTAINS
cyboxCommon:String_Value =
"\DISK&VEN_KINGSTON&PROD_DATATRAVELER_3.0&RE
V_PMAP\8606E6D418ABE6077179FAE&0]
                    OR
     WinRegistryKeyObj:Name = HardwareID
                    IN
WinRegistryKeyObj:Hive = HKEY_LOCAL_MACHINE
WinRegistryKeyObj:Key =
"\SYSTEM\CurrentControlSet\Enum\USBSTOR\
Disk&Ven_Kingston&Prod_DataTraveler_3.0&Rev_
PMAP\08606E6D418ABE6077179FAE&0"
                  CONTAINS
WinRegistryKeyObj:Data =
"USBSTOR\DiskKingstonDataTraveler_3.0PMAP"
```

In addition to capturing details associated with such a `USB Device Connected` event in a specific case, the DFAX representation of this information shows digital investigators what kinds of artifacts to look for in other cases involving similar actions.

Beyond searching for a specific USB device being connected, a DFAX representation can be used to search for any `USB Device Connected` event regardless of the specific serial number or model of USB device. Some forensic tools are adding features to support such searches for generalized activities of interest that comprise various low-level artifacts. For instance, the tagging feature in Plaso (https://code.google.com/p/plaso/) can group certain combinations of digital artifacts into event categories such as `Application Execution`, `Document Opened`, and `File Downloaded` that can be queried to return the underlying low-level digital artifacts associated with these events. DFAX provides a standardized way to represent these kinds of actions. Furthermore, beyond simply categorizing low-level artifacts, DFAX can be used to define relationships between actions, thus enabling more structured searches and refined analysis.

Representing action patterns using DFAX

Some actions can be described using higher-level terms, potentially reflecting associated behaviors and objectives, which can be an important part of forensic analysis. For instance, a specific set of digital artifacts could be expressed as belonging to `File Wiping` or `Disk Cleaning`, and both of these actions could in turn be categorized as `Data Destruction`.

Similarly, a specific set of digital artifacts could be expressed as belonging to `Encryption` or `Steganography`, and both could be categorized at a higher level as `Data Hiding`. Another set of digital artifacts could be expressed as `Clock Changing`, and be categorized at a higher level as `Data Staging`. An overarching category that encompasses all of these actions can be termed `Concealment`.

Using an approach similar to STIX, as depicted in Fig. 4, DFAX supports a set of vocabularies to tie low-level digital objects such as those represented using CybOX, to relevant actions and then contextualize those actions into higher level action patterns that can assist in establishing modus operandi, which can be significant to digital investigators. STIX refers to these behaviors as tactics, techniques, and procedures (TTP). In addition to concealment activities, this schema can be used to categorize actions associated with planning (e.g., Web searches on how to poison a person), presurveillance of a victim or crime scene, grooming (e.g., sending a victim pornography to reduce resistance to sexual activities), and various other kinds of behavior encountered in digital investigations. Such categorization allows for querying data on the basis of high-level behaviors, which can be more powerful than just searching for low-level digital artifacts.

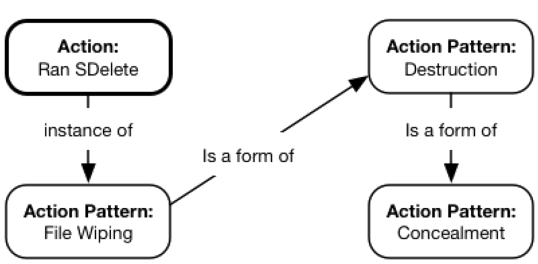

Fig. 4. Depiction of action patterns (graph).

Representing changes using CybOX

To represent the digital traces associated with certain actions such as a specific application being installed on a system, it is often necessary to compare the system state before and after the application is installed. In a digital investigation, discerning the specific alterations between multiple versions of a file can be important. In a digital context, differences can be described specifically (e.g., new file created, updated date-time stamp) or statistically (e.g., similarity digests). Such information can be represented as discrete Actions in CybOX.

For example, the NIST Diskprint project is expanding the NSRL metadata reference set by recording changes made to a system by an application over its lifecycle. As a way of communicating these changes, NIST outputs the file metadata as CybOX (http://www.nsrl.nist.gov/dskprt/DPexample.html). Specific digital traces generated when a particular software application or piece of malware is installed can be represented as Actions in CybOX, comparing the before and after states of the system, to support digital forensic investigations. This type of Diskprint information can be useful for automatically determining whether certain applications were installed and used on a system, even after uninstallation, which can be pertinent in some digital investigations.

Representing an absence of digital evidence

As offenders become more aware of digital forensic capabilities, they take precautions to cover their digital traces using a variety of concealment methods such as file wiping and encryption. Therefore, it is also necessary to represent explicitly the absence of information on a computer in order to support further analysis of potential evidence destruction or concealment such as cleared Web browser history or deleted system/security event logs. This can be accomplished in CybOX using the value `Does Not Exist` for the `State` element within any object.

Conclusions and future work

To be effective, digital forensic information needs to be represented and shared in a form that is consistent across all applicable contexts and tools. In order to be adopted, it is necessary for a language to cover the information that needs to be represented.

The focus of this work is on leveraging CybOX to support a new Digital Forensic Analysis exchange eXpression (DFAX) schema that provides a standard approach for representing digital facts, their relationships, associated provenance details, and higher level behaviors. The

expanding use of CybOX in digital forensics and related domains makes it a strong candidate for standardized representation of digital forensic information.

There are ongoing efforts to add more types of digital objects and associated attributes into CybOX, and to refine how file systems, smartphones, and network connections are represented within CybOX. As a community driven effort, such changes are formally proposed and reviewed through community meetings and online forums, ensuring broad acceptance and adoption.

The new DFAX schema under community development is required to ensure that information specific to the digital forensic domain can be represented. Community consensus is being sought in various forums, including a mailing list and online repository (https://github.com/dfax/dfax), to define the DFAX ontology and elements outlined in Fig. 1. After DFAX matures and an explicit ontology has been agreed upon within the digital forensic community, other more flexible ways to represent the data such RDF XML/OWL can be explored.

Various organizations are currently sharing cyber threat and malware related information in CybOX format using Trusted Automated eXchange of Indicator Information (TAXII), which includes federated sharing models. Similar exchange mechanisms could be used to exchange DFAX.

Acknowledgments

This work has been encouraged and supported by William Eber, Barbara Guttman, Mary Laamanen, Alex Nelson, Penny Chase, IvanKirillov, Dave Baker, Jon Baker, and Charles Schmidt.

References

Alink W, Bhoedjang R, Boncz P, de Vries A. XIRAF—XML-based indexing and querying for digital forensicsProceedings of DFRWS2006. Digit Investig 2006;3(Suppl.):S50–8. Elsevier.

Barnum S. Structured threat information eXpression. The MITRE Corporation; 2012. stix.mitre.org/about/documents/STIX_Whitepaper_v1.0.pdf.

Bhoedjang RAF, van Ballegooijb AR, van Beeka HMA, van Schiea JC, Dillemab FW, van Baara RB, et al. Engineering an online computer forensic service. Digital Investigation 2012;9(2).

Casey E. Reinforcing the scientific method in digital investigations using a case-based reasoning (CBR) system [Ph.D. dissertation]. Ireland: University College Dublin; 2013.

Casey E, Katz G, Lewthwaite J. Honing digital forensic processes. Digit Investig 2013;10(2).

Cohen M, Schatz B, Garfinkel S. Extending the advanced forensic format to accommodate multiple data sources, logical evidence, arbitrary information and forensic workflowProceedings of DFRWS2009. Digit Investig 2009;6(Suppl.):S57–68. Elsevier.

Cohen M. Hunting in the enterprise: forensic triage and incident response. Digit Investig 2013;10(2).

Eaglin R, Craiger JP. Data sharing and the digital evidence markup language. 2005. Presented at 1st annual GJXDM users conference, Atlanta.

EVIDENCE Project. EVIDENCE semantic structure. European Informatics Data Exchange Framework for Courts and Evidence; 2013. http://s.evidenceproject.eu/p/e/v/evidence-ga-608185-d02-1-final-31072014-136.pdf.

Flaglien AO, Mallasvik A, Mustorp M, Arnes A. Storage and exchange formats for digital evidence. Digit Investig 2011;8(2):122–8.

Garfinkel SL. Automating disk forensic processing with SleuthKit. In: XML and Python, systematic approaches to digital forensics engineering (IEEE/SADFE 2009), Oakland, California; 2009.

Garfinkel SL. Digital forensics XML and the DFXML toolset. Digit Investig 2012a;8(3–4):161–74. Elsevier.

Garfinkel SL. Cross-drive analysisProceedings of DFRWS2006. Digit Investig 2012b;3(Suppl.):S71–81. Elsevier.

Hargreaves C, Patterson J. An automated timeline reconstruction approach for digital forensic investigationsProceedings of DFRWS2012. Digit Investig 2012;9(Suppl.):S69–79. Elsevier.

Lee S, Park T, Shin S, Un S, Hong D. A new forensic image format for high capacity disk storage. In: Information security and assurance, 2008. ISA 2008. International conference on information security and assurance, 24–26 April. IEEE Computer Society; 2008.

Levine BN, Liberatore M. DEX: digital evidence provenance supporting reproducibility and comparison. Digit Investig 2009;6(Suppl.):S48–56. Elsevier, https://github.com/umass-forensics/DEX-forensics.

Levy J. Leveraging CybOX with volatility. Volatility Labs Blog; 2013. http://volatility-labs.blogspot.com/2013/09/leveraging-cybox-with-volatility.html [viewed 16.02.14].

National Forensic Science Technology Center. Forensic information data exchange (FIDEX) final project report. 2010.

Schatz B. Digital evidence: representation and assurance [PhD dissertation]. Queensland University of Technology; 2007., http://eprints.qut.edu.au/16507/1/Bradley_Schatz_Thesis.pdf.

Turner P. Unification of digital evidence from disparate sources (digital evidence bags)Proceedings of DFRWS2005. Digit Investig 2005a;2(3):223–8. Elsevier.

Turner P. Digital provenance – interpretation, verification and corroboration. Digit Investig 2005b;2(1):45–9.

Turner P. Selective and intelligent imaging using digital evidence bagsProceedings of DFRWS2006. Digit Investig 2006;3(Suppl.):59–64.