Chapter for EU EVIDENCE Project publication

Title: **The Evolution of Expressing and Exchanging Cyber-investigation Information in a Standardized Form**

Eoghan Casey (University of Lausanne)[1]*, Sean Barnum(MITRE), Ryan Griffith (DC3), Jonathan Snyder (DC3), Harm van Beek (Netherlands Forensic Institute), Erwin van Eijk (Netherlands Forensic Institute), Ruud van Baar (Netherlands Forensic Institute), Alex Nelson (National Intitute of Standards and Technology)[2]

Motivation

This paper describes the evolution of a community-developed, standardized specification language for representing and exchanging information in the broadest possible range of cyber-investigation domains, including digital forensic science, incident response, and counter terrorism. This initiative was originally called the Digital Forensic Analysis eXpression (DFAX), which has evolved into the Cyber-investigation Analysis Standard Expression (CASE). These standardization efforts include development of the Unified Cyber Ontology (UCO) to provide the scaffolding necessary to satisfy the requirements of a wide range of use cases in multiple specializations.

A primary motivation for this community driven initiative is interoperability - to enable the exchange of cyber-investigation information between tools, organizations, and countries. The CASE specification language and UCO ontology are a rational progression from the foundational work on Digital Forensic Analysis eXpression (DFAX), which focused on digital forensic information and provenance context (Casey, Back, Barnum, 2015).

"*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 automate the comparison of similar results from multiple tools facilitates dual-tool verification. When crime spans borders, sharing of information between investigative agencies 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.*" (Casey, Back, Barnum, 2015)

The power of such a standard is that it supports automated normalization, combination and correlation of information, which means less time extracting and combining data, and more time analyzing information. In addition, the exchange of information in a standardized format helps breakdown data silos, increasing visibility across disparate data sources as depicted in Figure 1.

* Corresponding author. Tel.: +41-21-692-46-12; fax: +41-21-692-46-05.
E-mail address: eoghan.casey@unil.ch.

[2] Any mention of a vendor or product is not an endorsement or recommendation.

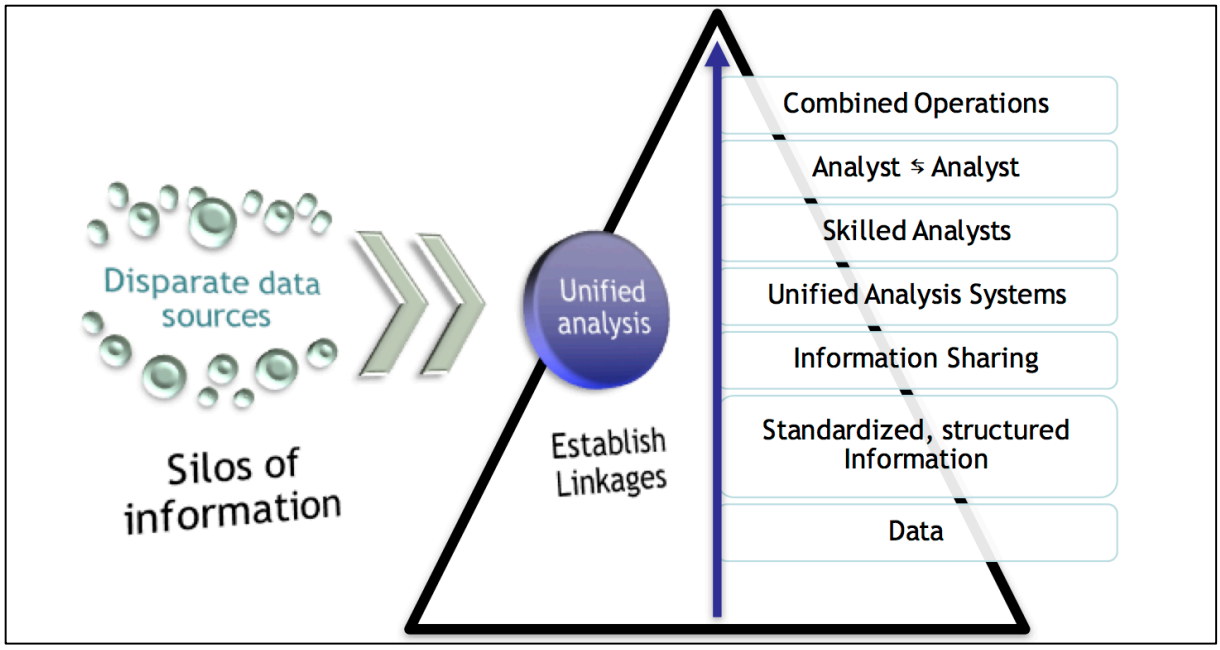

Figure 1: Standardizing helps break down data silos and increase visibility across multiple data sources, enabling more comprehensive correlation and analysis.

Another primary motivation for CASE is to enable more advanced and comprehensive correlation and analysis. In addition to fusing together disparate sources of information, CASE expresses information in a fully structured form that supports a multitude of analysis methods. In addition to searching for specific keywords or characteristics within a single case or across multiple cases, having a structured representation of cyber-investigation information allows pattern searching, graph query, data mining, and other sophisticated analytics. Improved capabilities to find important items can help solve a case, and more effective approaches to finding non-obvious similarities between cases can help overcome linkage blindness.[3]

> *Linkage blindness is an investigative failure to recognize a pattern which « links » one crime with another crime in a series of offenses…* (Vernon J. Geberth, Practical Homicide Investigation)

Overcoming linkage blindness can find connections between cases involving the same criminal activity in different countries, or the same crime pattern hitting different regions.

The capabilities, flexibility, and overall scope of CASE goes beyond all prior efforts to represent digital forensic information:

- The XML Data Encoding Specification for Intelligence Document and Media Exploitation (DOMEX) was developed by the U.S. government to share certain types of information, including mobile device details (ODNI, 2016). Although some elements in the DOMEX standard are used to keep track of provenance, the

[3] Linkage blindness: A term coined by a criminologist Steve Egger in the context of serial homicides to describe the failure to recognize that crimes were committed by the same offender because they occurred in different jurisdictions.

lack of supporting ontology and the inability to capture relationships limit the expressivity and flexibility of this standard.
- CybOX has been rebranded as STIX Cyber Observables, and concentrates on representing cyber threat intelligence (CTI OASIS).
- Digital Forensics XML or DFXML is a schema that is used by several tools to represent file system information (Garfinkel, 2009, 2012) and cross-verify extracted metadata (Nelson, 2014).

The initial draft of CASE has been released for broader community use and development (https://github.com/casework) along with the supporting UCO (https://github.com/ucoproject).

This paper provides a brief history of CASE and UCO, followed by an overview of the ontology and specification language.

Background

The foundation for this community driven standard was called DFAX, which included initial work on the UCO and concentrated on representing and sharing digital forensic information. DFAX utilized Cyber Observable eXpression (CybOX) to represent the purely technical information, such as binary artifacts and sources of matched search patterns (Casey, Back, Barnum, 2015).

CybOX was developed for cyber threat intelligence purposes, and had limitations in terms of representing digital forensic and cyber-investigation information. In 2016, CybOX was replaced by STIX Cyber Observables as an integrated component of the STIX standard, which focuses on cyber threat intelligence. The focus of STIX Cyber Observables which are embedded within the STIX schema makes it unsuitable as a foundation for representing various cyber-investigation use cases.

Leveraging all of the lessons learned from CybOX and DFAX, this standardization effort evolved into CASE and UCO to provide an improved data model and underlying ontology. CASE, as a specific profile of UCO, provides support for cyber-investigations in any context, including criminal, corporate and intelligence.

CASE and relevant portions of UCO build on the Hansken data model developed and implemented by the Netherlands Forensic Institute (NFI). Building on success of its precursor XIRAF, Hansken provides a robust platform that supports hundreds of investigations each year. The Hansken data model provides a solid foundation for developing CASE, including most common traces that are encountered in cyber-investigations, and flexible enough to add new types of traces.

Fully structuring data, which is supported by CASE and UCO, enables a wide range of analysis and correlation techniques. Other ontology-based efforts aimed at analyzing digital evidence have been narrowly focused, and can use CASE as their specification language. For instance, the Ontology for the Representation of Digital Incidents and Investigations (ORD2i) referenced DFAX and UCO, and provided a proof-of-concept implementation for timeline reconstruction and analysis (Chabot et al, 2015). The DESO ontology-based approach was proposed to represent known digital traces and to support triage searches of a digital crime scene for matching characteristics (Brady, et al

2015). The ParFor project also proposed an ontology-based approach to representing activities on computer systems (Turnbull, 2015). These efforts demonstrate that the need for a standardized way to represent and share cyber-investigation information is well recognized. Before UCO, there was limited agreement across the diverse community for such ontology. CASE and UCO address this gap with an ontology that can be used as a basis for community consensus and interoperability across tools and organizations.

The CASE specification language is not intended to define how tools or systems arrange data internally, but rather as a common language that applications can export and import to support interoperability and normalization. Developers of systems and applications can translate CASE to their internal implementations. The proposed JSON serialization is only one form of serialization, and the common format could be represented in XML, Turtle (RDF), protocol buffers, or other serializations.

UCO overview

The types of information to represent can be treated as layers, with the lowest layer representing raw information, the middle layer representing provenance, and the higher layer representing behavior.

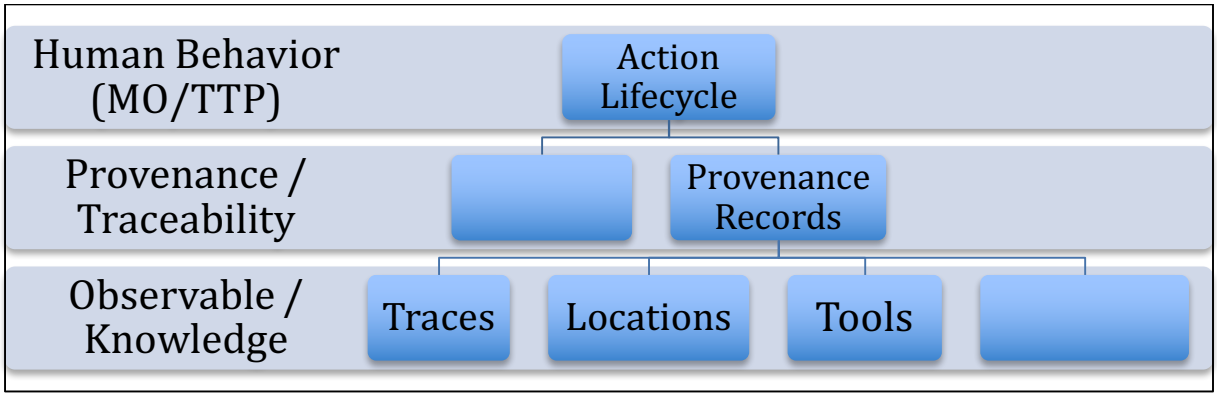

Figure A: Layers of representing cyber-investigation information

UCO provides an ontology that generalizes how this information is structured, and can be useful across multiple domains, including digital forensics, incident response, and counterterrorism.

CASE overview

A fundamental aspect of cyber-investigations is the extraction and analysis of traces, where the definition of a *trace* is any observable modification, including an absence of expected data, caused by an event in a digital crime scene (Casey, 2013). In the context of cyber-investigations, traces are used to address questions, which are generally described what, where, when, who, how and why. The state of a trace is also important to capture, such as whether it an item is allocated or deleted, or even whether an expected trace exists or does not exist.

The CASE specification language is flexible enough to represent a wide range of traces (corresponding to cyber-items as defined in the UCO) and their associated properties, including disks, devices, and file systems, providing a solid foundation for representing details within cyber-investigations. Figure 2 depicts a File object with multiple property bundles. The use of Property Bundles in CASE was inspired by the “duck” model

implemented in the Hansken system (van Baar, van Beek, van Eijk, 2014). Properties can include date-time stamps, the contents of a trace, and the hash values (e.g., MD5 and SHA256) of the data.

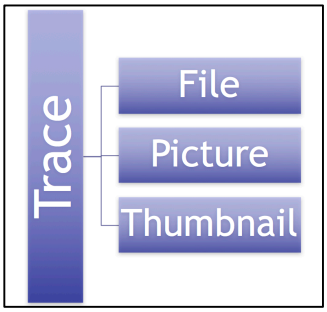

Figure 2: Duck model allows flexible representation of traces using various combination of property bundles.

Whereas DFAX utilized XML as its default serialization, in response to community input, CASE/UCO selected JSON-LD as the initial serialization binding (Lanthaler, Gütl, 2012). JSON is powerful and flexible but requires some scaffolding to support validation against an ontology. JSON-LD provides the necessary structure to support full validation of JSON content to its associated ontological specification as shown in Figure 3. The explicit validation enabled by JSON-LD yields assured integrity between the ontology and the serialization, and offers significant automation advantages including built-in API support for a range of languages (Python, Ruby, PHP, Go, C#, Java, etc.) and for lossless transformation between several serialization formats (JSON-LD, RDF-XML, Turtle-RDF, etc.).

```
{
 "@context": {
  "@vocab": "https://github.com/casework",
   },
 "@graph": [
   {
   "@id": "digital_photograph1",
   "@type": ”Trace",
   "propertyBundle": [
    {
     "@type": „File",
     "magicNumber": "/9j/4AAQSkZ",
     "mimeType": "image/jpg"
    },
    {
     "@type": “ContentData",
     “data": “/9j/4AAQSkZJRgABAQAAAQAB...",
     "size": 35000
    },
    {
     "@type": “RasterImage",
     "format": "jpg",
     "height": 12345,
     "width": 12345,
     "bitsPerPixel": 2
    },
    {
     "@type": "hash",
     "hashMethod": "MD5",
     "value": "3d137a188c1e82247b815209ce44af2c"
    },
    {
     "@type": “EXIF",
     "exifData": [
          "key": "Make“,
          "value": "Canon"
          …
  ]
 },
```

Figure 3: Example of CASE being used to represent a file. The JSON in this example is JSON-LD, which uses strict, namespaced @type values to specify the type for all JSON

objects, enabling their explicit traceability back to the specifications for these types in the UCO.

The ongoing community development of CASE is working to expand the specification language in order to cover other types of information such as Windows Registry entries, memory, and network traffic.

Provenance
For cyber-investigation purposes, to help establish the authenticity and reliability of information, it is important to capture where it originated or was found, as well as how it was handled after it was found. This lineage is collectively referred to as provenance.

Provenance includes collection documentation, chain of custody details, audit logs from forensic acquisition tools, and integrity records, which all help to establish the trustworthiness of cyber-investigation information. Provenance also involves tracking the data source and extraction method for each trace, such as a digital photograph obtained from a smartphone as shown in Figure 4.

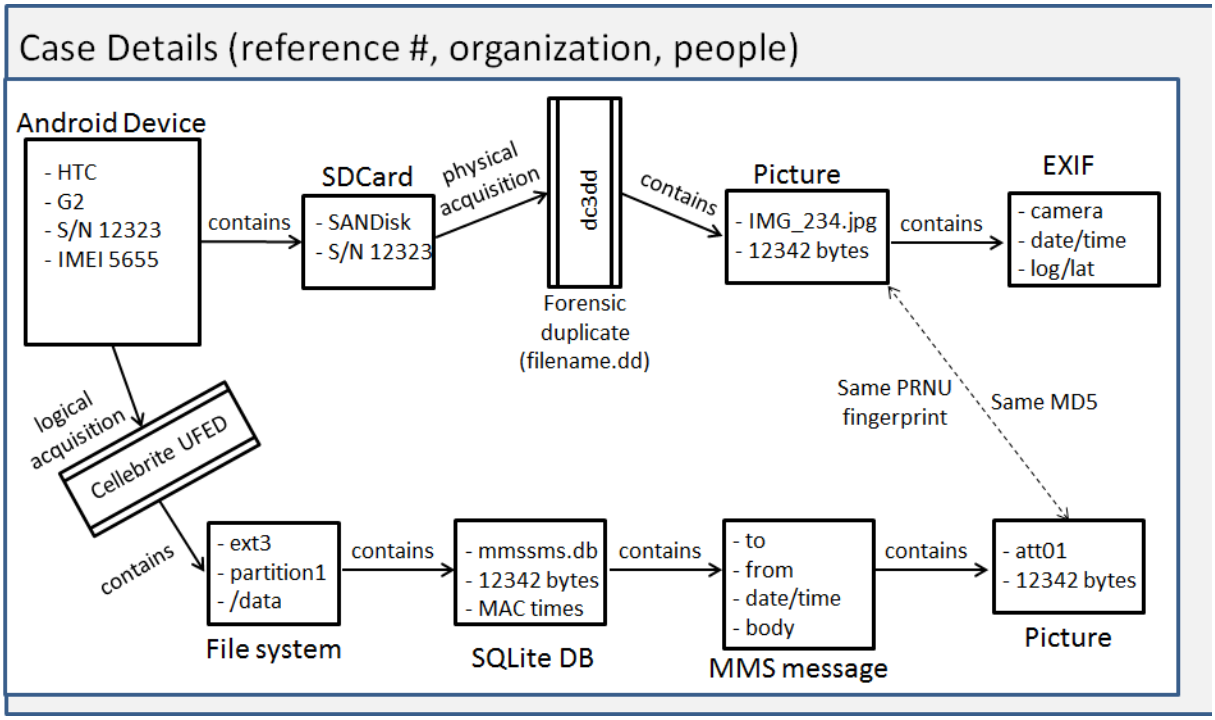

Figure 4: Provenance of a digital photograph extracted from an Android device.

CASE and UCO provide structures to represent all aspects of provenance in cyber-investigations, including chain-of-custody, case management, and forensic processing. CASE captures provenance information using Provenance Records which can include environmental characteristics such as the details of a crime scene or where the evidence was physically located. In addition, CASE captures information about any Forensic Action associated with each Provenance Record, as well as tracking who performed each Forensic Action and when it was performed (conceptual depiction in Figure 5). A basic example of a Forensic Action and Provenance Record is provided in Figure 6. In addition to supporting provenance, Forensic Actions can give insight into which tools and methods are effective in particular circumstances.

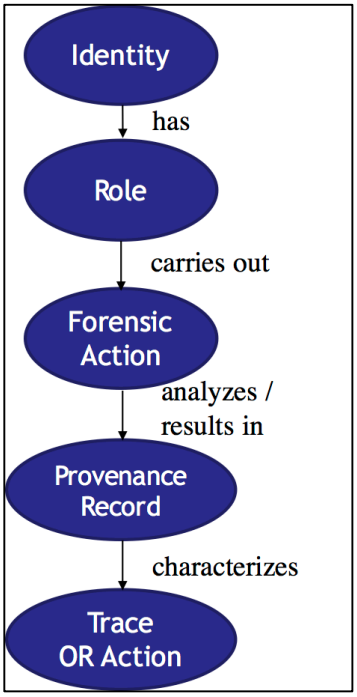

Figure 5: Conceptual depiction of representing provenance


```
  {
   "@id": "provenance_record1",
   "@type": "ProvenanceRecord",
   "description": "Android Smartphone",
   "exhibitNumber": "ACME-676553402357",
   "object": "device1"
  },
{
   "@id": "annotation1",
   "@type": "Annotation",
   "description": "Make forensic image of Android device »,
   "tags": [
    "forensic"
   ],
   "object": [
    "forensic_action23"
   ]
  },
  {
   "@id": "forensic_action23",
   "@type": "ForensicAction",
   "name": "imaged",
   "startTime": "2017-01-15T17:59:43.25Z",
   "endTime": "2017-01-15T19:59:43.25Z",
   "propertyBundle": [
```

```
    {
     "@type": "ActionReferences",
     "performer": "investigator1",
     "instrument": "tool1",
     "object": [
      "device1"
     ],
     "result": [
      "provenance_record1"
     ],
     "location": "forensic_lab1",
     "environment": "forensic_lab_computer1"
    },
    {
     "@type": "acme:ToolArguments",
     "aquisitionType": "Logical",
     "method": "ADB"
    }
   ]
  },
```

Figure 6: Example of forensic action and provenance representing using CASE.

Complete technical representation of the physical location where evidence was obtained and the people associated with the evidence can be covered by existing schemas. Therefore, rather than recreating a new representation of such information, it may be more effective to leverage an existing schema for such data. CASE and UCO have been designed to accommodate such re-use. Rather than include their own geolocation schema, they define an extension point where an existing schema can be used.

Fully-structured data in CASE

In addition to representing individual traces, it is important to capture their context and relationship with other traces and entities, for provenance and investigative purposes. CASE represents the linkages between items using a combination of embedded references for indelible properties and relationships for things that can change.

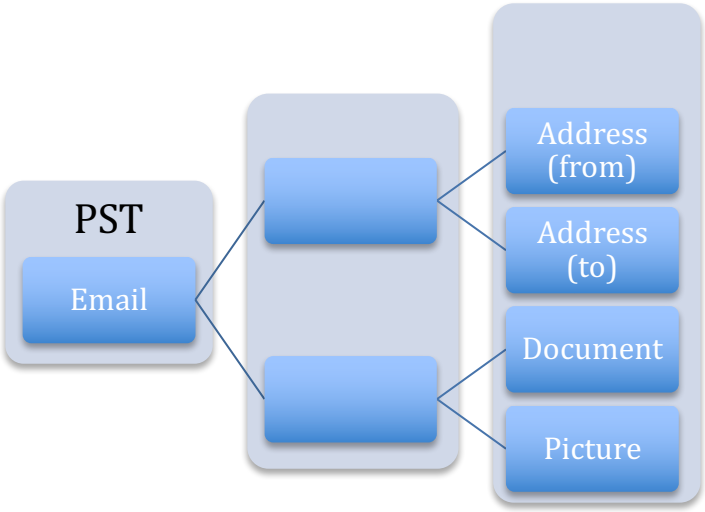

Figure 7: Structured representation of an Outlook PST file that contains an email message with multiple attachements

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 Figure 7 could be utilized to perform a graph search for all e-mail messages with a picture attachment from the subject to the victim (Casey, Back, Barnum, 2015).

Representing Actions in CASE

CASE and UCO provide a simple and adaptable way to represent an action or multiple related actions, which can be useful for sharing knowledge and supporting more advanced forensic analysis. In the context of cyber-investigations, CASE can represent actions involving digital traces such as a USB device being inserted into a computer and its associated traces (Casey, Back, Barnum, 2015). CASE can also be used to represent offender and victim actions and the associated traces. This type of abstraction can provide higher-level, human understandable portrayals of activities for more efficient forensic analysis.

Some forensic tools are adding features to support such abstractions 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. CASE provides a standardized way to represent these kinds of actions. Furthermore, beyond simply categorizing low-level artifacts, CASE can be used to define relationships between actions and traces, thus enabling more structured searches and refined analysis.

Action lifecycle

The Action Lifecycle from UCO can be adapted within CASE to define phases of a forensic investigation (e.g., documentation, preservation, examination, analysis, presentation) as

shown in Table 1. This generalized approach can be used to classify each action in a case, which provides context to support further analysis.

Forensic Process #1	Forensic Process #2	Forensic Process #3	Forensic Process #4
		Authorization	
	Planning	Planning	Preparation
Identification	Identification	Notification	
Preservation	Reconnaissance	Search	Incident Response
Collection	Transport & Storage	Collection	Data Collection
Examination		Transport	
Analysis	Analysis	Storage	Data Analysis
Presentation	Proof and Defense	Examination	Presentation of Findings
	Archive Storage	Presentation	
		Proof/defense	Incident Closure
		Dissemination	

Table 1: Forensic processes with different phases can be represented as an Action Lifecycle to categorize actions using CASE.

The SADFC phases in ORD2i (i.e., Extraction, Settlement, Enhancement, Analysis) can be represented in CASE as an Action Lifecycle. This type of information can be used to address various questions such as how much time was taken by each phase of an investigation, determining which tools are most useful for a given phase, and isolating which results were generated at different phases.

As shown in Table 2, the Action Lifecycle can also be used to categorize criminal activities such as a sexual predator's grooming of victims or a network intruder's method of operation (e.g., kill chain phases).

Grooming (Sexual Assault)	Kill Chain (Intrusion)
Victim selection	Reconnaissance
Establish trust	Development
Desensitization to sexual activity/abuse	Delivery
Maintain secrecy (persuasion/threats)	Exploitation
Arrange meeting	Configuration
Conceal evidence	Beaconing and C2

Table 2: The Action Lifecycle construct can be used to represent types of offender activities.

Guiding Principles

This initiative to structure and share cyber-investigation information strives to adhere to an implement a core set of guiding principles that community consensus has deemed necessary. Using lessons learned from DFAX, CybOX, and STIX, these principles are as follows:

Expressivity

In order to fully support the diversity of use cases in digital forensics, this initiative aims to address all defined use cases rather than focusing on a specific one. This goal involves covering all types of information relevant to digital forensics for various purposes.

Integrate rather than duplicate

Build on existing standardized representations, rather than create a separate one, to avoid redundancy and duplication of effort.

Flexibility

Avoid mandatory features to allow users to employ any portions of the standardized representation that are relevant for a given context.

Extensibility

Support community driven refinement and evolution of the language by building in extension mechanisms for domain specific use, for localized use, for user-driven refinements and evolution, and for ease of centralized refinement and evolution.

Automatability

Intentionally seek to maximize structure and consistency to support machine processable automation.

Readability

Create content structures to not only be machine-consumable and processable but also to, as much as possible, be human-readable. This human readability is necessary for clarity and comprehensibility during the early stages of development and adoption, and for sustained use in diverse environments going forward.

Conclusions and Next Steps

The CASE specification language and UCO support standardization and interoperability of tools, organizations, and countries dealing with cyber-investigations. In addition to sharing cyber-investigation information on a specific case, sharing traces or patterns of particular activities in a standardized format can help others find similar traces and patterns in new cases. Standardized representation of traces can also be useful for application footprinting by recording all traces of a given action (e.g., install, execute, uninstall). Sharing this kind of footprint information is a powerful means of facilitating digital forensic analysis and tool development.

Codifying and sharing information in a standardized manner enables digital investigators to search for similar patterns in their cases. 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).

Community development is ongoing to expand the types of information that CASE can represent. In addition, an API/library is under development to enable tools to “speak” CASE.

REFERENCES

Alink W, Bhoedjang R, Boncz P., de Vries A (2006) "Xiraf–xml-based indexing and querying for digital forensics" Digital Investigation

Barnum S (2014) “Whitepaper: Standardizing Cyber Threat Intelligence Information with the Structured Threat Information eXpression (STIX)” February 20, 2014, Version 1.1, Revision 1, http://stixproject.github.io/getting-started/whitepaper

van Beek HMA, van Eijk EJ, van Baar RB, Ugen M, Bodde JNC, Siemelink AJ (2015) “Digital Forensics as a Service: Game On” Digital Investigation, Special Issue on Big Data and Intelligent Data Analysis, Volume 15, pp 20-38. Elsevier

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

Casey E, Back G, Barnum S (2015) “Leveraging CybOX to standardize representation and exchange of digital forensic information” Proceedings of the 2nd annual DFRWS EU Conference, Digital Investigation, Volume 12, Supplement 1, Elsevier

Casey E (2013) “Reinforcing the Scientific Method in Digital Investigations using a Case-Based Reasoning (CBR) System” PhD Dissertation, University College Dublin

Chabot Y, Bertaux A, Nicolle C, Kechadi T (2015) "An Ontology-Based Approach for the Reconstruction and Analysis of Digital Incidents Timelines" Digital Investigation: Volume 15, December 2015, Pages 83-100, Elsevier: London (DOI: 10.1016/j.diin.2015.07.005)

Garfinkel, Simson. Automating Disk Forensic Processing with SleuthKit, XML and Python, Systematic Approaches to Digital Forensics Engineering (IEEE/SADFE 2009), Oakland, California.

Garfinkel S (2012) “Digital forensics XML and the DFXML toolset” Digital Investigation, Vol 8 161–174, Elsevier

Garfinkel SL, Malan DJ, Dubec K, Stevens CC, Pham C (2006) “Disk imaging with the advanced forensic format, library and tools” Research advances in digital forensics (Second Annual IFIP WG 11.9 International Conference on Digital Forensics) January 2006, Springer

Hargreaves C & Patterson J (2012) “An automated timeline reconstruction approach for digital forensic investigations”, Digital Investigation, Volume 9, Supplement (DFRWS2012 Proceedings)

Nelson AJ, Steggall EQ, Long DDE (2014) “Cooperative mode: Comparative storage metadata verification applied to the Xbox 360” Proceedings of the 14th annual DFRWS USA Conference, Digital Investigation, Volume 11, Supplement 1, Elsevier

Office of the Director of National Intelligence, "XML Data Encoding Specification for Intelligence Document and Media Exploitation,” https://www.dni.gov/index.php/about/organization/chief-information-officer/information-security-marking-access?id=1204