Contents lists available at ScienceDirect

Digital Investigation

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

# Advancing coordinated cyber-investigations and tool interoperability using a community developed specification language

Eoghan Casey [a, *], Sean Barnum [b], Ryan Griffith [c], Jonathan Snyder [c], Harm van Beek [d], Alex Nelson [e, 1]

[a] University of Lausanne, Switzerland
[b] MITRE, USA
[c] DoD Cyber Crime Center, USA
[d] Netherlands Forensic Institute, Netherlands
[e] National Institute of Standards and Technology, USA

ARTICLE INFO

*Article history:*
Received 3 March 2017
Received in revised form
4 August 2017
Accepted 5 August 2017
Available online 26 August 2017

*Keywords:*
Cyber-investigation
Digital forensics
Specification language
Standard representation
Unified cyber ontology
Information sharing
Digital evidence exchange
Evidence provenance
DFAX
DFXML
CybOX

ABSTRACT

Any investigation can have a digital dimension, often involving information from multiple data sources, organizations and jurisdictions. Existing approaches to representing and exchanging cyber-investigation information are inadequate, particularly when combining data sources from numerous organizations or dealing with large amounts of data from various tools. To conduct investigations effectively, there is a pressing need to harmonize how this information is represented and exchanged. This paper addresses this need for information exchange and tool interoperability with an open community-developed specification language called Cyber-investigation Analysis Standard Expression (CASE). To further promote a common structure, CASE aligns with and extends the Unified Cyber Ontology (UCO) construct, which provides a format for representing information in all cyber domains. This ontology abstracts objects and concepts that are not CASE-specific, so that they can be used across other cyber disciplines that may extend UCO. This work is a rational evolution of the Digital Forensic Analysis eXpression (DFAX) for representing digital forensic information and provenance. CASE is more flexible than DFAX and can be utilized in any context, including criminal, corporate and intelligence. CASE also builds on the Hansken data model developed and implemented by the Netherlands Forensic Institute (NFI). CASE enables the fusion of information from different organizations, data sources, and forensic tools to foster more comprehensive and cohesive analysis. This paper includes illustrative examples of how CASE can be implemented and used to capture information in a structured form to advance sharing, interoperability and analysis in cyber-investigations. In addition to capturing technical details and relationships between objects, CASE provides structure for representing and sharing details about how cyber-information was handled, transferred, processed, analyzed, and interpreted. CASE also supports data marking for sharing information at different levels of trust and classification, and for protecting sensitive and private information. Furthermore, CASE supports the sharing of knowledge related to cyber-investigations, including distinctive patterns of activity/behavior that are common across cases. This paper features a proof-of-concept Application Program Interface (API) to facilitate implementation of CASE in tools. Community members are encouraged to participate in the development and implementation of CASE and UCO.

## Introduction

Any investigation can have a digital dimension, often involving information from multiple data sources, organizations, and jurisdictions. Whether in court, battlefield or boardroom, decision makers need to have confidence that the information provided to them is trustworthy. Cyber-investigations support this need and,

[*] Corresponding author.
*E-mail address:* eoghan.casey@unil.ch (E. Casey).

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

http://dx.doi.org/10.1016/j.diin.2017.08.002
1742-2876/© 2017 Published by Elsevier Ltd.

in that role, are integrated with other domains, including digital forensic science, incident response, counter-terrorism, criminal justice, forensic intelligence and situational awareness. Therefore, to be effective, cyber-investigation information needs to be represented and shared in a form that is usable in any of these contexts, and is flexible enough to accommodate evolving requirements.

This paper describes a community-developed specification language called Cyber-investigation Analysis Standard Expression (CASE), which is intended to serve these needs. The primary motivation for CASE is interoperability – to advance the exchange of cyber-investigation information between tools and organizations (Casey et al., 2017a,b). The power of such a standard is that it provides a common language and structure to support automated normalization, combination, correlation, and validation of information, which means less time extracting and combining data, and more time analyzing information. CASE also supports data marking for sharing information at different levels of trust and classification, and for protecting sensitive and private information (Casey et al., 2017a,b).

CASE is a rational progression from the foundational work on Digital Forensic Analysis eXpression (DFAX), which focused on digital forensic information (Casey et al., 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 et al., 2015)

DFAX was created to represent and exchange digital forensic information, using Cyber Observable eXpression (CybOX) to represent the purely technical information, such as digital traces. Although intended as a representation for cyber observables independent of any particular usage context, the initial development priority of CybOX focused on supporting cyber-attack pattern detection and cyber threat intelligence. Because of this, CybOX had limitations in terms of representing some technical content specifically relevant to digital forensic and cyber-investigation information. Since its transfer to the OASIS standards body, CybOX has become much more closely coupled with STIX (Barnum, 2014) reducing its utility and flexibility for information representations other than STIX. In 2016, the independent CybOX was replaced by STIX Observables as an integrated component of the STIX standard, which focuses on cyber threat intelligence (Barnum, 2014). STIX Observables focus on objects relevant to attacks on computer systems, including executable files, processes, Registry keys, email messages, IP addresses, domain names, and URLs. In addition, STIX Observables are embedded within and dependent on the cyber threat intelligence context-specific structure of the STIX schema, which does not cover related domains such as incident response and digital forensic science. In short, STIX does not provide a suitable foundation for representing various cyber-investigation use cases that require more comprehensive expressivity for a wider range of digital traces and their context (e.g., file systems and smartphone apps), and that are bolstered by an ontological approach.

CASE is being developed in unison with the Unified Cyber Ontology (UCO). Leveraging the lessons learned from CybOX and DFAX, UCO provides an improved data model and underlying ontology from which contextually specific cyber-related representations can be defined. Enhancements to UCO have been made to support information representation across multiple cyber domains (e.g., incident response, digital forensic science, counter-terrorism), and to facilitate cross-domain exchange of cyber forensic intelligence. 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 the success of its precursor XIRAF, Hansken provides a robust platform that supports hundreds of investigations each year. The Hansken data model is a solid foundation for developing CASE, including most common traces that are encountered in cyber-investigations, and is flexible enough to add new types of traces (van Beek et al., 2015).

The novel contributions of this work include:

- Open community-developed specification language and ontology, with a proof-of-concept Application Program Interface (API) implementation, and examples of how to use CASE to support information exchange and tool interoperability;
- Alignment of ontology and data structures with existing forensic systems/tools to facilitate implementation and adoption by tool/system developers;
- Flexible data model (based on duck typing) that can be easily extended to represent any cyber-information and its properties;
- Formalized mechanisms to categorize and annotate *Traces* and *Actions*; including tracking forensic activities central to provenance in cyber-investigations;
- Use of JSON-LD as a default serialization to support full structural and semantic validation of all information in JSON serialized CASE content to the underlying ontological specification.

This paper starts with an overview of prior work and the evolution of CASE, and focuses on several use cases, encompassing the representation and exchange of extracted data and associated provenance details. An overview is provided of the kinds of information that can be represented by CASE, and the role of the underlying UCO is presented. Selection of JSON-LD as the initial serialization of CASE is explained.

The investigative scenario developed for this paper imagines The Oresteia by Aeschylus in the age of mobile devices. The purpose of this scenario is to show how CASE is used to capture information in cyber-investigations involving multiple related crimes to advance sharing, interoperability and analysis. A unifying CASE bundle representing this investigative scenario is provided in Appendix 2, and portions of the JSON are highlighted within the paper to illustrate specific aspects of CASE. The recommended identifier format is based on UUID, because the global uniqueness enables relationships to be defined across multiple cases and data sources. For readability, the examples for this paper use simplified labels instead of realistic UUIDs.

Example 1 shows the beginning of a CASE bundle containing multiple *Investigations*. Each *Investigation* contains a list of the associated elements that are defined in the remainder of the CASE bundle. To reduce repetitive examples in this paper, not every person in the scenario is explicitly represented using a complete *Identity* object. For illustrative purposes, each object that is referenced in this scenario uses the associated person's name in the simplified UUID (e.g. cassandra-device-uuid).