Government and Industry
Technical Briefing Memorandum

OPTIMIZING E-CRIME INVESTIGATION EFFICIENCY AND COOPERATION BY USING A COMMON DATA FORMAT

Electronic crime investigations are different from that of conventional law enforcement. The data that may be relevant to an electronic crime investigation is often voluminous and redundant. That data is almost always scattered across disparate jurisdictions and venues. Most obstructive is the fact that all these data, even when they are detected and collected by law enforcement, are archived in disparate file formats, precluding machine-based sharing and processing.

Evidentiary data mobilization for electronic crime forensics requires two fundamental developments. First is the establishment of a common data format in which evidentiary data can be archived, shared and processed by humans and, more importantly, by machines executing routines to pre-process them for forensic applications. Secondly, and much harder to address, the legal, regulatory and industrial policies require clarification to resolve real and perceived conflicts between data privacy laws and regulations and the needs of industry and law enforcement to exchange and process ecrime-related event data.

The latter imperative requires a legislative or treaty venue with the standing to address privacy laws. Toward the former, however, technology is available today to standardize the file formats used to archive, exchange and machine-process ecrime report records. The APWG has worked with its partners across its global membership base of 1800 institutions to develop an XML-based schema for reporting technical aspects of phishing, fraud, and other electronic crimes in a clear, consistent method. This schema, *Extensions to IODEF-Document Class for Reporting Phishing, Fraud, and Other Non-Network Layer Reports*, will be receive an official RFC number this year from the Internet Engineering Task Force (IETF), establishing it as a standard that can be employed without licensing fees or other encumbrances.

The goal of the data model is to allow an investigator to share relevant details of a possible criminal act with others in a data format that requires completeness, like local time-zone, while also providing multi-language support. Data shared in this format can be further processed easily by automation. For example, data about certain crimes can be automatically processed via computer upon arrival and redirected to the appropriate investigator in near-real time. Additionally, specific data elements can be controlled or encrypted to comply with evolving data privacy regimes.

The APWG has defined a set of extensions to the IETF Incident Object Data Exchange Format (IODEF) definitions as defined in IETF RFC 5070, a reporting standard for network events that was officially adopted by the IETF in December 2007. The IODEF is an XML-based data format designed to identify and describe network events such as virus infections, Denial of Service (DoS) attacks, or large scale malevolent scans by attackers. Each part of an IODEF report is specified through a schema definition indicating the data elements and their attributes. The schema also allows for implementers to specify which elements and attributes are required to assure that the essential ones are included within a report.

The APWG's ecrime reporting schema builds on the IODEF base specification by defining a set of data elements common to phishing, fraud, and other e-crime that allows the reporter of an event to specify the elements of the attempted crime, such as:

- The fraud source and target of a fraud attack, such a bank;
- The Web servers involved; data communication packets;
- Copy of the crimeware used in a specific e-crime event with a unique digital fingerprint used in a fraud scheme

- Domain Name Service (DNS) and registry information;
- Evidentiary files of a web site's content.

Technically, the schema is a series of defined report elements with associated data components that inform each element. Combinations of components within an element define the aspects of, for example, a phishing report and render that event as discrete pieces that can fit in the schema and, allow for machine manipulation and processing as well as consistent rendering for human interrogation.

The components of a PhraudReport are introduced in functional grouping as some parameters are related. Some elements may not make sense individually and have to be considered with other elements to completely illustrate a specific kind of e-crime event. A PhraudReport element, for example, is structured as follows.

Relevant information about a phishing or fraud event can be encoded by encoding the six components as follows:

```
+------------------+
|   PhraudReport   |
+------------------+
| STRING Version   |<>--{0..1}--[ PhishNameRef ]
| ENUM FraudType   |<>--{0..1}--[ PhishNameLocalRef ]
|                  |<>--{0..1}--[ FraudParameter ]
|                  |<>--{0..*}--[ FraudedBrandName ]
|                  |<>--{1..*}--[ LureSource ]
|                  |<>--{1..*}--[ OriginatingSensor ]
|                  |<>--{0..1}--[ EmailRecord ]
|                  |<>--{0..*}--[ DCSite ]
|                  |<>--{0..*}--[ TakeDownInfo ]
|                  |<>--{0..*}--[ ArchivedData ]
|                  |<>--{0..*}--[ RelatedData ]
|                  |<>--{0..*}--[ CorrelatedData ]
|                  |<>--{0..1}--[ PRComments ]
+------------------+
```

a. The *PhishNameRef* and *PhishNameLocalRef* elements are used to identify the fraud or class of fraud.

b. The *LureSource* element describes the source of the attack or phishing lure, including host information and any included malware.

c. The *DCSite* describes the technical details of the credential collection point.

d. The *Originating Sensor* element describes the means of detection.

The *RelatedData*, *ArchivedData*, and *TakeDownInfo* fields allow optional forensics and history data to be included. A specific phish/fraud activity can be identified using a combination of the *FraudType*, *FraudParameter*, *FraudedBrandName*, *LureSource*, and *PhishNameRef* elements.

As the extensions are XML-based, they can be processed with many freely available tools, and—as text—are readable without requiring special programs to display the data. (All web browsers will display XML formatted files as will most any popular word processor such as Microsoft Word. Any text editor (vi, emacs, nano, notepad, etc.) in any common operating system can display the XML-formatted content.)

The XML base also allows for significant improvements in report handling, as many of the report validation, collaboration, and distribution activities can be automated. With machine processing, many forensic routines requiring pains-taking hand processing can be executed as soon as relevant data becomes available to an application. In effect, the common data format can enable the automation of e-crime investigations, operating at the speed with which the electronic crimes themselves are executed.

With a common terminal format for e-crime reports, new forms of data sharing necessary to engage e-crime become possible in ways otherwise unimaginable without it:

- Private enterprises and their contractors can combine archived reports to detect larger trends and augment their fraud detection systems.
- Private enterprises and their contractors can share reports and e-crime event data in real-time to give all sharing parties earliest warning of new attacks that may concern them or their correspondents.
- Private enterprises and their contractors (e.g. banks and their security consultants) can quickly consolidate e-crime report databases to present a case to law enforcement.
- Private security firms can share data quickly and effectively to identify and track telling trends as well as indentify and characterize antagonists who are causing losses to their client companies.
- National computer emergency response teams, coordinating investigations into phishing attacks, can combine e-crime event databases to find corresponding data points in attacks launched in one country against targets in another.
- Public sector law enforcement agencies and private enterprises and security firms can combine e-crime event databases to analyze for trends and clues to inform case initialization.
- Public sector law enforcement agencies can quickly assemble relevant e-crime event data around a formerly unidentified suspect whose identity has been surmised as a correspondent in those electronic crimes.
- All parties to development of an existing law enforcement (or private security) case can program their systems to automatically direct reports of pre-determined characteristics to the appropriate investigators.

Ultimately the capacity to rapidly recruit, combine and analyze large disparate pools of e-crime data will suggest more automated mechanisms for e-crime detection and exposition. Furthermore, data fusion of summarized e-crime data with other established law enforcement data resources will redound, over time, to the development of potent e-crime investigative techniques that will make case initialization and development in the electronic realm as procedural as they for conventional law enforcement. Establishing a common data format is the first step toward that more efficient and cooperative future.

References

Extensions to the IODEF-Document Class for Reporting Phishing, Fraud, and Other Crimeware, Internet Engineering Task Force, July 2008. Click the "**View Document**" link:
https://datatracker.ietf.org/drafts/draft-cain-post-inch-phishingextns/

An open source software project relating to the development of tools for e-crime reporting can be found at: http://sourceforge.net/projects/ecrisp-x