An eCrime Reporting Lingua Franca:

Optimizing eCrime Investigation
Efficiency and Cooperation By
Using a Common Data Format

APWG

Committed to Wiping Out
Internet Scams and Fraud

March 2009

A Technical
Advisory for

Industry and
Government

Correspondent Author Contact Data:

Patrick Cain, APWG, pcain@antiphishing.org

Disclaimer: PLEASE NOTE: The APWG and its cooperating investigators, researchers, and service providers have provided this message as a public service, based upon aggregated professional experience and personal opinion. These recommendations are not a complete list of steps that may be taken to avoid harm from phishing. We offer no warranty as to the completeness, accuracy, or pertinence of these recommendations with respect to any particular registrar's operation, or with respect to any particular form of criminal attack. Please see the APWG website — http://www.apwg.org — for more information.

Principal Investigator:
Patrick Cain, Resident Research Fellow, APWG

Contributing Researchers

Dave Jevans, Chairman, APWG
Peter Cassidy, Secretary General, APWG

Summary

Historically, crime has been a local event; that is, the criminal is in close proximity to the victim. Electronic crime (e-crime) and variants on well-known criminal tactics that have been updated to use the Internet have removed this persistence of locality. The perpetrator of the crime and the victim may be separated by entire countries - or even continents. This adds new challenges for crime investigators as the party performing initial investigation may be quite remote from the actual crime "location" with different parties performing different parts of the investigation. The ability to convey accurate and complete investigative data—in multiple languages and styles—is now paramount to successful management of e-crime events, law enforcement case formation and subsequent prosecution.

To help with this information exchange, the APWG has worked with its partners across its global membership base of some 1800 institutions to develop an XML-based data model for reporting the technical aspects of phishing, fraud, and other electronic crimes to remote parties in a clear, consistent method. 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 quite easily by automation. For example, data about certain crimes can be automatically processed via computer on 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.

These factors make this data model an excellent vehicle to report, share, and interpret electronic crime events. Ultimately, the APWG believes that utilizing a common data format will allow many different useful forms of automated processing of forensic data, giving investigators and e-crime responders the kind of insights they require to transform large repositories of forensic data into actionable narratives that can animate potent e-crime management routines for private industry, as well as to assist in case formation, investigation and prosecution for law enforcement.

The Rationale For a Common eCrime Reporting Format

The rise in phishing and fraud activities via e-mail, instant messaging, DNS corruption and malicious code insertion has compelled corporations, Internet Service Providers (ISPs), consumer agencies and financial institutions to begin to collect, fuse, correlate and analyze phishing attack information and data related to e-crime events. The collected data allows them to better coordinate mitigation activities and support the pursuit and prosecution of attackers.

By using a common format, it becomes easier for an organization to engage in these coordination activities as well as to correlate information from multiple data sources or products into a cohesive view. As the number of data sources increases, a common format becomes even more important since multiple tools would be needed to interpret the different sources of data.

APWG SEES THE ENTERPRISE OF E-CRIME LAW ENFORCEMENT EVOLVING TO A MODEL RESEMBLING PUBLIC HEALTH INITIATIVES

The accumulation and correlation of information is also important to resolving phishing incidents detected externally, as the phished organization may not even be aware of the attack. Third parties aware of attacks may wish to notify the phished organization directly or through a central notification service or clearinghouse so that adequate responses could commence. The targeted organization's internal monitoring systems may also detect the attack and wish to take mitigation steps. If these systems cannot communicate adequately, there is no hope for attack mitigation or criminal prosecution.

APWG sees e-crime law enforcement endeavors gravitating over time to a model that more closely resembles public health initiatives like tracing the source of a contagious agent or a toxic substance in the food chain. It's all about how the case data are recruited. Instead of a report from one or a few sources used for case initialization and formation, in e-crime investigations a large number of data resources are collected, archived, collated and analyzed to build a summarized narrative to inform a new case and to contribute to and develop existing cases.

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 party of 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.

Criteria for Determination of Optimal Data Formats

When the APWG and its research correspondents began the development of a terminal format for e-crime reports, we attempted to find a suitable data model and format to adopt as a way of streamlining the effort — instead of starting from scratch. It became apparent that there was no existing, ready-to-adopt standard data format that met the criteria— so we developed one.

The important criteria for a worldwide, inter-domain data format become apparent very quickly. The format must allow for text data to be entered in a different language than the crime data. Take, for example, a Greek bank phishing e-mail message that is sourced from a

web server in Spain. The communications discussing the Greek text message between the investigating parties may be in Spanish. A second prime criterion allows for data elements to be marked as required or optional.

For example, a report about an Internet event is not very useful if the report does not include the attacker's Internet Protocol (IP) address. The bane of many international reports is the lack of required time zone markers on timestamps, e.g., the report states that the attack happened at two o'clock in the morning. Was this local time? Was it GMT? There must be a uniform way to precisely characterize time in order, for instance, to craft a coherent chronology of e-crime events in a case narrative.

Since the e-crime landscape is continually evolving, the data model and formats must be easily expandable to allow for the tracking of the new techniques discovered during investigations as well as the ability to add more data elements to existing reports.

A secondary, but still important criterion is to use a data format that does not require odd or expensive tools, nor encodes data in hard-to-decipher formats. Most all of parties exchanging investigative data have

> IN EFFECT, THE COMMON DATA FORMAT CAN ENABLE THE AUTOMATION OF E-CRIME MANAGEMENT AND LAW ENFORCEMENT, OPERATING AT THE SPEED WITH WHICH THE ELECTRONIC CRIMES THEMSELVES ARE EXECUTED

neither large budgets nor big support groups ready and able to figure out unclear reports. Thus, the ability to read and comprehend reports without complex tools is a necessity.

The IODEF Extensions for e-Crime Reporting

Failing to find an acceptable existing data format inspired APWG researchers to define a set of extensions to the IETF Incident Object Data Exchange Format (IODEF) definitions as defined in IETF RFC 5070[1], a reporting standard for network events that was officially adopted by the Internet Engineering Task Force (IETF) in December 2007.   (The IETF is an international body that, in part, develops technical and protocol standards for the operation and maintenance of the Internet.)

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 make sure that the important ones are included within a report.

The APWG's *Extension to IODEF-Document Class for Reporting Phishing, Fraud, and Other Non-Network Layer Reports* builds on the IODEF base specification by defining a set of data
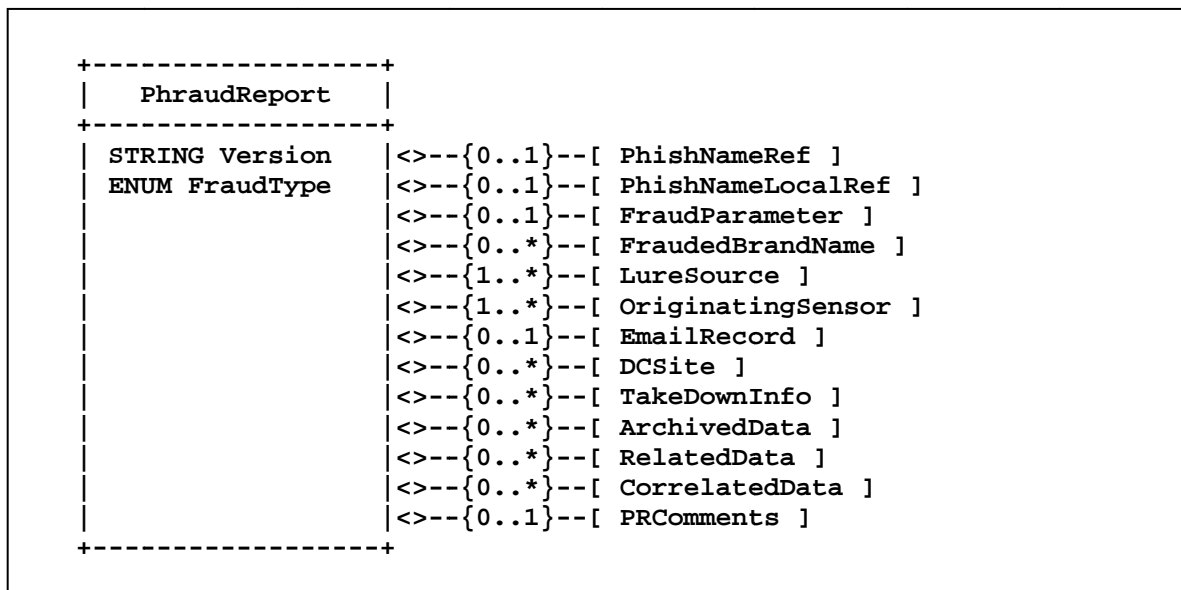
elements common to phishing, fraud, and other e-crime that allows the reporter of an event to specify the elements of the attempted crime, such as:

- The fraud source and target of an 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.

Figure 1: Report Element Example: The PhraudReport Element

```
    +------------------+
    |   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 ]
    +------------------+
```

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

 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 will display the XML formatted content as well.)

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 of the forensic routines that require pains-taking hand processing can be executed as soon as relevant data becomes available to a programmed forensic application. In effect, the common data format can enable the automation of e-crime management and law enforcement, operating at the speed with which the electronic crimes themselves are executed.

Enabling Robust Data Sharing

Significant operation efficiencies are possible if a common data format is shared amongst reporting and consuming parties. For example, once a reporter generates a report, it can be electronically sent to a database where the data could automatically consumed and redistributed.

Another party could request that data from the database, receive it in the common format, and use existing tools to decompose and examine it. This second party could also add additional data to the original report and return it to the original database using the same common format.

WITH A COMMON TERMINAL FORMAT FOR E-CRIME REPORTS, NEW FORMS OF DATA EXCHANGE TO ENGAGE E-CRIME BECOME POSSIBLE IN WAYS OTHERWISE UNIMAGINABLE WITHOUT IT

There are four communities currently using the IODEF data model and its extensions: national CERTs exchanging network incident data; a group of American and European financial institutions exchanging IP Addresses and associated fraud attack details; a number of ICT security companies and individuals reporting phishing attempts.

The enduring logistical challenges of international e-crime data sharing are:

- Finding a common data sharing and reporting format that supports multiple local languages;
- Providing adequate flexibility to evolve with the changing e-crime landscape;
- Ensuring that created reports contain sufficient and syntactically correct data.

The APWG believes that the *Extension to IODEF-Document Class for Reporting Phishing, Fraud, and Other Non-Network Layer Reports* meet these challenges and will continue to do so for a deep time horizon and will adapt to industrial and law enforcement needs for the foreseeable future.

The schema was developed to solve a few specific, but growing, identified problems such as exchanging information with speakers of other languages and trying to minimize the back-and-forth negotiation of capturing critical data. A few example use cases follow on page 11.

Figure 2: APWG e-Crime Reporting Tool
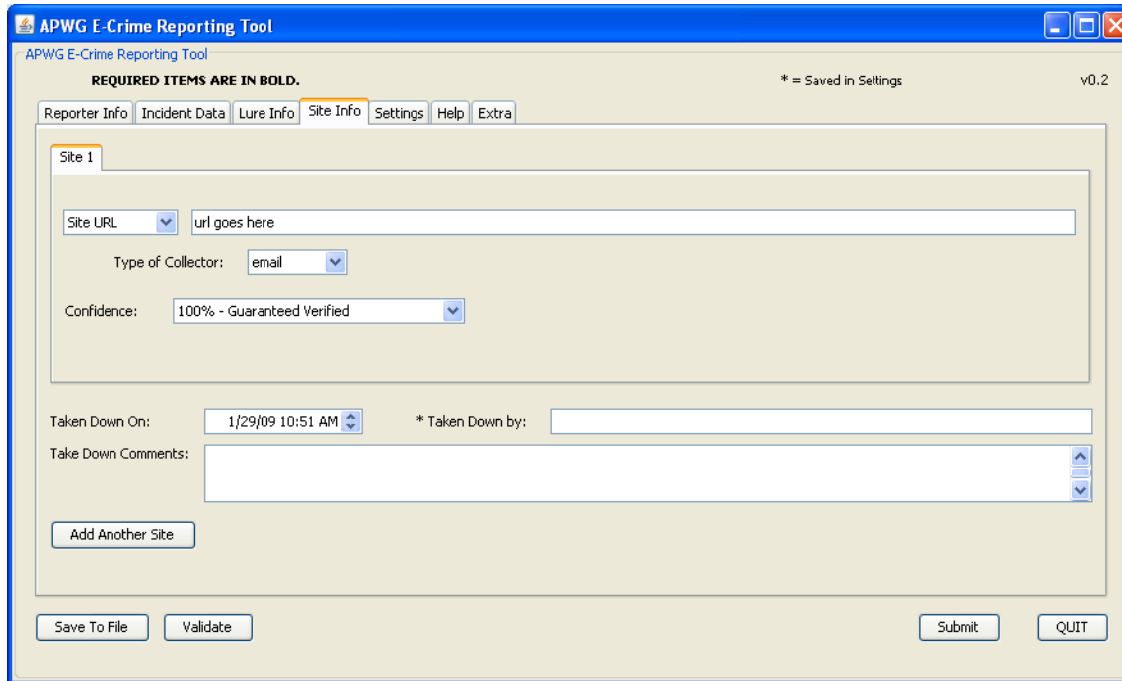
XML makes reports human readable and assists in editing, adding data and organizing human-driven workflows. The standardized format also allows for rapid consolidation of data and machine processing of reports for different forensic application scenarios

The APWG has established working betas of a compliant eCrime Reporting Tool for US-EN, UK-EN and ES-ES (Spain-native Spanish). More languages are to come. Goal: create a version of the APWG e-Crime Reporting Tool available in every language in which electronic crime is a problem to help establish and feed private sector, public sector and non-profit e-crime data repositories

A working beta of the APWG e-Crime Reporting Tool, a console that allows for detailed manual reporting and archiving of e-Crime events is available in US-EN at: http://sourceforge.net/projects/ecrisp-x

Use Case Scenarios and Associated Benefits

1. **Untrained sources who are preparing and sending crime data can be directed to send complete reports.** Veteran investigators and receivers of incident data all have horror stories of receiving incomplete data about an important crime event (e.g., the source data is missing; the timestamp has no time zone information; the data payloads are empty.) The IODEF exchange format can require certain critical data fields to ensure that a received report is complete. This ability to require certain fields makes it much easier to receive incident data from a wider reporting audience.

2. **Exchanging e-crime data with speakers of another language is simplified**. Many electronic crime events happen in multiple jurisdictions. An e-crime reporter may need to describe an event to an investigator that understands and/or reads a different language. The text areas in an IODEF format message contain a language marker that allows a receiver to use tools to semi-automatically translate the sentences, paragraphs and assorted information into a language that the receiver understands. Additionally, the sender of an IODEF-formatted report may also translate — and mark appropriately — the anticipated language of the receiver, making international information exchanges conducted in different languages quicker and simpler.

3. **The back-and-forth conversations when e-crime data is channeled via a third party can be reduced or eliminated**. Some portion of exchanged incident data is channeled through a third party before delivery to the intended receiver so the original data can be desensitized, anonymized or aggregated. For example, a group of banks may send their incident data to a clearinghouse at which it is compiled into a generalized report before forwarding to an investigator. If additional data is required during investigation of the compiled report, traditionally a series of back-and-forth communications ensues for the correspondents to determine which data should be used to inform reports and satisfy the report's requirements. When using the IODEF formats, one can send the entire — or pieces of— the compiled report back to the clearinghouse , which can send it back to the originator. The originator can quickly find their submitted reports using the included IODEF Incident Identifier instead of searching through all the data they submitted to the clearinghouse trying to match up the returned report and their internal data.

4. **Additional standard security mechanisms can be applied to the exchanged data as part of the data exchange process**. Some parties are sensitive to the data that is exchanged with other parties, such as ensuring that only the received party can read the message or requiring that the original data contain an integrity mechanism or digital signature. As the IODEF format complies with the XML encoding rules, any standard security

mechanism can be easily applied to the IODEF data, such as PGP, S/MIME, SSL/TLS, or generalized encipherment, and transported and recovered or validated at the receiving party. No additional security development is required to use these security services.

5. **International reporting requirements can be implemented**. E-crimes span multiple jurisdictions with multiple privacy, release, retention, and confidentiality requirements. Using the IODEF format to share data across national boundaries allows for the application of other standard XML security mechanisms -- irrespective of how the data is transported to the receiving party -- such as digital signatures and confidentiality (encipherment). The format also allows the ability to remove, encipher, or obfuscate certain data fields before further distribution to support compliance with national policy or regulation.

Looking Ahead: IODEF Extensions Development Arc

The establishment of a common terminal data format for e-crime reports is an essential element in the construction of a global counter e-crime data exchange infrastructure, indeed, a counter for the global e-crime plexus that is growing in competence and power with every passing day.

Still, there are other data logistics challenges that must be engaged in order to mobilize forensic data that is islanded in repositories maintained by industry, law enforcement agencies, government agencies, CERTs, NGOs and independent clearinghouses of e-crime report data.

First, tools for reporting and for translating existing data sets into IODEF-compatible reports will have to be established to take advantage of the common data schema. The APWG has already programmed a manual reporting console that will allow anyone who can operate a general purpose computer to construct a complete e-crime event report and archive it. [See Figure 2, p. 10.]

The APWG has established an open-source software project specifically to design and construct data translation tools to convert data sets in non-compliant formats of correspondents' systems into the IODEF Extensions for e-Crime Reporting format. Many of the conversion software routines are identical with only minor differences, but most all correspondent systems would require their own converters.

Given the eagerness of all stakeholders engaged in e-crime response and investigations to forensic exchange data, the APWG expects rapid organization of correspondence and the subsequent discovery of new advantageous data fusion and analysis schemes that will yield telling, conclusive intelligence for forensic artisans.

Researchers, investigators and responders from industry and the public sector will doubtless begin demanding more data and more frequent correspondence. Then, when the data-logistics barriers to facile, automated e-crime data exchange are removed, a global counter e-crime data infrastructure will have been established - and will flourish through clearinghouses and legally rational correspondence agreements.

Before that juncture, however, e-crime data correspondents will have to confront the legal, regulatory and complicated ethical questions that attend the movement of e-crime data across international frontiers and often through the hands of correspondents who may or may not have specific authorization or legal standing to handle those data. That process is underway in governmental bodies worldwide and is as vital to the development of a global counter e-crime data exchange as is the development of universal data schema for e-crime reports.

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 at:
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

---

[1] Danyliw, R., Meijer, J., and Y. Demchenko, *The Incident Object Description Exchange Format*, *RFC 5070*, December 2007. http://www.ietf.org/rfc/rfc5070.txt