Marking Data for Forwarding and Re-Sharing

Patrick Cain
Resident Research Fellow, APWG
President, The Cooper-Cain Group, Inc.

1.0	April, 2014	First Version
1.4	September, 2014	Modified for lessons learned in first pilot
1.5	May, 2015	Added picture and data model information

1 Introduction

Many parties collect Internet event data such as data such as IP Addresses, originator identification, or communications content to track network congestion, comply with regulatory regimes, or to detect malicious activity. Many times the data collected is not truly 'public' data but has handling and distribution restrictions or caveats on it. The APWG shares some data that carries some further sharing restrictions and is currently exploring ways to mark this data.

Most data or event sharing schemes include the ability to add a document sensitivity or classification marking to alert the recipient of the sensitivity of the data or its handling restrictions. For example, the IETF's IODEF XML format has an attribute at the top-level to choose one of four sensitivity markings – 'default', 'public', 'private', and 'need-to-know'. Those four choices are also available for marking specific sections of event logs or data, so a report can be marked with an overall sensitivity but have portions marked differently. Other data sharing formats (e.g., STIX, REN-ISAC) have equivalent functionality in the same or more – maybe 6 – markings. Other schemes have only three levels and invite creative combinations of the three values (e.g., TLP).

As data exchanging becomes more automated the challenge is to devise a marking scheme that can be unambiguously interpreted by a machine – without the need for human assistance. As an example, one may receive 10,000 or so reports of malicious web sites every day. Human review to determine data sensitivity of the reports' data items will significantly slow down the processing rate of the reports and possibly doom the data

http://www.apwg.org ● info@apwg.org
PMB 246, 405 Waltham Street, Lexington MA USA 02421

exchange. This paper presents a means to mark data to share within known groups that would support automation mechanisms.

2 The Problem

"The Problem" is really two distinct problems. First, a scheme is needed to properly mark data as it is received by the recipient to note its sensitivity. This (sensitivity) marking needs to be flexible enough to support a wide community of users, be not overly complicated to understand – particularly by automation systems, and be easily expandable as marks change and evolve over time. The sensitivity marks tell the recipient how to locally protect, and possibly re-share, the data. The second part of the problem is to devise a way to convey additional restrictions on the recipient. Both markings should unambiguously tell the recipient what they can do with the data after they receive it, for example, can they share it with others in their team or disclose details to other parties (who may be a victim of the event).

There is no way for those two problems to be solved with a relatively small - four, six, or eight – set of identifiers. And there is even a slimmer chance that multiple data sharing communities could agree as to the definitions of those identifiers. The next sections introduce a way to deal with both of the identified problems.

Note that our problem definition does not use these data sharing markings as a means to convey content sensitivity. Other marks are expected to be used for this purpose.

3 Our Data Sharing Model

To understand our problem and possible solutions requires some understanding of how the APWG receives and distributes data. In short, the APWG is a data clearinghouse: very little processing of the received data is performed before the data is forwarded to others. Our goal is to be a common point of data collection to make it easier to collect data.

The APWG forwards data to a set of recipients who are allowed to use the data for various purposes or to share the data further as explained in a contractual agreement.

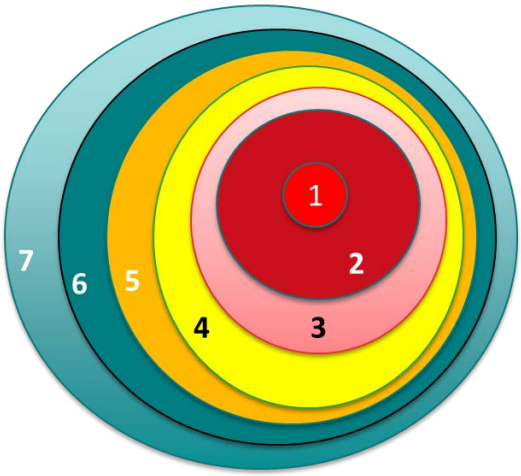
The purposes allowed to receivers of APWG data are roughly as follows. The data is:

- only for the recipient's use and should not be shared further.
- may be shared with the recipient's security team
- may be shared with other members of the recipient's organization
- may be used in products

http://www.apwg.org ● info@apwg.org
PMB 246, 405 Waltham Street, Lexington MA USA 02421

- may be shared with other security groups
- may be shared with the public

Pictorially, the purposes can be shown as a set of concentric circles, where each purpose is assigned a numerical value, such as:



- 1 - 'recipient only' or 'no further sharing'
- 2 - Coworkers in the security group
- 3 - Data incorporated into products
- 4 - Shared with affected users
- 5 - Shared within the company
- 6 - Forwarded to other security groups
- 7 - Shared with the public

Each circle includes the lower numbered circles There are more complex diagrams to show other relationships. For example, circle 2 could be split into two parts, one for friends of Pat (#2a) and one for enemies (#2b) of Pat. Data would be shared with the friends of Pat (#2a) but not his enemies (#2b). But the data could not be further shared as some enemies of Pat (in #2b) would get the data as part of circle #3 since the larger circles include the inner sets. Support for this more complex usage has been deferred until the concentric circle approach has been thoroughly tested.

http://www.apwg.org ● info@apwg.org
PMB 246, 405 Waltham Street, Lexington MA USA 02421

4 The Requirements

Means to express both recipient and re-sharing constraints leads one to a small set of requirements.

1. The solution should inform the recipient of the data what they can do with it. For example, can they share it with others in their company, disclose it publicly, etc. This is called the "sharing tag".
2. The solution should allow the sharer to add extra guidance, as in "Do not touch this system as it's under surveillance", or "Do not share it with Bob as we think he's a bad guy" or even "Public disclosure is embargoed until Tuesday at dawn". Recently the "share this data but don't include attribution" has become fashionable as more sensitive data flows among parties. This extra guidance or cautionary detail to be considered when evaluating, interpreting, or doing something is called a "caveat".
3. The apwg shares data between individuals, within groups, with other groups, and with the public. The solution needs to support all four without burdening the APWG operations staff.
4. The tags should be usable in multiple languages.
5. The tag should be easy to use in XML, CSV, or any other format-of-the-day.

The tags do not have to include all the policy implications of the data as sharing groups should have guidelines, maybe even contracts, to convey what the tags would imply. The sharing markings also do not have to convey data sensitivity marks. In many cases the "who can see it" implies certain sensitivities, and should be covered in the sharing group agreements.

5 Shoehorning Markings into Existing Structures

Our problem became visible when we started to share IODEF XML formatted data, which has four predefined tags. One solution was to redefine the restriction class in the IODEF schema to include other enumerations than the four defined in the standard. This has been tried with varying success. Many XML validation tools will mark the XML document as invalid since the IODEF schema doesn't except the non-standard enumerations. In some cases the standard IODEF schema can be modified to get around this problem but that requires all tools used by data sharers to use the new schema and a new version of the standard to be produced.

A second idea tried to redefine what the four classes meant, e.g., 'public' meant share with anyone, 'restricted' meant the recipient could share it with trusted parties, etc.. But it soon

http://www.apwg.org ● info@apwg.org
PMB 246, 405 Waltham Street, Lexington MA USA 02421

became evident that redefining the four markers would only add confusion as not everyone knew or agreed with the new interpretations.

Ignoring the IODEF constraintissues and looking at other commonly-used schemes was not fruitful either. A current favourite marking scheme is based on the Traffic Light Protocol (TLP) which defines four levels of sharing and sensitivity. Although the levels are 'red' (no sharing), 'amber' (some sharing) and' green' (more sharing) and 'white' (no restrictions) there have been 'black' (which I infer as a burnt out traffic light) and confusion abounds as to what the actual colours mean for further re-sharing of the data. There isn't enough information in four levels to support our sharing model, either, and although we could probably shoe-horn our groups into four levels there is still no way to add the localized caveats.

A real concern is having data marked as 'private' or 'amber' by two different communities with different numbers of tags and unequal definitions of 'private' and conflicting handling caveats and no means-contractually or programmatically to equate them. More operational experience and study will be necessary to alleviate this concern.

6 A DataMarkings Structure

As existing marking schemes seem inappropriate to our needs, a totally new structure was designed to hold all the data marking information. The marking scheme is structured as an XML blob since that allows for some easy testing and validation but the structure should work in other formats. The thing, labeled 'DataMarkings', would contain a sequence of markings for a particular community. Each 'community' element includes sensitivity and sharing tag identifiers as defined by and for that community. Different communities could define their own equivalency rules to deal with data crossing group boundaries.

For example, a dataMarkings structure that looks like:

```
<dataMarkings>
    <community name="apwg" version="1.0">
        <tag>3 - Friends</tag>
        <tag>2 – Enemies of Pat</tag>
    </community>
</dataMarkings>
```

would convey to a recipient that the data should be controlled and further shared as a level '3 - Friends' and a level '2 – Enemies of Pat' in the apwg community. Now, although the '2' and the '3' are the authoritative markers and are intended to help the automation systems,

http://www.apwg.org ● info@apwg.org
PMB 246, 405 Waltham Street, Lexington MA USA 02421

they may not have apparent meaning to a human so the <tag> could also be a defined data marking label like 'no sharing outside group' or 'sharing with public allowed'. The <tag> structure doesn't need to know this detail. Additionally, there are some paranoid communities where the community name may be sensitive so the structure also allows any text to be used -- e.g., community names generated by a hash or encryption or even random values.

The community string also carries a version identifier so communities can change, add, or remove markings without having to pick a different community name. The hope is that the version attribute will reduce the number of 'apwg', 'apwg-1', 'apwg-2' … 'apwg-1367' distinct community identifiers necessary in the future as the markings evolve.

Some thought has been given to defining two other attributes – 'until' and 'after' – to deal with embargoed data. For example, data may be 'no sharing allowed' until a point that an investigation is completed, then that data set becomes 'share with trusted groups'. Although the XML additions are straightforward, it has not been made part of the <dataMarkings> class until development of an acceptable CONOPS and use case is complete. In real operations it may be easier to re-share the embargoed data with a new mark at the embargo expiration than to have to support complex caveat logic.

6.1 Hierarchical versus distinct markings

The <dataMarkings> structure supports hierarchical and distinct marking schemes although the first pilots use hierarchical marks.. A community could design their marks to be very specific, e.g., 0 – recipients, 1-friends of Pat, and 2 – friends of Bob. If we wanted to share with friends of Pat and friends of Bob the mark would need both an entry for'1' and for '2'. There is no means to generate an "only trusted insiders" mark as it seems illogical as how would one know? The only case where this seems to make sense is to mark data as "only the infected system owner" if you are sharing the data with someone who has contact information for the infectee. The <dataMarkings> structure may be simplified if such a tag is really implemented as a caveat, which is our current plan.

7 Carrying Complex Markings into XML Documents

Another attribute of the community element is the 'alias' attribute. In IODEF and other XML formats, the generator of a report may mark specific parts of the report with more restrictive markings. For example, a spam report may mark the whole report with a 'public' mark but mark the <History> element with a 'good guys only' as the history may include active investigative data.

http://www.apwg.org ● info@apwg.org
PMB 246, 405 Waltham Street, Lexington MA USA 02421

The alias attribute allows the report originator to designate a short-hand marking for use later in the document. A more complex example is:

```
<dataMarkings>
    <community name=”apwg” version=”1.2”
alias=”private”><tag>3</tag><tag>restrictive</tag></community>
</dataMarkings>
```

Note that the <alias> class performs the same functions as the ‘shoehorning’ mentioned above, except by reusing existing <restriction> enumerations there is no need to modify the existing IODEF or STIX schemas. The bad news is that there are still only four choices to ‘alias’ and the access control routines that process the report need to be aware of the equivalent markings. So although the structure supports it there are not many actual uses expected.

Although proposed as more of a test feature, it has many advantages over adding additional <dataMarkings> structures and reissuing all the format standards.

8 New XML Data Classes

This section defines the <dataMarkings> structure as an XML-Document. Although it can be used in other formats XML allows for some testing and guided implementations.

8.1 The structure

The overall structure is two lists of values:

```
BEGIN
    List of sharing tags (identifier, sharing-value)
    List of caveats (identifier, value)
END
```

The initial sharing tags in the APWG community would be:

0 - Recipient only
1 - Community
11 - Internal Summary
13 - Internal Details

21 - Impacted Party Summary
23 - Impacted Party Details
33 – Used in Products
41 - Trusted Summary

http://www.apwg.org ● info@apwg.org
PMB 246, 405 Waltham Street, Lexington MA USA 02421

43 - Trusted Details
91 - Public Summary
99 - No Restrictions

This list supports our requirement to support the APWG sharing model in a hierarchical way. The numerical values were picked to allow easy (and fast) comparison in software. A higher value tag implies the lower values, so a tag value of 31 – Trusted Summary, implies that the data can be shared with the community and internal groups.

Trying to define an initial set of caveats was more challenging. Although there are a number of sharing constraints it is unclear which of those constraints are valid in the APWG sharing model. An initial set of caveats are below but generating an acceptable caveat list will probably take quite some time . The use of non-numerical values should reduce confusion with tag values.

NA - No originator attribution
HI – Historical Data
AI – Active Investigation, do not disturb or contact

8.2 A More International-Friendly Syntax

One concern is that non-English speakers may not adequately comprehend the descriptive portions of the sharing tags. A slight modification to the syntax could help this by modifying the descriptive portion of the tag, as in: <tag>11 – Internal Summary</tag>
would change into <tag value=”11” lang=”en”>Internal Summary</tag>
or for a Spanish version: <tag value=”11” lang=”sp”>Resumen interna</tag>

This new encoding would allow the descriptive field to be translated into local languages but the actual tag value would stay the same to optimize processing. Note that this modification would be useful for XML-encoded data markings where the extra bytes needed to encode the language tag do not significantly add to the length of the tag which is untrue for other non-XML encodings.

8.3 XML Schema Definition

To help the tag definitions an XML schema is being developed. It is not final but is included here for information.

```
<?xml version="1.0" encoding="UTF-8"?>
<!--
```


http://www.apwg.org ● info@apwg.org
PMB 246, 405 Waltham Street, Lexington MA USA 02421

```
 ============================================================
 ===      Top Level Class:  dataRestrictions                ===
 ============================================================

 This schema was developed by Pat Cain, APWG. <pcain@apwg.org>

 This schema defines data classes and elements incorporated within an
 XML element.to express the data sensitivity and further distribution markings for an event
report.

 The markings are expected to be used thusly:
 <dataRestrictions>
   <community name="apwg"><tag> only</tag></community>
   <community name="apwg" alias="six" version="1.0"><tag>6</tag><tag>members
only</tag></community>"
   <caveat>HI</caveat>
 </dataRestrictions>
   History:
.1    08/2011 - First cut
.2    01/2012 - Validation succeeds; accompyaning document written
.3    09/2014 - Add more levels to the apwg1Tags element.
.5....06/2115 - Modify markings from pilot trials.

-->
<xs:schema elementFormDefault="qualified"
        targetNamespace="http://apwg.org/schemas/dataMarking-1.0"
        version="1.0" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
        xmlns:xs="http://www.w3.org/2001/XMLSchema"
        xmlns:marking="http://data-marking.mitre.org/Marking-1"
        xmlns:hfp="http://www.w3.org/2001/XMLSchema-hasFacetAndProperty"
        xmlns:apwgMarkings="http://apwg.org/schemas/dataMarking-1.0">
 <xs:annotation>
   <xs:documentation>This document is copyright © 2012, 2014, 2015 by the APWG,
   www.apwg.org. Comments ans suggestons can be submitted to the principal
   research fellow pcain@apwg.org.</xs:documentation>

   <xs:documentation>This APWG developed this document as a means to mark
   shared data as not all datum submitted to a data cleainghouse may be
   appropriate to share with a wide audience.Initial trials with existing
   marking sets led us to define a more flexible, extensible, set of multiple marking
   options.</xs:documentation>

   <xs:documentation>This set of marks allows for a "community" mark to distinguish different
   marking sets (tags) . Each community defines a set of tags to mark data in accordance
```


http://www.apwg.org • info@apwg.org
PMB 246, 405 Waltham Street, Lexington MA USA 02421

```
  with their policies and operating model.
  Communities may also develop, as neccessary, optional 'caveat' tags that allow for more
  restrictive multi-lingual guidance. Communities are encouraged to develop
  their own sets of community and caveat structures.</xs:documentation>
 </xs:annotation>

<xs:annotation>
 <xs:documentation>The following import is to support STIX encodings, where the markings
  need to be an extension of a defined class.
 </xs:documentation></xs:annotation>
 <xs:import namespace="http://data-marking.mitre.org/Marking-1"
        schemaLocation="../../../STIX/data_marking.xsd"></xs:import>

<xs:annotation><xs:documentation>The goal is to get something like this output:
<dataMarkings><tag value="1">Recipient Only</tag><caveats>tag value="HI">Historical
Data</caveats></dataMarkings>
</xs:documentation></xs:annotation>
 <xs:complexType name="dataMarkingStructureType">
  <xs:complexContent>
   <xs:extension base="marking:MarkingStructureType">
    <xs:sequence>
     <xs:element maxOccurs="unbounded" name="tag"
            type="apwgMarkings:apwg1Tags" xml:lang="en-US"></xs:element>
     <xs:element maxOccurs="unbounded" minOccurs="0" name="caveat"
            type="apwgMarkings:CaveatType"></xs:element>
    </xs:sequence>

    <xs:attribute default="0.5" name="version" type="xs:string"></xs:attribute>
   </xs:extension>
  </xs:complexContent>
 </xs:complexType>
<xs:annotation><xs:documentation>This definition is here so we don't have to import all of the
IETF IODEF schema.</xs:documentation></xs:annotation>
 <xs:complexType name="MLStringType">
  <xs:simpleContent>
   <xs:extension base="xs:string">
    <xs:attribute default="en-US" name="lang" type="xs:language"
            use="optional"></xs:attribute>
   </xs:extension>
  </xs:simpleContent>
 </xs:complexType>

 <xs:complexType name="CaveatType">
  <xs:simpleContent>
```


http://www.apwg.org • info@apwg.org
PMB 246, 405 Waltham Street, Lexington MA USA 02421

```
    <xs:extension base="xs:string">
          <xs:annotation><xs:documentation>
          The choices here are
          "No Original Attribution", "Historical Information","Active Investigation"
          </xs:documentation></xs:annotation>
     <xs:attribute default="en-US" name="lang" type="xs:language"
use="optional"></xs:attribute>
      <xs:attribute name="tag">
       <xs:simpleType>
        <xs:restriction base="xs:string">
         <xs:enumeration value="NA"/>
         <xs:enumeration value="HI"/>
         <xs:enumeration value="AI"/>
        </xs:restriction>
       </xs:simpleType>
      </xs:attribute>
      </xs:extension>
   </xs:simpleContent>
 </xs:complexType>

 <xs:complexType name="apwg1Tags">
 <xs:simpleContent>
  <xs:extension base="xs:string">
  <xs:annotation><xs:documentation>
    The permitted values are:
    tag #  Contents
     0 - Recipient only
     1 - Community
    11 - Internal Summary
    13 - Internal Details
    21 - Affected Party Summary
    23 - Affected Party Details
    33 - In Products
    41 - Trusted Summary
    43 - Trusted Details
    81 - Public Summary
    99 - No Restrictions
    </xs:documentation></xs:annotation>
   <xs:attribute default="en-US" name="lang" type="xs:language"
use="optional"></xs:attribute>
      <xs:attribute name="tag">
       <xs:simpleType>
        <xs:restriction base="xs:integer">
                <xs:minExclusive value="0"/>
```


http://www.apwg.org ● info@apwg.org
PMB 246, 405 Waltham Street, Lexington MA USA 02421

```
                <xs:maxExclusive value="99"/>
            </xs:restriction>
          </xs:simpleType>
        </xs:attribute>
      </xs:extension>
    </xs:simpleContent>
  </xs:complexType>
</xs:schema>
```

Note: The schema is probably broken, being it is XML. Check the github for updates.

9 A Staged STIX Example

The following STIX-Document shows placement and an example use of the markings. Some fields have been compacted for display.

```
<STIX_Header>
   <Title>Example Report for Scanning for open ssh servers</Title>
   <Package_Intent xsi:type="stixVocabs:PackageIntentVocab-1.0">Indicators -
Network Activity</Package_Intent>
      <Profiles>
         <stixCommon:Profile>apwg.org:scan-general-1</stixCommon:Profile>
      </Profiles>
      <Handling>
         <marking:Marking>
            <marking:Marking_Structure marking_model_ref="apwg1"
            xsi:type="apwgMarkings:apwgMarkingStructureType">
               <apwgMarkings:tag value =”99”>No
Restrictions</apwgMarkings:tag>
            </marking:Marking_Structure>
         </marking:Marking>
      </Handling>
      <Information_Source>
      …
```

10 Use in CSV formats

Although we specified the tags and caveats in XML they should work in CSV sharing communities. The community, tag, and caveats could be encoded as community/tag/caveats followed by a comma. As in

,apwg/11 – Internal Summary/no attribution .

Some sharing communities may be able to specify shortcuts. If the community uses the apwg tags, and really wants to save space, the data marking could be

http://www.apwg.org ● info@apwg.org
PMB 246, 405 Waltham Street, Lexington MA USA 02421

,11/NA,

Other formats should be able to support our markings in a similar manner.

11 APWG Pilot Use of <dataMarkings>

APWG researchers have proposed multiple communities for the collection and sharing of data and incorporated the marks into a test data repository. Some of the actual policy guidance to mark data are still under development and are repository and community dependent and the definitions are quite fluid; do not rely on them for operational use.

The current XML schema and CSV guidance are available at github.com/patCain/ecrisp.

12 Further Considerations

The use of these marking is still in development and the operational situations are still evolving. Although a draft CONOPS is in the works, comments, suggestions for improvement, and operations models that break the concept are always appreciated – particularly if you share data in a compatible data model as the APWG's.

13 References

Danyliw, R., Meijer, J., & Demchenko, Y. (2007, December). *The Incident Object Description Exchange Format (RFC 5070).* Retrieved January 2012, from Internet Engineering Task Force: ftp://ftp.isi.edu/in-notes/rfc5070.txt

Traffic Light Protocol, http://en.wikipedia.org/wiki/Traffic_Light_Protocol

Structured Threat Information eXchange, http://stix.mitre.org