Marking Data for Forwarding and Re-Sharing

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

September, 2014

Version 1.4

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.

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

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

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.

3 The Requirements

Looking into ways to express both sensitivity 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 sensitivity 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 guidance is sometimes 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.

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

4 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 base IODEF schema can be modified to get around this problem but that requires all tools used by data sharers to use the new schema.

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 became evident that redefining the four markers would only add confusion as not everyone kept up with the new interpretations.

Ignoring the IODEF constraints 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', 'amber' and' green' 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. There isn't enough information in four levels for our sharing model, and although we could probably shoe-horn our groups into four levels there is still no way to add the localized sensitivity markings.

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 the concern.

5 A DataMarkings Structure

A possibly solution is to craft a totally new structure to hold all the data marking information. This is our current plan. We structured it 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:

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

```
<dataMarkings>
    <community name=”apwg” version=”1.0”><tag>3 - Friends</tag></community>
</dataMarkings>
```

would convey to a recipient that the data should be controlled and further shared as a level ‘3 - Friends’ in the apwg community. Now, although the ‘3’ is the authoritative marker and is intended to help the automation systems, it 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.

5.1 Hierarchical versus distinct markings

The <dataMarkings> structure only supports hierarchical marking schemes. There is no means to generate an “only trusted insiders” mark as it seems illogical. 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.

6 Carrying 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’

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

mark but mark the <History> element with a 'good guys only' as the history may include active investigative data.

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></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 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 iti don't expect many actual uses.

Although proposed as more of a test feature, it has many advantages over adding additional <dataMarkings> structures to almost every place where sensitive data could be populated in an IODEF-Document. Or a STIX document. Or anybody else's format.

7 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.

7.1 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 could be:

0 - Recipient only
1 - Community
11 - Internal Summary
13 - Internal Details
21 - Trusted Summary
23 - Trusted Details

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

41 - Affected Party Summary
43 - Affected Party Details
81 - 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 21 – 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 additional and local values are expected. The use of non-numerical values should reduce confusion with tag values.

NA - No attribution
NP - No public sharing until
AP - Only share with affected party.

Generating an acceptable caveat list will probably take quite some time. The community marker will be important for the caveats as they are expected to be quite a fluid set.

7.2 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>

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.

The apwg1 example markings would now look like this:

Tag value	Tag text description
0	Recipient Only
1	Community
11	Internal Summary
13	Internal Details

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

21	Trusted Summary
23	Trusted Details
41	Affected Party Summary
43	Affected Party Details
81	Public Summary
99	No Restrictions

7.3 XML Schema Definition

To help the tag definition an XML schema was developed. The latest version is available at github.com/patCain/ecrisp in the schema folder. The structure as an XML class called <dataMarking> could be defined as follows.

Note: The schema is probably broken. Check the github for one that should work.

```
<xs:schema xmlns:xs=http://www.w3.org/2001/XMLSchema elementFormDefault="qualified"
    targetNamespace="apwg.org/schemas/dataMarking-1.0"
    xmlns:marker="apwg.org/schemas/dataMarking-1.0"
    xmlns:iodef="urn:ietf:params:xml:ns:iodef-1.0">
    <xs:import namespace="urn:ietf:params:xml:ns:iodef-1.0"
        schemaLocation="iodef-1.0.xsd" />
<xs:complexType name="apwgMarkingStructureType">
    <xs:complexContent>
    <xs:extension base="marking:MarkingStructureType">
    <xs:sequence>
        <xs:element maxOccurs="unbounded" name="tag" type="apwgMarkings:apwg1Tags" xml:lang="en-US"/>
        <xs:element maxOccurs="unbounded" minOccurs="0" name="caveat" type="apwgMarkings:CaveatType"/>
    </xs:sequence>
    <xs:attribute default="1.0" name="version" type="xs:string"></xs:attribute>
    </xs:extension>
    </xs:complexContent>
</xs:complexType>
<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>
<xs:extension base="xs:string">
<xs:attribute default="en-US" name="lang" type="xs:language" use="optional"/>
<xs:attribute default="Do" name="shareWith">
<xs:simpleType>
<xs:restriction base="xs:string">
<xs:enumeration value="Do"></xs:enumeration>
<xs:enumeration value="Do Not"></xs:enumeration>
</xs:restriction>
</xs:simpleType>
</xs:attribute name="Until" type=xs:Date use="optional" />
```


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

```
</xs:extension>
</xs:simpleContent>
</xs:complexType>

<xs:simpleType name="apwg1Tags">
<xs:restriction base="xs:string">
<xs:enumeration value="0 - Recipient only"></xs:enumeration>
<xs:enumeration value="1 - Community"></xs:enumeration>
<xs:enumeration value="11 - Internal Summary"></xs:enumeration>
<xs:enumeration value="13 - Internal Details"></xs:enumeration>
<xs:enumeration value="21 - Trusted Summary"></xs:enumeration>
<xs:enumeration value="23 - Trusted Details"></xs:enumeration>
<xs:enumeration value="41 - Affected Party Summary"></xs:enumeration>
<xs:enumeration value="43 - Affected Party Details"></xs:enumeration>
<xs:enumeration value="81 - Public Summary"></xs:enumeration>
<xs:enumeration value="99 - No Restrictions"></xs:enumeration>
</xs:restriction>
</xs:simpleType>
</xs:schema>
```

8 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>
      …
```

9 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

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

,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

,11/no attribution,

Since we do not share lots of csv format data, a working example has not been tried, but we have faith that it will work.

10 APWG Pilot Use of <dataMarkings>

APWG researchers have proposed multiple communities for the collection and sharing of data. Some of the actual guidance policies to mark data are still under development and are repository and community dependent. These 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.

11 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.

12 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

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