CTI-TC Working Session

Meeting Date:	**December 4, 2018**
Time:	**3:00 p.m. EST**
Purpose:	**Weekly Working Session**

Attendees:

Allan Thomson – Moderator
Trey Darley
Sarah Kelley
Gary Katz
John-Mark Gurney
Rich Piazza
Sean Barnum
Nicholas Hayden
Tom Vaughn
Dr. Masato
Jason Keirstead
Chris Ricard
Jane Ginn – Recorder
Vivek Jain

Agenda:

- **Cyber Observables – How handle moving forward**

Meeting Notes:

Allan Thomson

There has been a mini-group on the two key proposals – we are discussing here with
Larger group today, during the regular working call

Problem Summary

We **agree** that an ID is required for SCOs with the following properties

It should be possible to deterministically compute on both creation (producer side) and useful for search (consumer side)
Its easy to create (for both sides)
It can be referenced by relationships across transactional/individual units of intel (i.e. bundles)
The ID will be computed on a subset of SCO properties <- mini-group consensus last week

We **need to work** on

A) How does each implementation interoperate including what needs to be defined in the spec for preferred subset
B) How an ID is computed for the set of properties chosen for subset of properties

Commonalities of Two Proposals

Producer SHOULD use an identifier-template defined in the STIX specification for the SCO
Exact format of identifier template is tbc (later slides)
Producer MAY use a different identifier-template than that defined in the STIX specification for the SCO
Producer MUST pass an identifier-map of non-standard templates used as part of a STIX Bundle either directly as a STIX object or via reference to an externally published version for use by external organizations
Identifiers MUST use an identifier-template to specify how the id is generated.

John-Mark Gurney

[*Asked a question for clarification on two vendors and the use of IDs*]
Then, asked if there was only 1 ID on these objects

Allan Thomson

[*Noted that in the future – there could be multiple IDs – but we want to get agreement on this first*]
[*Went over the Pros & Cons*]

ID Proposals (pro/con not agreed)

Option Summary	Description	Pro (*as provided by proposer and does not represent agreement*)	Con (*as provided by proposer and does not represent agreement*)
Base 85 + SHA1 Hash-based	1) SCO Ids are Hash values based on defined set of properties 2) Define STIX pattern grammar like definition of all attributes that contribute to ID (see next slide) 3) Include in grammar what special character is inserted to avoid	1) Smaller size than UUIDs 2) Fast to compute 3) Reuses existing definition language (with minor tweaks) that is used for SCO 4) Is deterministic	1) Different IDs from Objects (this could be considered a pro as well)
UUID-based	1) ALL ids continue to use the same form: <object-type>—UUID 2) The UUID portion of the id is either a UUID4 or a UUID5	1) Does not result in large number of IDs for the same object compared to UUID4 [Required by high volume sensor providers]. Even less than for current proposal as producers can use less granular semantic equivalence-based determinism 2) Allows full integrity for versioning of cyber observables 3) Does not depend upon receiving final universal consensus across all stakeholders on semantic equivalence for each object. Very quick to implement 4) MUCH smaller change from status quo. Does not require any current implementation to change unless they want to	1) Does not guarantee auto-correlation of exact match objects across producers

Sean Barnum

[*Clarified that the Pros and Cons for the UUID proposal are overall, not compared to the other proposal*]

John-Mark Gurney

The SHA1 standard is compromised – we may need to consider

Gary Katz

Would that we relevant in this case?

John-Mark Gurney

If you keep all of the data for the Object, then the SHA1 could be used, if you don't, then it could be A problem, even for this case.

Gary Katz

Could you send out some links on this afterwards

Chris Ricard

Are the two approaches being debated to ensure semantic equivalency

Allan Thomson

We want an approach where we can have an ID that can be used in multiple Use Cases

Chris Ricard

Then, it is deterministic…?

OK, then, gave an example of a scenario… Noted that different users would assign different IDs

Allan Thomson

The mini-group has discussed this about the mapping… We agreed that there will never be consensus On what parameters to be used… that is why we are deferring to the STIXPreferred persona

The specification allows flexibility…

the Interoperability is where we have agreed upon for the Use Cases

Chris Ricard

If it is something that is negotiated on a case-by-case basis…

then you can have agreement on a specific Use Case

Allan Thomson

[Had problem with word 'negotiation'… but, agreed in principle]

Gary Katz

Gave a clarifying point about how different vendors will use their own parameters could be with a library

Chris Ricard

If this is for searching…Why not just hash the value?

Sean Barnum

The primary Use Case is for de-duping… not query. The secondary Use Case is between vendors

Then what we all do to make it easier for the users… but, primary Use Case is de-dupe

Jason Keirstead

What Chris is bringing up is very relevant… a legitimate problem with the whole idea…

This could be a problem… I realize this is a compromise

Allan Thomson

We all agree… Let's try and find a compromise…. From a standards POV, what we need to do is

Find a compromise

Chris Ricard

I don't know that this solves anything

Jason Keirstead

What this seems like to me is that we need to flush this out further…

or we will have a problem with TC

The problem I have with this…. What is the Business Value….If the IDs are different from producer-to-Producer

Allan Thomson

In a large community, we need a mapping….[*Gave example of AIS feeds*]

Jason Keirstead

Made argument that the intra-vendor Use Cases have not been articulated

Gary Katz

Here is one: High Speed sensor – need a way to use same ID for the same object

A second one – As a producer of intel, as a producer, I have made a determination that these are

Correlated… you need to provide that as a service

The third Use Case… I have multiple, different vendors… in that case, I'll need to correlate

With these proposals… we have a discovery process with a mapping structure

Jason Keirstead

Gave an argument – Do not agree

Sean Barnum

Asked about how to keep track of relationships

Jason Keirstead

They are internal issues… it is not about sharing.

I understand that internally, that you guys have a graph

Allan Thomson

This is not about individual organizations… Made a point about custom properties – will not fix issue

If that is where we disagree… Is that what you are actually disputing

Jason Keirstead

We can have STIX top-level objects that would be linked to a Cyber Observable… then, can't use STIX

Gary Katz

It is about how it is computed… not what the ID is

Sean Barnum

We do this to scale… so it is implementable… [*Gave examples of file, network traffic, email different*]

We've seen different players have different perspectives…

What Allan was saying was that Optionality will help with specific Use Cases… without trying to

Achieve the 5% without breaking the 95%

Chris Ricard

It seems to me the ID should not be indicating semantic equivalence… it should be a different object

[Gave a proposed solution of a 'Semantic Equivalence' object that could relate to the CO]

Made argument that they should not be de-deduped

John-Mark Gurney
I agree with Chris
Gary Katz
For those that are opposing this… could you please outline how your organization would do this
Please outline
Jason Keirstead
I keep having a problem. We would have to throw these things away
I still have to match UUID4s – We still have to do string matches….
Allan Thomson
At the beginning of this mini-group… there are some Use Cases that it will help
Sarah Kelley
If it solves some Use Cases, then let's use it
John-Mark Gurney
If we can solve the security issue… the other problem is that not all fields are hashed
It solves one problem and raises another problem
Gary Katz
Can I respond to that real quick? [*Gave an example of different vendors using different hashes*]
You are still getting a correlation between different producers… if have a mapping
John-Mark Gurney
The problem is that for additional context properties that you are linking to… you cannot use
That – the hash will be different
Gary Katz
Have an identifier map for that Producer that would allow me to distinguish it from others
John-Mark Gurney
Correct… but it is not handled in this proposals
That is why I like Jason's idea of having a custom property…
Trey Darley
Gave example, being inside CERT, we are having problems with correlating things… It didn't work
For Malware, Infrastructure and Incident… What we are aiming at is the best alternative
To no agreement
Allan Thomson
So, what was proposed was an attempt to provide Optionality… but, it sounds like some are having
Problems…
We collectively have to find a compromise that works for everybody…
Trey Darley
The problem is getting worse for all of the market sectors… we have a societal imperative to find a
Solution
Sean Barnum
How do we move forward that does not block for some of us to move forward… For those of you
That are having problems… Please be specific
We got to this point in the Mini-group for some very specific Use Cases
There are things with STIX that we can't do…
John-Mark Gurney
I heard a proposal here… you create a third-party object [*Chris's suggestion as given above*]
Allan Thomson
It was discussed in the Mini-Group, and then was discounted.
John-Mark Gurney
It should be presented to the larger group, so we could debate it.
Gary Katz
The reason it was discounted was that it did not meet the specific Use Case of the high-speed sensor.

Allan Thomson

We are running out of time… we'll have to discuss this later. Thank you all for joining us.

Rest of Slide Deck Information:

Identifier Templates: Option #1

- Use STIX Pattern Grammar variation
  - Use the terms defined in the STIX pattern grammar and the concatenation terms
  - and define SHA1 hash on the result of the expression
    - [email.type:value FOLLOWEDBY email.is_multipart:value]
- For optional fields we could have
  - [email.type:value FOLLOWEDBY (email.is_multipart:value OR "\ff")]
  - If is_multipart was an optional field value
  - \ff or similar would be chosen to avoid ambiguity

Identifier Template: Option #2

Option #2

```
// Identifier-Map Example:
{
  type: "identifier-map",
  namespace: "acmecorp.com",
  identifier_map: {
    arglebargle: "${fieldName1}, ${fieldName2}"
  }
}
```

```
// Object Example:
{
  type: "arglebargle",
  fieldName1: "foo",
  fieldName2: "fee",
}
```

The STIX genid library could take the object and identifier-map as arguments and return a valid id

The genid call into the python UUID5 library would look something like this:

```
uuid.uuid5("acmecorp.com", "foofee")
```

Meeting Terminated

****