Labeling Cybersecurity Data for AI Automation (Single- and Multi-Modal)

Industry Connections Activity Initiation Document (ICAID)

Version: 1.0, 9 September 2020

IC20-010-01 Approved by the IE&SS SMDC 8 October 2020

Instructions

- Instructions on how to fill out this form are shown in red. It is recommended to leave the instructions in the final document and simply add the requested information where indicated.
- Shaded Text indicates a placeholder that should be replaced with information specific to this ICAID, and the shading removed.
- Completed forms, in Word format, or any questions should be sent to the IEEE Standards Association (IEEE-SA) Industry Connections Committee (ICCom) Administrator at the following address: industryconnections@ieee.org.
- The version number above, along with the date, may be used by the submitter to distinguish successive updates of this document. A separate, unique Industry Connections (IC) Activity Number will be assigned when the document is submitted to the ICCom Administrator.

1. Contact

Provide the name and contact information of the primary contact person for this IC activity. Affiliation is any entity that provides the person financial or other substantive support, for which the person may feel an obligation. If necessary, a second/alternate contact person's information may also be provided.

Name: Dejan Milojicic
Email Address: dejan.milojicic@hpe.com
Employer: Hewlett Packard Enterprise
Affiliation: Hewlett Packard Labs

IEEE collects personal data on this form, which is made publicly available, to allow communication by materially interested parties and with Activity Oversight Committee and Activity officers who are responsible for IEEE work items.

2. Participation and Voting Model

Specify whether this activity will be entity-based (participants are entities, which may have multiple representatives, one-entity-one-vote), or individual-based (participants represent themselves, one-person-one-vote).

Individual-Based

3. Purpose

3.1 Motivation and Goal

Briefly explain the context and motivation for starting this IC activity, and the overall purpose or goal to be accomplished.

Cyberanalysts are becoming a bottleneck in analyzing ever increasing amounts of data. Automating cyberanalysts actions using AI can help reduce amounts of work for analysts and thereby reduce time to outcome dramatically, record actions in knowledgebases for training of new cyberanalsts and in general open up the field for new opportunities. As a result, the state of cybersecurity will improve.

IEEE is engaging in running the Grand Challenge of Applying AI/ML to Cybersecurity (see the links provided in Section 3.3, also: https://ieeexplore.ieee.org/document/8909930). Running Grand Challenge will enable collecting data and behaviors of actors participating. This data can then be used to train AI.

To effectively use data, requires its labeling, which is the initial focus of standardization in this area.

It is envisioned that this program will bring together industry stakeholders to engage in building consensus on priority issues for standardization activities on these topics, and providing a platform for IEEE thought leadership to industry. The activity will liaise with IEEE-USA and its Grand Challenge initiative to enable collaboration and attract additional participants where there are common interests between the activity and Grand Challenge participants. As standards opportunities are identified, the activity participants will partner with IEEE Societies to place new standards projects that align with Society areas of expertise.

3.2 Related Work

Provide a brief comparison of this activity to existing, related efforts or standards of which you are aware (industry associations, consortia, standardization activities, etc.).

MITRE STIX/TAXII: it is part of OASIS, one of the largest technical committees in history (299 members as of last year), with broad participation (Multi-nationals, start-ups, universities, government agencies, consultants) and representatives from many industries (Financial services, healthcare, software, research, public sector).

STIX stands for Structured Threat Information Expression, it is a language for sharing cyber threat intelligence. It is based on JSON and designed for sharing. STIX has RESTful API, which simplifies development of TAXII 2 applications. It supports Domain objects, such as: *Attack Pattern, Campaign, Course of Action, Identity, Indicator, Intrusion Set, Malware, Observed Data, Report, Threat Actor, Tool, Vulnerability*. It observes objects such as: *Artifact, AS, Directory, Domain Name, Email Address, Email Message, File, IPv4 Address, IPv6 Address, MAC Address, Mutex, Network Traffic,*

Process, Software, User Account, URL, Windows Registry Key, X.509 Certificate. More details are available at https://www.oasis-open.org/committees/cti/

TAXII 2: TAXII is an open protocol for the communication of cyber threat information. TAXII enables authenticated and secure communication of cyber threat information across products and organizations. (from: https://taxiiproject.github.io/taxii2/). It is also part of the OASIS CTI TC.

Ontology Working Groups ISO: A big part of the labeling effort is to standardize the process for creating labeling rules which can then be applied automatically to captured data. In order to do this, we will need to translate the knowledge around what the label means and how its applied. One way to do this is by using ontologies to capture the associations between observed data and their meanings. Converting these ontologies to labeling rules requires that the ontologies follow a standard structure and format.

CDF (Based on Elastic Common Schema or ECS): The Common Data format is an attempt within the US government to standardize on a common schema so that the data is structured in a way so that it can be used to train and evaluate AI. Currently, each vendor has their own unique format for logs. This hampers the ability to aggregate the data since something as simple as date is not agreed on.

Enterprise Data Header 2 (EDH2)/Trusted Data Format (TDF): The EDH2/TDF effort is to secure the transport of sensor data across the network. If the data from the sensor is corrupted, delayed, or maliciously changed, it would serious hamper an analyst ability to make decisions. The EDH2/TDF effort is to guarantee that the data was securely transferred from a verified source to its intended destination.

3.3 Previously Published Material

Provide a list of any known previously published material intended for inclusion in the proposed deliverables of this activity.

We had published a report https://www.ieee.org/content/dam/ieee-org/ieee/web/org/about/industry/ieee_confluence_report.pdf that is high level description of applying AI/ML to Cybersecurity, so we will potentially use some aspects of this paper as an introduction to standard.

We also had published an IEEE Computer paper (https://ieeexplore.ieee.org/document/8909930), which describes the approach of the grand challenge for which we intend to use standards.

3.4 Potential Markets Served

Indicate the main beneficiaries of this work, and what the potential impact might be.

There are multiple markets, primarily starting from the security community in industry, followed by the benefits to governments agencies, and finally academia for doing research.

3.5 How will the activity benefit the IEEE?

This activity will enhance IEEE visibility in these key and ground breaking cybersecurity topics. It will allow the development of a vibrant technology industry community, complemented by the high visibility of the IEEE USA Grand Challenge initiative.

There will also be longer term revenue opportunities for IEEE, initially via the Grand Challenge and associated government funding, and longer term through this Industry Connections activity as appropriate opportunities are identified for memberships, funded workshops and associated new products (e.g. education).

4. Estimated Timeframe

Indicate approximately how long you expect this activity to operate to achieve its proposed results (e.g., time to completion of all deliverables).

Expected Completion Date:

09/2022

IC activities are chartered for two years at a time. Activities are eligible for extension upon request and review by ICCom and the IEEE-SA Standards Board. Should an extension be required, please notify the ICCom Administrator prior to the two-year mark.

5. Proposed Deliverables

Outline the anticipated deliverables and output from this IC activity, such as documents (e.g., white papers, reports), proposals for standards, conferences and workshops, databases, computer code, etc., and indicate the expected timeframe for each.

The initial activity will be to build a cybersecurity industry community of participation during late 2020, and drafting of proposed project activities with associated timelines. It is envisioned that preliminary recommendations, and initiation of new standards proposals can be achieved by mid-2021. This will align well with the Grand Challenge that will run in 2021.

This will be followed by development of schemas for labeling data for cybersecurity, and ensuring compatibility with the standards described above, such as STIX. Consensus on tests for labeling data will also be included.

5.1 Open Source Software Development

Indicate whether this IC Activity will develop or incorporate open source software in the deliverables. All contributions of open source software for use in Industry Connections activities shall be accompanied by an approved IEEE Contributor License Agreement (CLA) appropriate for the open source license under which the Work Product will be made available. CLAs, once accepted, are irrevocable.

Will the activity develop or incorporate open source software (either normatively or informatively) in the deliverables?:

Yes, at minimum informatively to analyze labels and work with them.

6. Funding Requirements

Outline any contracted services or other expenses that are currently anticipated, beyond the basic support services provided to all IC activities. Indicate how those funds are expected to be obtained (e.g., through participant fees, sponsorships, government or other grants, etc.). Activities needing substantial funding may require additional reviews and approvals beyond ICCom.

This activity will primarily require the typical Industry Connections activity support provided by SA staff. Any needs for longer term financial support, if identified, will be addressed by a funding plan as necessary (e.g., participation fees, workshop fees, etc.)

It is expected that some of the government sources will sponsor the Grand Challenge, and that aspect will be managed by IEEE USA, with liaison to this IC activity for information exchange and community development.

7. Management and Procedures

7.1 Activity Oversight Committee

Indicate whether an IEEE committee of some form (e.g., a Standards committee) has agreed to oversee this activity and its procedures.

Has an IEEE committee agreed to oversee this activity?: No

Note that liaison relationships will be established with IEEE-USA who will oversee the overall effort of Grand Challenge. Liaison relationships will also be established with IEEE Communications Society, IEEE Computer Society and other IEEE societies as they align with anticipated standards needs and recommendations.

If yes, indicate the IEEE committee's name and its chair's contact information.

N/A

IEEE Committee Name: Chair's Name:
Chair's Email Address:

IEEE collects personal data on this form, which is made publicly available, to allow communication by materially interested parties and with Activity Oversight Committee and Activity officers who are responsible for IEEE work items.

7.2 Activity Management

If no Activity Oversight Committee has been identified in 7.1 above, indicate how this activity will manage itself on a day-to-day basis (e.g., executive committee, officers, etc).

There will be officers (chair/vice-chair) and/or an executive committee with officers.

7.3 Procedures

Indicate what documented procedures will be used to guide the operations of this activity; either (a) modified baseline *Industry Connections Activity Policies and Procedures*, (b) Standards Committee policies and procedures accepted by the IEEE-SA Standards

Board, or (c) Working Group policies and procedures accepted by the Working Group's Standards Committee. If option (a) is chosen, then ICCom review and approval of the P&P is required. If option (b) or (c) is chosen, then ICCom approval of the use of the P&P is required.

The activity will use the modified baseline Industry Connections P&Ps

8. Participants

8.1 Stakeholder Communities

Indicate the stakeholder communities (the types of companies or other entities, or the different groups of individuals) that are expected to be interested in this IC activity, and will be invited to participate.

Government agencies, e.g., DHS, NSA, NREL
Industry, e.g., Hewlet Packard Enterprise,
Academia, e.g., GaTech

8.2 Expected Number of Participants

Indicate the approximate number of entities (if entity-based) or individuals (if individual-based) expected to be actively involved in this activity.

20

8.3 Initial Participants

Provide a number of the entities or individuals that will be participating from the outset. It is recommended there be at least three initial participants for an entity-based activity, or five initial participants (each with a different affiliation) for an individual-based activity.

Use the following table for an entity-based activity:

Entity	Primary Contact	Additional Representatives

Entity Name	Contact Name	Name

Use the following table for an individual-based activity:

Individual		**Employer**	**Affiliation**
Daniel Bennet		NREL	
Kirk Bresniker		Hewlett Packard Enterprise	Hewlett Packard Labs
Tom Coughlin		Contractor	
Ada Gavrilovska		GaTech	CERCS
Ethan Hamilton		Entrepreneur	
James Holt		Laboratory of Physical Sciences	
Paul Jones		Turing Institute	
John Johnson			
Dejan Milojicic		Hewlett Packard Enterprise	Hewlett Packard Labs
Vane Russel		DHS	
Mark Sherman		CMU SEI	CERT
Trung Tran		Amplio	