ATHENA: A FAIR approach to publish and evaluate cybersecurity datasets*

Thaisa da S. Hernandez[1,4,*,†], Caroline Duarte Gandolfi[1,†], Pedro Henrique Bulcão[1,†], Luiz Bonino da Silva Santos[2,†], Anderson F. P. dos Santos[1,3,†] and Maria Cláudia Reis Cavalcanti[1,†]

[1] *Instituto Militar de Engenharia, Praça Gen. Tibúrcio 80, Urca, Rio de Janeiro, RJ, 22290-270*

[2] *University of Twente, Drienerlolaan 5, 7522 NB Enschede, The Netherlands*

[3] *Venturus Centro de Inovação Tecnológica, Av. G. V. di Napoli 1185, Bosque das Palmeiras, Campinas, SP, 13086-530*

[4] *Diretoria de Comunicações e Tecnologia da Informação da Marinha, Rua 1º de Março, 118, Centro, Rio de Janeiro, RJ, 13086-530*

Abstract

The massive increase in the attack surface caused by an exponential volume of data has highlighted the importance of continuous research in the field of cybersecurity, which in turn has become increasingly data-driven. The availability and quality of cybersecurity datasets are therefore fundamental for the reliability of predictions and the implications in innovation in this domain. However, there are numerous challenges regarding the availability of good-quality cybersecurity datasets. This work addresses these challenges by proposing an approach to publish cybersecurity dataset metadata and to assess the quality of these datasets, considering their specific properties. The differential of our approach is the integration of the FAIR (Findable, Accessible, Interoperable, Reusable) principles into the evaluation process. This approach was implemented as a composite of software modules. First, a FAIR Data Point repository was instantiated to publish metadata about cybersecurity datasets. Secondly, the Athena Evaluator module was implemented to analyze the metadata published in the repository based on a set of specific quality metrics and on metrics aligned with the FAIR principles. Additionally, to support the creation and management of different metadata schemas for the various types of cybersecurity datasets, we have also developed an easy-to-use form design tool, named FAIR Data Point metadAta Schema ediTor (FAST), that provides agility and flexibility to the metadata repository platform. Last but not least, we created a metadata schema for network traffic datasets based on the lightweight Athena-o ontology, which provides a semantic basis for describing the properties of these datasets.

Keywords

FAIR principles, dataset evaluation, metadata, information security

1. Introduction

The ever-increasing number of digital threats requires a continuous advance in cybersecurity research and practices, which are becoming increasingly data-driven [1]. In this scenario, cybersecurity datasets play a key role, serving as the basis for training machine learning models, validating intrusion detection systems, analyzing malware, and investigating new vulnerabilities [2]. The availability and quality of these datasets are therefore fundamental to the reliability of predictions [3] and to driving innovation in the field of cybersecurity.

However, there are still a few quality cybersecurity datasets available to be reused [4]. The main concerns about sharing cybersecurity data are the challenges of preserving privacy and standardizing the data publication format [4]. The relative scarcity of cybersecurity datasets is compounded by the

Proceedings of the 18th Seminar on Ontology Research in Brazil (ONTOBRAS 2025) and 9th Doctoral and Masters Consortium on Ontologies (WTDO 2025), São José dos Campos (SP), Brazil, September 29 – October 02, 2025.

[*] Corresponding author.

[†] These authors contributed equally.

thaisa.silva@ime.eb.br (T. d. S. Hernandez); caroline.gandolfi29@ime.eb.br (C. D. Gandolfi); pedrohgbulcao@ime.eb.br (P. H. Bulcão); l.o.boninodasilvasantos@utwente.nl (L. B. d. S. Santos); anderson@ime.eb.br (A. F. P. d. Santos); yoko@ime.eb.br (M. C. R. Cavalcanti)

0009-0000-8310-796X (T. d. S. Hernandez); 0009-0006-1173-8600 (C. D. Gandolfi); 0009-0008-6666-0104 (P. H. Bulcão); 0000-0002-1164-1351 (L. B. d. S. Santos); 0000-0002-6754-4809 (A. F. P. d. Santos); 0000-0003-4965-9941 (M. C. R. Cavalcanti)

lack of a central registry and inconsistent provenance information. In addition, most cybersecurity datasets are outdated, and much of the information related to attack data is redundant [5]. With regard to the quality of a cybersecurity dataset, there are clear challenges in obtaining, maintaining, and publishing it. Besides, there is a shortage of consistent metrics, and researchers limit themselves to evaluating quality based on the reputation of the authors [5].

These challenges result in a central problem: the lack of a formal procedure to publish metadata and evaluate the quality and reliability of cybersecurity datasets. This work directly addresses this problem by proposing an approach to publish dataset metadata and evaluate the quality of these datasets, considering their specific properties. A differential of our approach is the integration of the FAIR principles [6] into the evaluation process. They provide guidelines for the publication of digital resources such as datasets, in a way that makes them Findable, Accessible, Interoperable, and Reusable [6]. By incorporating the FAIR principles, we not only aim to measure the technical quality of the cybersecurity datasets but also to promote better data management and reuse practices. As a secondary objective, we aim to contribute to increasing the availability of and trust in high-quality cybersecurity datasets.

To achieve these goals, a metadata repository has been implemented to support flexible schemas, adapted to the specific properties of the various types of cybersecurity datasets. Quality evaluation is carried out by the Athena Evaluator software, which analyzes the metadata published in the repository based on a set of specific quality metrics and also metrics aligned with the FAIR principles. To support the creation and management of these metadata schemas, we have also developed a lightweight ontology, which provides a semantic basis for describing the properties of cybersecurity datasets.

This article is organized as follows: Section 2 presents related work. Section 3 presents the Athena approach. Section 4 presents the implementation of this approach in the context of network traffic datasets. In Section 5, we present a case study on the evaluation of the CIC-DDoS2019 dataset. Finally, in Section 6, we conclude the paper and discuss the next steps in our research.

2. Related Works

The related work was organized to cover research that evaluates the quality of cybersecurity datasets and research that focuses on FAIR data management. Data quality assessment is a well-established field of research in several areas [7], but its specific application to cybersecurity datasets presents unique challenges related to the dynamic, heterogeneous, and sensitive nature of these data [8]. Gharib et al. [9] conducted a study of existing cybersecurity datasets between 1998 and 2016, and presented an evaluation framework for cybersecurity datasets with eleven proposed criteria: complete network configuration, complete traffic, labeled dataset, complete interaction, complete capture, available protocols, attack diversity, anonymity, heterogeneity, feature set, and metadata. These eleven criteria are evaluated according to a weight that can be defined on the basis of the organization's request or the type of Intrusion Detection System (IDS) selected for the test. In Sharafaldin et al. [10], a specific cybersecurity dataset was developed, and the quality of this dataset was compared to other synthetically generated datasets. This comparison was based on the eleven criteria proposed by Gharib et al. [9] in his framework. However, although the evaluation structure proposed by Gharib et al. is quite complete, containing a range of quality criteria and a quantitative approach to evaluating these criteria, there is a gap related to checking the timeliness of the dataset. In Ring et al.[11], a survey focused on cybersecurity datasets was carried out, where a collection of fifteen properties was established as a basis for identifying and comparing these datasets. These properties cover a range of criteria and are grouped into five categories: general information, nature of the data, volume of data, recording environment, and evaluation, but do not create a scoring structure to evaluate these criteria. Furthermore, despite agreeing with the FAIR Principles, this work does not go into these principles.

Regarding related work on FAIR data management in the field of cybersecurity datasets, Raza et al.[12] uses the FAIR Principles as a framework for data management and evaluation, reinforcing the importance of making data Findable, Accessible, Interoperable, and Reusable. The article proposes a

Table 1
Related works

	[9]	[10]	[12]	[11]	[13]	[14]	[4]	**This work**
Cybersecurity Dataset quality	x	x		x	x			x
FAIR Principles adequacy			x	x		x	x	x
Customizable cybersecurity metadata						x		x
Lightweight ontology							x	x

methodology for developing and evaluating fair-compliant datasets, although it is in a different domain of cybersecurity, focused on Large Language Models (LLMs). Silva et al. [4] proposed an approach to support cybersecurity dataset publishing for machine learning tasks following FAIR principles and involving, among others, anonymization and preprocessing of data. This approach addresses the limited availability of cybersecurity datasets, providing an environment to facilitate and motivate the creation of these datasets for publication. However, the emphasis of the approach is on generating higher-quality data in line with the FAIR principles, rather than covering a process of evaluating the dataset prior to its publication. The research carried out by Göbel et al. [13] has a focus on the creation and optimization of datasets in the context of cybersecurity, with an emphasis on digital forensics. It addresses the challenges and best practices for creating high-quality datasets, although it does not explicitly address the fairness of datasets. Mombelli et al. [14] addresses the application of the FAIR Principles and metadata quality in the field of digital forensics. The paper evaluates metadata completeness and compliance with the FAIR Principles in 212 datasets from NIST's Computer Forensic Reference Dataset Portal (CFReDS). The results indicate deficiencies in metadata quality and the need for better data management standards. Providing important insights into the ongoing need to improve metadata management in cybersecurity datasets.

Unlike the aforementioned works, this paper combines data management and quality assessment in the field of cybersecurity. The Athena approach was based on three fundamental pillars: a customizable FAIR Data Point repository, a lightweight support ontology called Athena-o, and the Athena Evaluator software for evaluating specific cybersecurity metrics and FAIR metrics. By incorporating FAIR principles as a new dimension of quality assessment, we aim not only to measure the technical quality of cybersecurity datasets but also to promote best practices in data management and reuse. Table 1 summarizes the differential of the Athena approach.

3. Athena Approach

The Athena approach aims to evaluate the quality of cybersecurity datasets and to help promote better data management and sharing practices. To this end, we integrate the analysis of the intrinsic properties of cybersecurity datasets with the evaluation of their compliance with the FAIR principles. Our approach is extensible, capable of adapting to the diversity of datasets in the cybersecurity domain, such as network traffic and malware datasets. The Athena approach is based on three fundamental pillars: a customized metadata repository, a lightweight support ontology called Athena-o and the Athena Evaluator software. Figure 1 gives an overview of the Athena approach, and each stage is detailed below.

Quality evaluation begins with the publication of the dataset through a set of descriptive metadata, stored in our customized repository. This repository has been implemented to be flexible, allowing the definition of specific metadata schemes for different types of cybersecurity datasets. The central idea is that quality is not an absolute concept, "quality data must be intrinsically good, contextually appropriate for the task and clearly represented to the data consumer" [15]. In the step **Select a Cybersecurity Dataset Type**, the person responsible for publishing the cybersecurity dataset metadata, in this approach called the Publisher, selects the most appropriate metadata schema to describe their type of cybersecurity dataset.

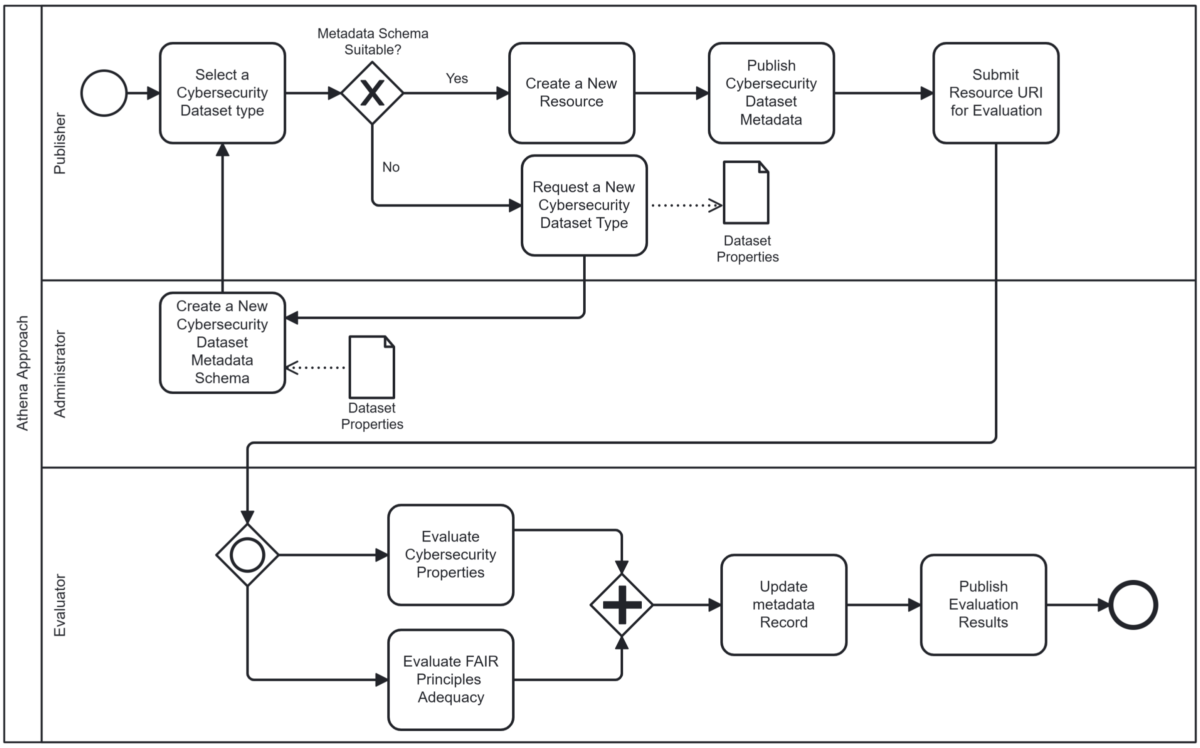

Athena Approach
Publisher
Administrator
Evaluator
Select a Cybersecurity Dataset type
Metadata Schema Suitable?
Yes
No
Create a New Resource
Publish Cybersecurity Dataset Metadata
Submit Resource URI for Evaluation
Request a New Cybersecurity Dataset Type
Dataset Properties
Create a New Cybersecurity Dataset Metadata Schema
Dataset Properties
Evaluate Cybersecurity Properties
Evaluate FAIR Principles Adequacy
Update metadata Record
Publish Evaluation Results

Figure 1: Athena approach - publication and evaluation process

The Administrator plays the role of the person responsible for creating metadata schemas. If there is no suitable metadata schema, the step **Request a New Metadata Schema** can be triggered and the Administrator can **Create a New Cybersecurity Dataset Metadata Schema** informing the necessary properties. The definition of these schemas is supported by our lightweight ontology, which provides semantic relationships to describe the properties consistently. Moreover, this task should provide a form editor facility, which allows users to create these schemas, facilitating the extensibility of the approach. Details on the implementation of the metadata repository, the lightweight ontology, and the form editor will be presented in section 4.

In our approach, a dataset is registered in the repository as a resource. So, in the next step **Create a new Resource**, a new digital resource is created containing metadata records from a specific dataset according to the selected metadata schema. Once the resource has been created, in the step **Publish Cybersecurity Dataset Metadata** the Cybersecurity Dataset metadata is recorded and made available for evaluation.

When the Publisher **Submit Resource URI for Evaluation**, the Evaluator software starts interacting with the metadata repository with the URI of a resource provided. At this stage, the dataset will undergo two types of evaluation. In the step **Evaluate Cybersecurity Properties**, the dataset will be evaluated based on a set of metrics defined on the basis of specific cybersecurity properties. These properties, from a general perspective, can cover aspects such as data timeliness and relevance. The selection and weighting of the metrics can be adjusted depending on the type of dataset, giving flexibility to the process. In addition, in the stage **Evaluate FAIR principles adequacy** the dataset is subjected to maturity tests using FAIR Metrics[1] to evaluate its level of compliance with the FAIR principles. The results of the evaluations are published together with the other metadata records.

[1] https://github.com/FAIRMetrics/Metrics/

4. Implementation

To implement the Athena approach, a FAIR Data Point repository[2] was customized to support a specific metadata schema, adapted to the specific properties of the various types of cybersecurity datasets. The FAIR Data Point follows the Data Catalog Vocabulary (DCAT)[3], and one of its main differential characteristics is its flexibility, i.e., it may be customized to describe different types of digital objects, which are defined as sub-classes of DCAT Resource. This work takes advantage of this feature, using the inheritance of DCAT's general properties and focusing only on specific features of cybersecurity datasets.

Although the Athena approach aims to cover a variety of cybersecurity datasets, its initial implementation and validation focused on network traffic datasets. In the context of cybersecurity, network traffic datasets have specific characteristics that need to be considered, such as the year in which the traffic was generated, which is different from the year in which the dataset itself was published; the incidence of malicious traffic and the corresponding types of attack; whether the traffic was labeled or not; and the type of network on which the traffic was generated. Ring et al. [11] summarized this set of properties into five categories: general information (year of traffic creation, public availability, normal traffic, attack traffic), nature of the data (metadata, format, anonymity), data volume (count and duration), recording environment (traffic type, network type, complete network) and evaluation (predefined, balanced, labeled divisions).

In this section, we first describe the lightweight ontology named Athena-o (subsection 4.1), which reused concepts of existing ontologies, conforming to the Interoperability principle. From this ontology, we derived the Athena metadata schema (subsection 4.2), which included the properties already mentioned by Ring et al. [11], extending the DCAT schema. Finally, the Athena Evaluator (subsection 4.3) was implemented using the FAIR metrics API[1], which already implements metrics to evaluate datasets concerning the FAIR principles. However, we implemented new specific metrics to evaluate the network traffic datasets, based on the specific properties defined in the Athena metadata schema.

4.1. Athena-o

The Athena-o lightweight ontology, shown in Figure 2, was developed to provide a semantic basis and interoperability in the selected properties that describe cybersecurity datasets. This ontology defines new concepts, as well as reuses existing classes from well-known ontologies and vocabularies, such as Dublin Core (DC)[4], TOUCAN Ontology (ToCo)[5], Unified Cyber Ontology (UCO)[6] and National Institute of Standards and Technology (NIST) glossary[7].

By extending the **DCAT Dataset** concept, Athena-o reuses the already well-established properties relating to dataset metadata such as *dcterms:format* and *dcat:byteSize*. In this article, we focus on the specific features of cybersecurity datasets. Athena-o introduces the **Cybersecurity Dataset** concept (*at:CybersecurityDataset*), which specializes the **DCAT dataset** concept (*dcat:Dataset*), of which, in turn, the **Network Traffic Dataset** (*at:NetworkTrafficDataset*) concept is specialized. Furthermore, specific properties have been defined for the **Network Traffic Dataset** concept conforming to the properties defined by Ring et al. [11], such as the year of traffic creation (*time:yearOfTrafficCreation*), and the kind of traffic (*at:kindOfTraffic*), whose values may be *real*, *emulated*, or *synthetic*. The **Traffic** concept inherits from the **UCO Network Flow** concept (*uco:NetworkFlow*), which can be specialized into two concepts: **Normal network traffic** (*at:NormalTraffic*) and **Anomalous network traffic** (*at:AttackTraffic*). The **Attack Traffic** concept is connected to the **Attack Type** concept (*nist:attack*), reused from the NIST Glossary, through the property *at:isClassifiedBy*. This concept has two properties that represent the attackers' IP (*at:AttackerIP*) and the victims' IP (*at:VictimIP*).

[2] https://app.fairdatapoint.org/

[3] https://www.w3.org/TR/vocab-dcat-3/

[4] https://www.dublincore.org/

[5] https://github.com/QianruZhou333/toco_ontology/

[6] https://unifiedcyberontology.org/

[7] https://csrc.nist.gov/glossary/

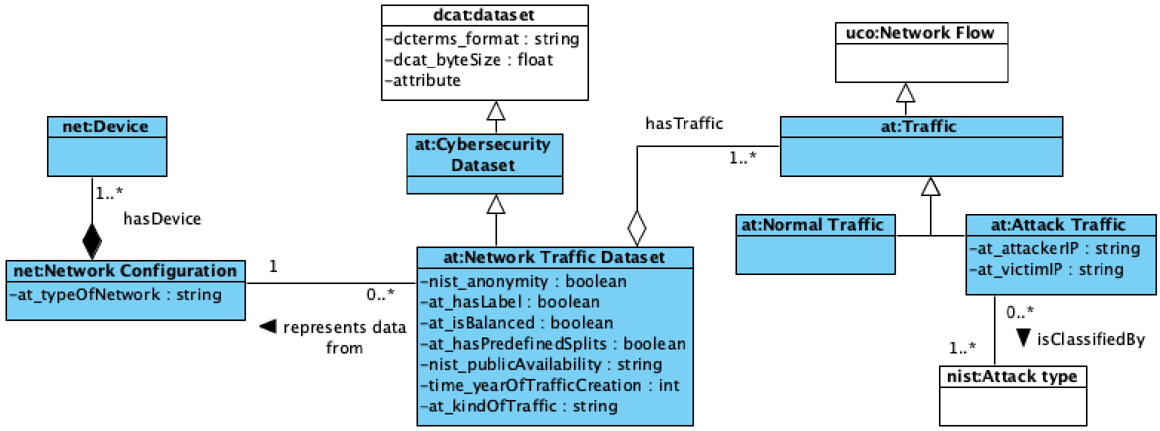


Figure 2: Athena-o: a lightweight ontology for the Athena approach

According to Ring et al., a dataset description must include the network configuration through which the traffic flowed. Thus, Athena-o reuses the **Physical infrastructure** (*net:PhysicalInfrastructure*) and **Device** (*net:Device*) concepts from the Toco Ontology. The former includes a property that represents the type of network from which the data in a dataset was collected (*at:typeOfNetwork*), and the latter represents the devices that are part of a network infrastructure. In addition, *nist:hasPublicAvailability* and *nist: Anonymity* properties represent the availability and anonymization of a dataset, respectively. Finally, *at:hasPredefinedSplits*, *at:hasLabel*, and *at:isBalanced* properties represent metadata that are useful for performing effective machine learning tasks. These properties provide respectively information if a dataset includes predefined subsets for training and evaluation, if datasets are labeled or not and if datasets are balanced with respect to their class labels.

The applicability of the approach to other types of cybersecurity datasets (e.g., malware) is possible through the extension of the Athena-o ontology, in which case a new class would be added as a subclass of Cybersecurity Datasets. This extension of the ontology and the subsequent creation of a metadata schema are performed by the Administrator, based on the new properties submitted by the publisher, as described in Section 3.

4.2. Athena Metadata Schema

The Athena metadata schema is expressed in RDF (Resource Description Framework) using the Shapes Constraint Language (SHACL) [16] and the Data Shapes Vocabulary (DASH)[8]. The former is rich in establishing constraints for validating the schema instantiations, while the latter is an extension of SHACL with new constraints and target types, and also includes components to fix constraint violations. Moreover, SHACL includes constructs such as *sh:order* and *sh:group* that can aid in the construction of form layouts, and DASH also includes constructs that are particularly useful for form configuration, such as *dash:TextFieldEditor*.

Athena-o guided the creation of the corresponding metadata schema, but some simplifications were made. The choice for the enrichment of the schema with SHACL and DASH languages was required by the FAIR Data Point implementation. Listing 1 shows a fragment of the Athena metadata schema created for describing the Network Traffic datasets. Note that it begins with the declaration of the *DatasetShape* element that has the *dcat:Dataset* as its target class. For simplification reasons, besides the dataset element, all the other elements of Athena-o were mapped into properties associated to the dataset element. The DASH constructs inform the FAIR Data Point user interface elements, so it can organize and configure the properties in a form for capturing metadata values.

For example, the *publicAvailability* attribute is defined as a drop down menu (*Select field*), which guides the user in choosing one of the pre-defined options. According to Athena-o, a Network Traffic

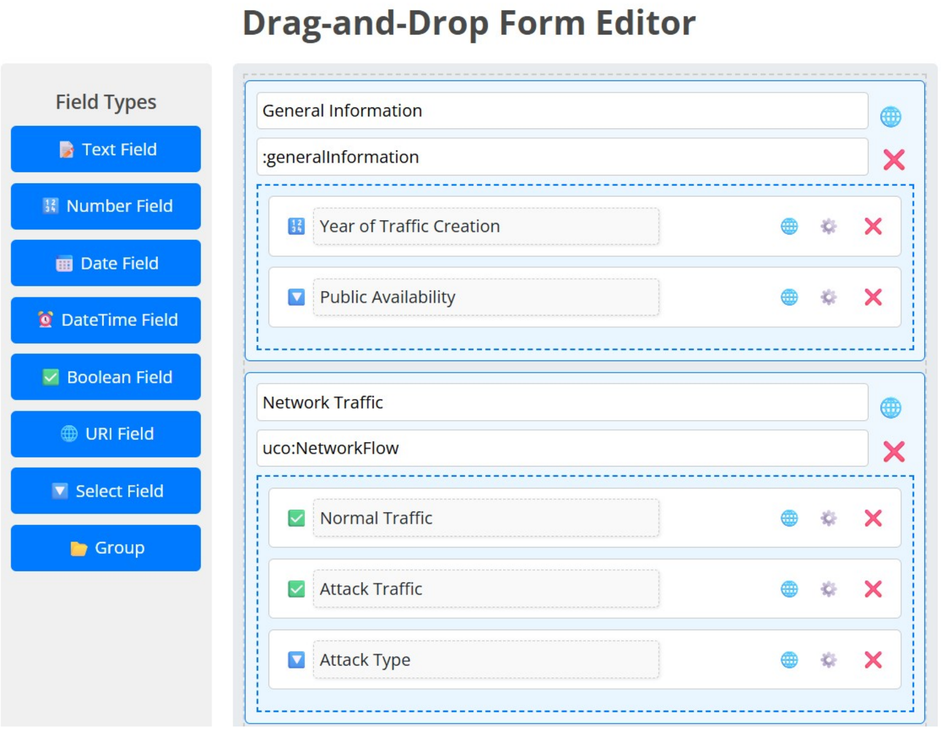
[8] https://datashapes.org/dash/

Dataset contains Traffic that may be Normal or Attack Traffic. Note that, to describe the dataset metadata, there is no need to represent its content, but it is important to indicate if it includes Normal or Attack traffic. Thus, *Attack/Normal traffic* properties are defined as boolean datatypes. Similarly, the *Attack type* is also defined as a property associated directly with the dataset, indicating what types of attack it includes.

Listing 1: Metadata Schema Fragment for Network Traffic Datasets

```
:DatasetShape a sh:NodeShape ;
    sh:targetClass dcat:Dataset ;
    sh:property [
        sh:path time:year ;
        sh:name "Year of Traffic Creation" ;
        sh:datatype xsd:integer ;
        dash:editor dash:TextFieldEditor ;
        dash:viewer dash:LiteralViewer ;
        sh:minCount 0 ;
        sh:maxCount 1 ;
        sh:group :generalInformation ;
        sh:order 0 ;
    ] ;
    sh:property [
        sh:path at:publicAvailability ;
        sh:name "Public Availability" ;
        sh:datatype xsd:string ;
        sh:in ( "No" "On request (o.r.)" "Yes" ) ;
        dash:editor dash:EnumSelectEditor ;
        dash:viewer dash:LiteralViewer ;
        sh:minCount 0 ;
        sh:maxCount 1 ;
        sh:group :generalInformation ;
        sh:order 1 ;
    ] ;
    sh:property [
        sh:path at:NormalTraffic ;
        sh:name "Normal Traffic" ;
        sh:datatype xsd:boolean ;
        dash:editor dash:BooleanSelectEditor ;
        dash:viewer dash:LiteralViewer ;
        sh:minCount 0 ;
        sh:maxCount 1 ;
        sh:group uco:NetworkFlow ;
        sh:order 2 ;
    ] ;
    sh:property [
        sh:path at:AttackTraffic ;
        sh:name "Attack Traffic" ;
        sh:datatype xsd:boolean ;
        dash:editor dash:BooleanSelectEditor ;
        dash:viewer dash:LiteralViewer ;
        sh:minCount 0 ;
        sh:maxCount 1 ;
        sh:group uco:NetworkFlow ;
        sh:order 3 ;
    ] ;
    sh:property [
        sh:path nist:attack ;
        sh:name "Attack Type" ;
        sh:datatype xsd:string ;
        dash:editor dash:InstancesSelectEditor ;
        dash:viewer dash:LiteralViewer ;
        sh:minCount 0 ;
        sh:maxCount 100 ;
        sh:group uco:NetworkFlow ;
        sh:order 4 ;
    ] ;
```

Finally, we highlight that the created schema is easily extended or adapted to other types of cybersecurity datasets. A user-friendly form design tool, named FAIR Data Point metadAta Schema ediTor (FAST), was implemented to automate the schema generation. It allows schema designers to configure a user interface form by dragging and dropping interface components into a canvas in a visual way. Then, the form is automatically transformed into a SHACL/DASH specification of the schema, which in turn is the input to the FAIR Data Point schema configuration. Figure 3 shows an example of the FAST interface tool, where, for example, the *Attack Traffic* property is defined with a Boolean field type. While adding all properties and their respective field types, the SHACL/DASH code can be viewed and edited.



General Information
:generalInformation
Year of Traffic Creation
Public Availability
Network Traffic
uco:NetworkFlow
Normal Traffic
Attack Traffic
Attack Type

Figure 3: Form editor for Network Traffic datasets metadata schema.

4.3. Athena Evaluator

Athena Evaluator is an application developed in Python whose function is to evaluate cybersecurity datasets in two aspects: the first aspect relates to the intrinsic properties of these specific types of datasets, described through a metadata schema, and the second relates to the compliance of these datasets with the FAIR Principles. To do this, Athena Evaluator interacts with the FAIR Data Point API[9] and executes the evaluation metrics based on the published metadata of a given dataset.

In our implementation with focus on Network Traffic datasets, Athena Evaluator applies the quality assessment metrics based on the values of the following metadata: Year of Traffic Creation, Public availability, Normal Traffic, Attack Traffic, Anonymity, Complete Network, Predefined Splits, Balanced, Labeled according to the metadata schema.

Since new attack scenarios emerge every day, the age of a cybersecurity dataset plays a very important role [11]. Older datasets may not fully reflect the risks that exist today, since attacks have new variants launched all the time. To evaluate timeliness, the logic *Fuzzy* [17] is used to define the degree of pertinence of the year of creation of the data set in “Old”, “Medium” and “Recent” categories. For the purpose of this research, we established a specific range of time covers from 1998 to nowadays (2025) because the relevance of the datasets generated in this period with the following intervals: Old [1998 to 2007], Medium [2003 to 2019] and Recent [2016 to 2025]. The pertinence of the year the dataset traffic was created to one of the sets is calculated using a triangular pertinence function [17]. Regarding the anonymization metric, the problems of compromised privacy occur when the payload is not encrypted in a dataset with real traffic. So, most datasets have their payloads removed or anonymized, which decreases the usefulness of the dataset but maintains the privacy of the information [9]. Datasets with synthetic or emulated traffic do not suffer from this issue and can keep this information available. Therefore, the evaluation of this metric is directly related to the type of traffic in the dataset, which can have three values: Real, Emulated, and Synthetic. If a dataset has a real traffic type, it means that the data needs to be anonymized; otherwise, if the traffic type is emulated or synthetic, it makes no sense for the data to be anonymized.

Moreover, the dataset is evaluated based on the presence of the following properties: Public availability, Normal and Attack Traffic, Complete Network, Predefined Splits, Balanced, Labeled according to the metadata schema. Finally, the relevance of a dataset is evaluated by a metric that weights the number of its citations (obtained through its DOI). In this metric, the number of citations is attenuated and correlated with the score assigned to the year in which the traffic was created. This approach helps that a dataset’s historical popularity does not overshadow the usage-based relevance of more recent datasets. Concerning the FAIR principles, Athena Evaluator implements a selected set of FAIR metrics[1]. The score for each sub-principle is the average of the corresponding FAIR metrics. The compliance of the evaluated dataset with each of the principles is as follows:

Findable: Metrics are used to verify the existence of globally unique and persistent identifiers associated with the dataset in order for them to be found and resolved by computers. Globally unique means that the identifier is guaranteed to refer unambiguously to exactly one resource in the world, and persistence refers to the requirement that this globally unique identifier is never reused in another context and continues to identify the same resource, even if that resource no longer exists (F1) [18]. In addition, metrics are used to verify the richness of the metadata description (F2). According to Jacobsen et al. [18], it is hard to generally define the minimally required “richness” of this metadata, except that the more generous it is, both for humans and computers, the more specifically findable it becomes in refined searches. Furthermore, the principles (F3) metadata clearly and explicitly include the identifier of the data it describes, and (F4) metadata are registered or indexed in a searchable resource are also evaluated.

Accessible: One of the main objectives of identifying a digital resource is to simultaneously provide the ability to retrieve the record of that digital resource, in a given format, using a clearly defined mechanism: thus, retrievability is a facet of FAIR accessibility [18]. In this case, a set of metrics is used to check the level of recoverability of the data, including authentication/authorization protocols if necessary (A1.1 and A1.). In addition, the FM-A2 metric is used to verify that metadata is accessible, even when the data are no longer available (A2). It is important that consumers have, at the very least, access to high-quality metadata that describes those resources sufficiently to minimally understand

[9] https://app.fairdatapoint.org/swagger-ui/index.html/

their nature and their provenance, even when the relevant data are not available anymore. There is a continued focus on keeping relevant digital resources available in the future [18].

Interoperability: Achieving a “common understanding” of digital resources through a globally understood “language” for machines is the purpose of principle I1. To evaluate this principle, we used the FAIR Metrics to verify the use of a knowledge representation language, vocabularies and ontologies (I1 and I2). In addition, references to other related resources are included in order to verify that the knowledge representing one resource is linked to that of other resources to create a significantly interconnected network of data and services (I3) [18].

Reusable: Digital resources and their metadata must always, without exception, include a license that describes under what conditions the resource can be used, even if it is “unconditional”. Here, metrics are used to verify the presence of a clear and accessible license (R1.1) and a detailed description of the provenance of the dataset (R1.2).

5. Case Study

For the case study, we selected CIC-DDoS2019[10] because it is widely recognized for intrusion detection research, especially for Distributed Denial of Service (DDoS) attacks, contains a wide variety of DDoS attacks in real time and is used by researchers to find the best characteristics and the best model to detect this type of attack with minimal execution time and cost [19]. For this dataset, we collected the metadata needed to populate our FAIR Data Point repository, using the support of the schema defined for network traffic datasets (Section 4.2). We then submitted the dataset for evaluation by the Athena Evaluator software. Figure 4 shows the metadata collected and published according to the created metadata schema and Figure 5 shows the results of the evaluations carried out by Athena Evaluator.

In Figure 4, we point out that the Network Traffic Datasets metadata schema is informed using the *conformsTo* predicate of the Dublin Core Terms [20]. In addition, metadata from the DCAT Resources and Datasets classes, such as *dcterms:license* and *dcterms:rights* are inherited to compose, together with the Network Traffic Datasets schema, the metadata records of the CIC-DDoS2019 dataset.

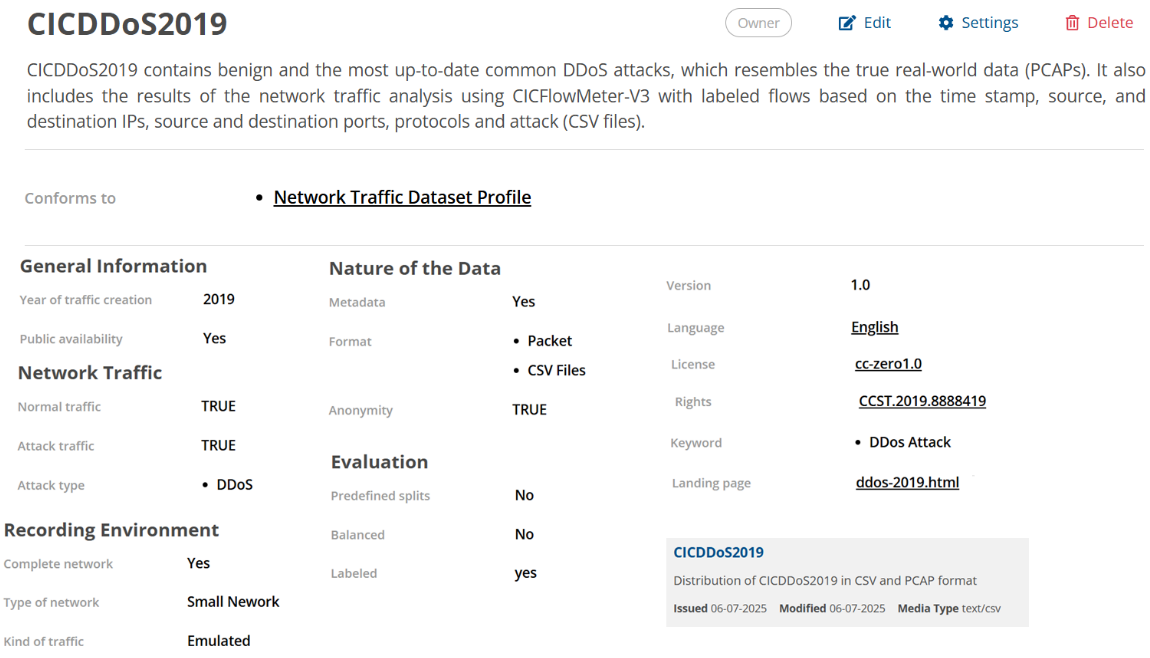
In the first part of the evaluation, metadata for the year of traffic creation, public availability, normal traffic, attack traffic, metadata, anonymity, complete network, predefined splits, labeled and balanced are collected by Athena Evaluator through the FAIR Data Point API[9] and submitted for evaluation according to the specific metrics detailed in Section 4.3. The degree of pertinence of the year of traffic creation of the CIC-DDoS2019 dataset in “Old”, “Medium” and “Recent” categories was calculated using a triangular pertinence function, receiving a higher score for having a higher degree of membership to the "Recent" set. In addition, the dataset is publicly available, contains benign traffic as well as more up-to-date DDoS attacks (DNS, SNMP, NTP, WebDDoS, MSSQL, UDP, LDAP, NetBIOS, SSDP, PortScan, UDP-Lag, and SYN), has a complete network configuration, and makes a good amount of metadata available to the community. Regarding the anonymity metric, since it is a dataset that contains an emulated traffic type, anonymization is not necessary. Furthermore, since it is a labeled dataset, it received the maximum score in this metric. On the other hand, because it is not balanced and does not contain predefined subsets, it did not score in these categories. Finally, its relevance was calculated considering the number of citations and the age score of the dataset.

In the second part of the evaluation, Athena Evaluator assessed the conformity of the CIC-DDoS2019 dataset to the FAIR principles, focusing on its metadata published in the FAIR Data Point repository. The CIC-DDoS2019 dataset performed excellently in the evaluations regarding the principles F1, F2, F3, and F4 due to its rich metadata description and a globally unique and persistent identifier through its DOI. Regarding the Accessibility principle, using HTTP as a communication protocol and publishing its metadata in the FAIR Data Point, which allows for an authentication and authorization procedure when necessary, enabled a good score in these principles (A1.1 and A1.2). Furthermore, the metadata records are available in RDF format, contributing to the principle (I1), and to the use of "vocabularies" such as Dublin Core[4], ToCo[5], UCO[6] and NIST glossary[7] (I2 and I3). For this last test, two metrics were

[10] https://www.unb.ca/cic/datasets/ddos-2019.html/

used, in which any Linked Data found was tested for the resolution of a subset of properties (predicates) present and whether these are handled for other Linked Data, failing only the latter. Finally, the dataset was evaluated concerning a clear and accessible data usage license through the *dcterms:license* and *dcterms:rights* (R1.1) and if associated with detailed provenance (R1.2) metadata through, for example, the *dcterms:publisher*, *dcat:contactPoint*, *dcat:landingPage* metadata present in the FAIR Data Point.

The aim is not to score on all the principles, but to encourage the community to provide more accessible, interoperable, and reusable datasets for the advancement of cybersecurity research. By evaluating cybersecurity datasets from this perspective, our study not only contributes to understanding the quality of this specific dataset but also exemplifies the practical application of the FAIR principles for promoting open science and data reuse in a cybersecurity context, encouraging their adoption in future datasets. The Athena Evaluator code is available at Github[11] for evaluation by the community and reproduction of the results presented here.



CICDDoS2019

CICDDoS2019 contains benign and the most up-to-date common DDoS attacks, which resembles the true real-world data (PCAPs). It also includes the results of the network traffic analysis using CICFlowMeter-V3 with labeled flows based on the time stamp, source, and destination IPs, source and destination ports, protocols and attack (CSV files).

Conforms to
- Network Traffic Dataset Profile

General Information

Year of traffic creation	2019
Public availability	Yes

Network Traffic

Normal traffic	TRUE
Attack traffic	TRUE
Attack type	• DDoS

Recording Environment

Complete network	Yes
Type of network	Small Nework
Kind of traffic	Emulated

Nature of the Data

Metadata	Yes
Format	• Packet • CSV Files
Anonymity	TRUE

Evaluation

Predefined splits	No
Balanced	No
Labeled	yes

Version	1.0
Language	English
License	cc-zero1.0
Rights	CCST.2019.8888419
Keyword	• DDos Attack
Landing page	ddos-2019.html

CICDDoS2019

Distribution of CICDDoS2019 in CSV and PCAP format

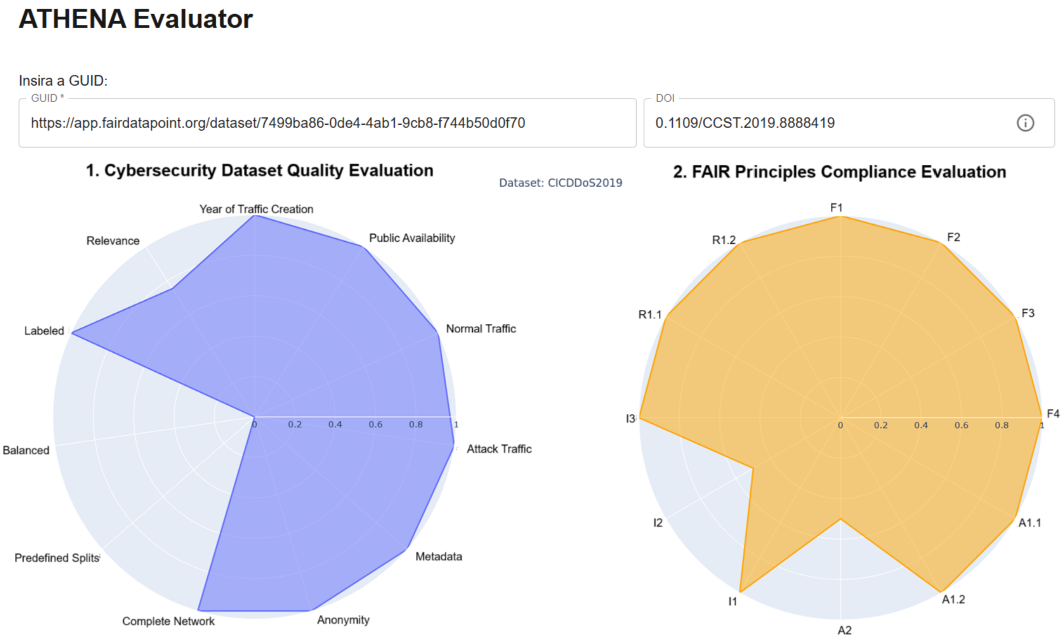
Issued 06-07-2025 **Modified** 06-07-2025 **Media Type** text/csv

Figure 4: FAIR Data Point repository with a specific metadata schema for cybersecurity datasets.

6. Conclusion

This article presented an approach to publish dataset metadata and evaluate the quality of these datasets, considering their specific properties and integration of the FAIR principles into the evaluation process. To this end, the Athena approach was based on three fundamental pillars: a customized FAIR Data Point repository, a lightweight support ontology called Athena-o and the Athena Evaluator software. The FAIR Data Point has been implemented to support flexible metadata schemas, adapted to the specific properties of the various types of cybersecurity datasets. Quality evaluation was carried out by the Athena Evaluator software, which analyzes the metadata published in the repository based on a set of specific quality metrics and also metrics aligned with the FAIR principles. To support the creation and management of these metadata schemas, we have also developed a lightweight ontology, which provides a semantic basis for describing the properties of cybersecurity datasets. We present a case study evaluating the CIC-DDoS2019 dataset, demonstrating the viability of integrating specific properties with FAIR principles, thus structuring a systematic approach with a formal procedure for

[11] https://github.com/comp-ime-eb-br/S2C2-IME/tree/main/deliverables/AthenaEvaluator/



ATHENA Evaluator

Insira a GUID:

GUID *

https://app.fairdatapoint.org/dataset/7499ba86-0de4-4ab1-9cb8-f744b50d0f70

DOI

0.1109/CCST.2019.8888419

Dataset: CICDDoS2019

Figure 5: Athena Evaluator with the result of the evaluation of the CICIDS2019 dataset in relation to its specific properties and its compliance with the FAIR principles.

evaluating cybersecurity datasets. The FAIR Data Point implementation is flexible and can be extended to accommodate new properties and other types of cybersecurity datasets.

By assessing specific properties of cybersecurity datasets as well as potential areas for improvement from a metadata perspective, we provide guidance for researchers involved in creating new datasets. Furthermore, by assessing cybersecurity datasets from this perspective, our study not only contributes to the understanding of the quality of this dataset but also exemplifies the practical application of the FAIR principles to promote open science and data reuse in a cybersecurity context, encouraging their adoption in future datasets. The goal is not to score on all principles, but to encourage the community to provide more findable, accessible, interoperable, reusable, and higher-quality datasets to advance cybersecurity research.

This study focused on the evaluation of dataset quality based on its metadata and the FAIR principles, without delving into the performance of models. For future work, we suggest carrying out a comparative analysis of the impact of the quality characteristics of the evaluated datasets on the performance of different intrusion detection algorithms. In addition, applying this evaluation methodology to other network security datasets could further enrich the understanding of data quality in the area. Finally, we intend to conduct empirical studies that provide further evidence of the applicability of our approach across diverse scenarios and perform a comparative evaluation of other automated frameworks, thereby allowing for a more comprehensive understanding of their performance and the potential advantages of our approach. In this paper, we briefly describe the metrics used to evaluate the datasets. A detailed description will be provided in a future publication.

7. Acknowledgments

This research has been partially supported by CAPES ("Coordenação de Aperfeiçoamento de Pessoal de Nível Superior", in Portuguese). Also, it has been funded by FINEP/DCT/FAPEB (ref.: 2904/20 contract no 01.20.0272.00) under the “System of Systems of Command and Control” project (“Sistema de Sistemas de Comando e Controle”, in Portuguese). Generative AI was used to improve English writing in a few specific points of the text.

References

[1] M. Zheng, H. Robbins, Z. Chai, P. Thapa, T. Moore, Cybersecurity research datasets: Taxonomy and empirical analysis, in: 11th USENIX Workshop on Cyber Security Experimentation and Test (CSET 18), USENIX Association, Baltimore, MD, 2018.

[2] A. Khraisat, I. Gondal, P. Vamplew, J. Kamruzzaman, Survey of intrusion detection systems: Techniques, datasets and challenges, Cybersecurity 2 (2019) 1–22.

[3] M. Macas, C. Wu, W. Fuertes, A survey on deep learning for cybersecurity: Progress, challenges, and opportunities, Computer Networks 212 (2022) 109032.

[4] M. L. e. Silva, K. de Faria Cordeiro, M. C. Cavalcanti, Sec4ml: An approach to support cybersecurity data publishing for machine learning tasks, in: 2021 IEEE 25th International Enterprise Distributed Object Computing Workshop (EDOCW), 2021, pp. 226–235. doi:10.1109/EDOCW52865.2021.00053.

[5] A. Kenyon, L. Deka, D. Elizondo, Are public intrusion datasets fit for purpose? characterising the state of the art in intrusion event datasets, Computers & Security 99 (2020) 102022.

[6] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al., The fair guiding principles for scientific data management and stewardship, Scientific Data 3 (2016) 1–9.

[7] O. Reda, N. C. Benabdellah, A. Zellou, A systematic literature review on data quality assessment, Bulletin of Electrical Engineering and Informatics 12 (2023) 3736–3757.

[8] J. Zhao, M. Shao, H. Wang, X. Yu, B. Li, X. Liu, Cyber threat prediction using dynamic heterogeneous graph learning, Knowledge-Based Systems 240 (2022) 108086.

[9] A. Gharib, I. Sharafaldin, A. H. Lashkari, A. A. Ghorbani, An evaluation framework for intrusion detection dataset, in: 2016 International Conference on Information Science and Security (ICISS), IEEE, 2016, pp. 1–6.

[10] I. Sharafaldin, A. H. Lashkari, A. A. Ghorbani, et al., Toward generating a new intrusion detection dataset and intrusion traffic characterization., ICISSp 1 (2018) 108–116.

[11] M. Ring, S. Wunderlich, D. Scheuring, D. Landes, A. Hotho, A survey of network-based intrusion detection data sets, Computers & Security 86 (2019) 147–167.

[12] S. Raza, S. Ghuge, C. Ding, E. Dolatabadi, D. Pandya, Fair enough: Develop and assess a fair-compliant dataset for large language model training?, Data Intelligence 6 (2024) 559–585. doi:10.1162/dint_a_00255.

[13] T. Göbel, F. Breitinger, H. Baier, Optimising data set creation in the cybersecurity landscape with a special focus on digital forensics: Principles, characteristics, and use cases, Forensic Science International: Digital Investigation 52 (2025) 301882. doi:https://doi.org/10.1016/j.fsidi.2025.301882.

[14] S. Mombelli, J. R. Lyle, F. Breitinger, Fairness in digital forensics datasets' metadata–and how to improve it, Forensic Science International: Digital Investigation 48 (2024) 301681.

[15] R. Y. Wang, D. M. Strong, Beyond accuracy: What data quality means to data consumers, Journal of management information systems 12 (1996) 5–33. 23 out. de 2023.

[16] W. W. W. C. (W3C), SHACL - shapes constraint language, https://www.w3.org/TR/shacl/, 2017. W3C Recommendation, 20 July 2017. Accessed June 2025.

[17] G. Klir, B. Yuan, Fuzzy sets and fuzzy logic, volume 4, Prentice hall New Jersey, 1995.

[18] A. Jacobsen, R. de Miranda Azevedo, N. Juty, D. Batista, S. Coles, R. Cornet, M. Courtot, M. Crosas, M. Dumontier, C. T. Evelo, et al., Fair principles: interpretations and implementation considerations, 2020.

[19] M. Ramzan, M. Shoaib, A. Altaf, S. Arshad, F. Iqbal, Á. K. Castilla, I. Ashraf, Distributed denial of service attack detection in network traffic using deep learning algorithm, Sensors 23 (2023) 8642.

[20] L. O. B. da Silva Santos, K. Burger, R. Kaliyaperumal, M. D. Wilkinson, Fair data point: A fair-oriented approach for metadata publication, Data Intelligence 5 (2023) 163–183.