

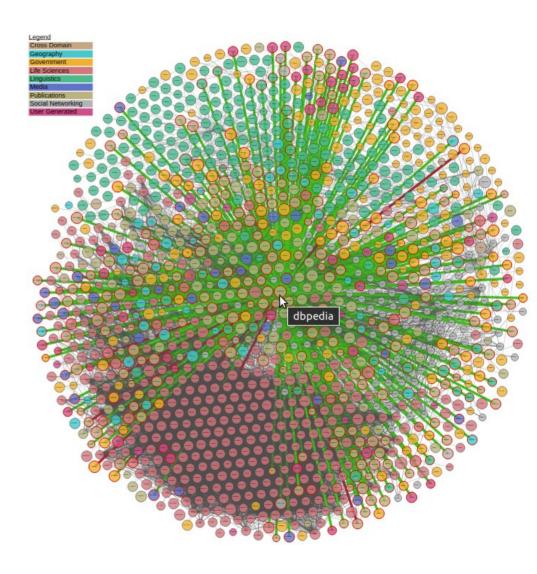
#### INSTITUTO MILITAR DE FNGENHARIA

Programa de Pós-Graduação em Sistemas e Computação

Disciplina: Web Semântica Período: 2023.2

Aluna: Thaisa da Silva Pinto

DESENVOLVIMENTO DE UM
MODELO DE DADOS EM RDF
PARA UM DATASET DE DETECÇÃO
DE INTRUSÃO: IMPLEMENTAÇÃO
E PUBLICAÇÃO EM UM FAIR
DATAPOINT



#### **OBJETIVO**

O objetivo deste trabalho é desenvolver um **modelo de dados** de um *dataset* de Sistemas de Detecção de Intrusão (IDS) utilizando o *framework* RDF, implementar esse modelo utilizando o GraphDB e, por fim, publicar seus metadados em um FAIR Data Point.

# SUMÁRIO

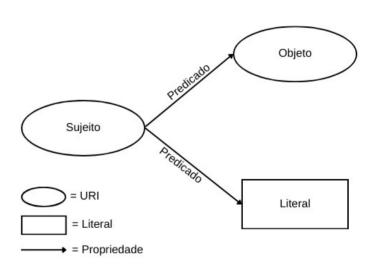
- Conceituação
- Apresentação do dataset CSE-CIC-IDS2018
- Visão geral dos artefatos produzidos
- Visão geral do modelo de dados
- Descrição do modelo de dados
- Representação do modelo RDF turtle
- Implementação do modelo de dados no GraphDB
- Exemplos de consultas no GraphDB
- Publicação no FAIR Data Point

# **CONCEITUAÇÃO**

## RESOURCE DESCRIPTION FRAMEWORK (RDF)

É um **arcabouço** para representar informações na **web** que pode ser processado por **máquinas**.

- Permite fazer afirmações sobre **recursos** (quaisquer coisas, tanto concretas quanto abstratas).
- O RDF possui três elementos (triplas):
   <sujeito> <pr
- Uma **tripla** pode ser representada como um grafo dirigido, do sujeito para o objeto (flexibilidade representação).
- RDF Schema (RDFS) É uma extensão do RDF permite descrever classes, propriedades e definição de domínios. Permite criar restrições sobre as triplas, além de inferências.



Permite adicionar regras e ontologias (representação semântica).

Fonte: https://www.w3.org/TR/rdf11-primer/

# CONCEITUAÇÃO

## **ONTOLOGY WEB LANGUAGE (OWL)**

OWL é uma linguagem que estende o RDF e oferece um conjunto amplo (ex: disjointWith, sameAs, equivalentClass) de tipos de restrições ao conjunto de triplas definidas.

## Por que definir restrições?

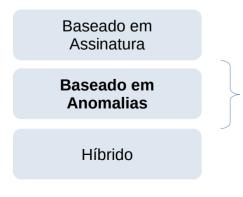
"uma das formas de se estender o que é semântica é pensar que o que faz com que diferentes pessoas possam entender o mesmo significado de algum conteúdo é **restringir o número de diferentes interpretações** possíveis sobre aquele conteúdo." (LAUFER, 2015 )

Fonte: https://www.w3.org/TR/owl-ref/

# **CONCEITUAÇÃO**

SISTEMAS DE DETECÇÃO DE INTRUSÃO (IDS)

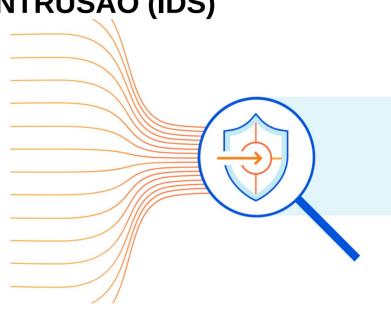
- Tem a função de monitorar e analisar os eventos que ocorrem em uma rede em busca de sinais de possíveis Incidentes de Segurança de Informação (ISI) (Mittal, 2016).
- ISIs podem ter muitas causas como acesso, utilização, divulgação, perturbação, modificação ou destruição não autorizada que afetem a confidencialidade, integridade ou a disponibilidade de um sistema computacional (NIST, 2023).
- Métodos de detecção:



#### Métodos de treinamento:

- Estatístico
- Baseado em conhecimento
- Aprendizado de Máquina

Usam conjuntos de dados



### **DATASET CSE-CIC-IDS2018**

- É um conjunto de dados de *benchmark* para Sistemas de Detecção de Intrusão (IDS) disponibilizado pela Universidade de **New Brunswik** (UNB).
- Contém representações abstratas de eventos ocorridos em uma rede.
- É composto pela coletânea de 10 arquivos que representam 7 categorias de ocorrências: Benigno, Brute Force, Botnet, DoS, DDoS, Web attacks e Infiltration.
- Para fins deste trabalho foi utilizado um extrato do arquivo *Thuesday-20-02-2018\_TrafficForML\_CICFlowMeter* contento ataques do tipo DDoS.

# VISÃO GERAL DOS ARTEFATOS PRODUZIDOS

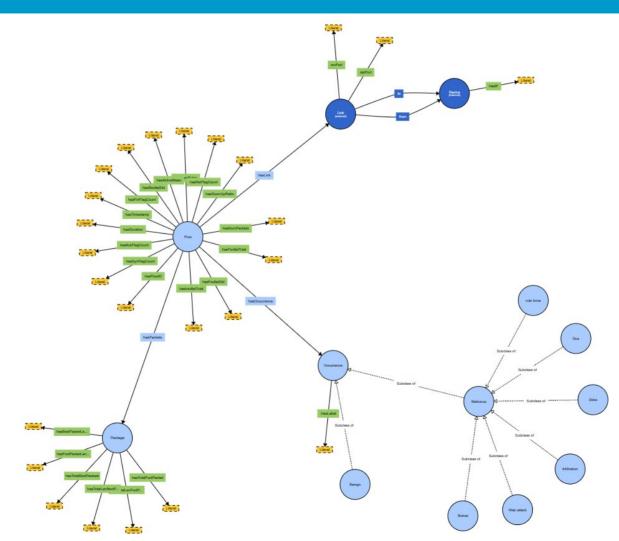
A ferramenta WebVOWL foi utilizada para a criação do modelo de dados.

- Principais artefatos produzido:
- 1) Modelo de dados (conjunto de entidades e relações observadas no domínio analisado)
  - 2) **Implementação** do modelo de dados no GraphDB

sistemas de redes de telecomunicações (prefixo "net").

- Para a representação semântica foi utilizada a ontologia ToCo (Toucan Ontology), desenvolvida para
- Características da modelagem:
- As **classes** e as **propriedades** foram definidas a partir das **características dos atributos** (agrupamento por funcionalidade).
- A quantidade de atributos foi reduzida por motivos didáticos.

# VISÃO GERAL DO MODELO DE DADOS



# MODELO DE DADOS - DESCRIÇÃO

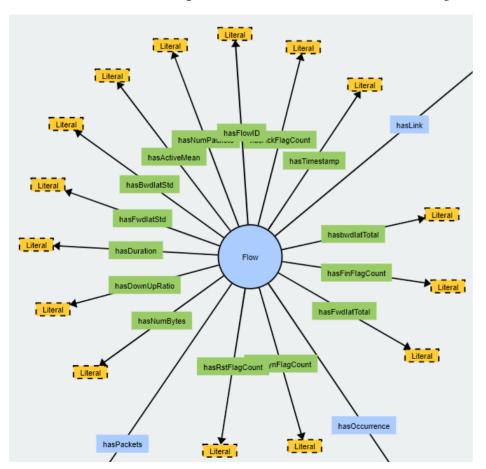
As **classes** e **subclasses** definidas são:

Classe	IRI	Descrição	Subclasse
Flow	ex:Flow	Descreve o fluxo bidirecional presente em cada tupla do dataset. Esta classe possui atributos que representam recursos estatísticos como duração, número de pacotes, número de bytes, comprimento dos pacotes, etc.	
Occurrence	ex:Occurrence	Descreve as ocorrências descritas em cada tupla do dataset.	Malicious Benign
Malicious	ex:Malicious	Descreve as ocorrências rotuladas como malignas em cada tupla do <u>dataset</u> .	Brute-force Botnet Dos Ddos Infiltration Web_attack
Benign	ex:Benign	Descreve as ocorrências rotuladas como benignas em cada tupla do dataset.	
Package	ex:Package	Descreve os pacotes atribuídos aos fluxos. Esta classe possui atributos que representam recursos estatísticos referentes ao pacote.	
Link	net:Link	Descreve um meio (por exemplo, cabo trançado, fibra óptica, onda eletromagnética) usado para conectar dois dispositivos na rede de telecomunicações.	
Device	net:Device	Define os dispositivos da infraestrutura física do sistema de telecomunicações.	

Tabela 1: Definição de classes. Fonte: Elaborada pela autora.

# **MODELO DE DADOS – DESCRIÇÃO**

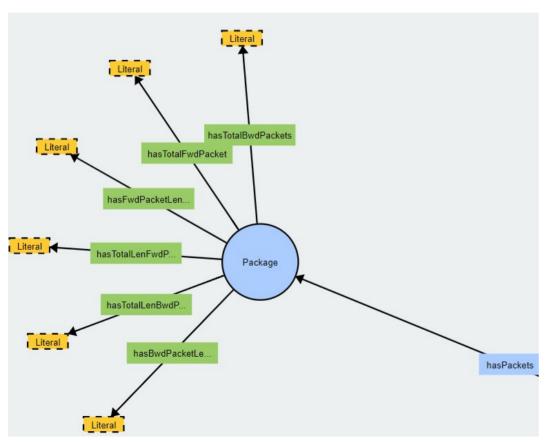
## Propriedades e Restrições – classe Flow



Propriedades	Domínio	Range
ex:hasTimestamp ex:hasDuration ex:hasByts ex:hasPkts ex:hasFwdlatTotal ex:hasFwdlatStd ex:hasBwdlatStd ex:hasFinFlagCount ex:hasSynFlagCount ex:hasRstFlagCount ex:hasAckFlagCount ex:hasAckFlagCount ex:hasAckFlagCount	Flow	Literal
ex:hasOccorrence	Flow	Occurrence
ex:hasPackets	Flow	Package
ex:hasLink	Flow	Link

# MODELO DE DADOS - DESCRIÇÃO

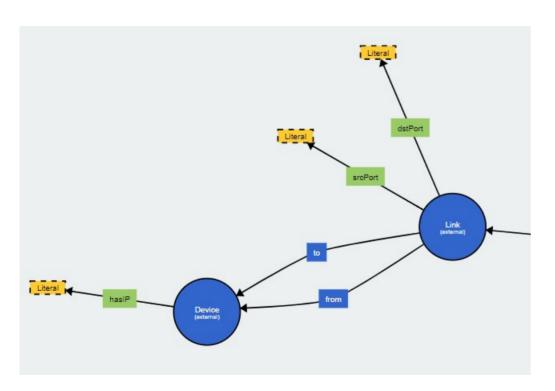
## Propriedades e Restrições – classe Package



Propriedades	Domínio	Range
ex:hasTotalFwdPacket ex:hasTotalBwdPackets ex:hasTotalLenFwdPacket ex:hasTotalLenBwdPacket ex:hasFwdPacketLengthStd ex:hasBwdPacketLenStd	Package	Literal

# MODELO DE DADOS - DESCRIÇÃO

## Propriedades e Restrições – classes Link e Device



#### Link

Propriedades	Domínio	Range
net:from* net:to*	Link	Device
ex:srcPort ex:dstPort	Link	Literal

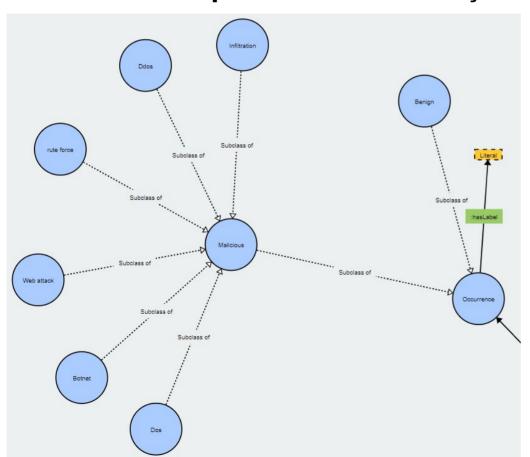
#### Device

Propriedades	Domínio	Range
ex:hasIP	Device	Literal

<sup>\*</sup> URI net: <a href="http://purl.org/toco/">

## **MODELO DE DADOS – DESCRIÇÃO**

## Propriedades e Restrições – classe Occurrence



Propriedades	Domínio	Range
ex:hasLabel	Occurrence	Literal

#### Observações:

- As subclasses herdam a propriedade da classe ocorrência.
- A classe Benign é disjunta da classe Malicious.

## <u>Definição dos prefixos</u>

Generated with the experimental alpha version of the TTL exporter of WebVOWL (version 1.1.7) http://visualdataweb.de/webvowl/

```
@prefix :
                     <http://example.com/CSE-CIC-IDS2018/> .
                     <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdf:
@prefix rdfs:
                      <http://www.w3.org/2000/01/rdf-schema#> .
@prefix owl:
                     <http://www.w3.org/2002/07/owl#> .
@prefix xsd:
                     <http://www.w3.org/2001/XMLSchema#> .
@prefix dc:
                     <http://purl.org/dc/elements/1.1/> .
@prefix xml:
                     <http://www.w3.org/XML/1998/namespace> .
@prefix wot:
                      <http://xmlns.com/wot/0.1/> .
@prefix vs:
                      <http://www.w3.org/2003/06/sw-vocab-status/ns#> .
Annafiv foaf.
                     /http://ymlns.com/fosf/0.1/>
@prefix ex:
                      <http://example.com/CSE-CIC-IDS2018/> .
@prefix :
                     <http://example.com/CSE-CIC-IDS2018/> .
@prefix net:
                      <http://purl.org/toco/> .
                      <http://example.com/CSE-CIC-IDS2018/> .
@base
<http://example.com/CSE-CIC-IDS2018/> rdf:type owl:Ontology ;
                                       dc:title "Modelo Formal Dataset CSE-CIC-IDS2018"@en;
                                       dc:description "Modelo Formal do dataset de IDS CSE-CIC-IDS2018, descrito usando o esquema W3C RDF e a Web Ontology
                                       Language. "@en;
                                       owl:versionInfo "1.0"@en;
                                       dc:creator "Thaisa" .
```

#### Exemplo de definição de classes

Classe <u>Malicious</u> é **subclasse** de <u>Occurence</u> e **disjunta** da classe <u>Benign</u>.

Exemplo da utilização da ontologia **TOCO** com o perfixo **net.**.

Exemplo de
enriquecimento de
dataset com a
utilização da
propriedade
owl:equivalentClass

rdfs:domain net:Link;
rdfs:range net:Device .

#### Exemplo de definição de propriedades

```
:hasTimestamp rdf:type owl:DatatypeProperty;
              rdfs:label "hasTimestamp"@IRI-based;
              rdfs:label "hasTimestamp"@iri-based;
              rdfs:label "hasTimestamp"@en;
              rdfs:domain :Flow:
              rdfs:range rdfs:Literal .
                        ----- Property 20----
:hasOccurrence rdf:type owl:ObjectProperty ;
              rdfs:label "hasOccurrence"@IRI-based;
              rdfs:label "hasOccurrence"@iri-based:
              rdfs:label "hasOccurrence"@en;
              rdfs:domain :Flow:
              rdfs:range :Occurrence .
          ----- Property 13--
 net:from rdf:type owl:ObjectProperty ;
         rdfs:label "from"@IRI-based;
         rdfs:label "from"@iri-based;
         rdfs:label "from"@en;
```

Exemplo de uma propriedade do tipo **DataTypeProperty** 

Exemplo de uma propriedade do tipo **ObjectProperty** 

Exemplo de utilização da ontologia **TOCO** com o perfixo **net.**.

#### Exemplo de instâncias

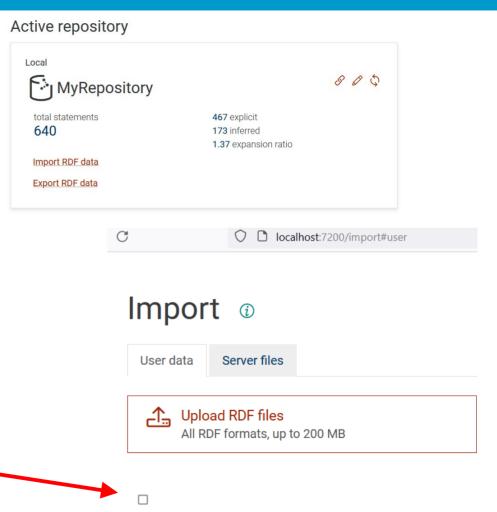
```
<http://example.com/CSE-CIC-IDS2018/flow 1> rdf:type :Flow;
                                          ex:hasFlowID "172.31.69.1-172.31.69.25-67-68-17" :
                                          ex:hasTimestamp "20/02/2018 08:50:51" :
                                          ex:hasDuration "716";
                                          ex:hasNumBytes "878491.6201";
                                          ex:hasNumPackets "2793.296089";
                                          ex:hasFwdIatTotal "0";
                                          ex:hasbwdIatTotal "0" :
                                          ex:hasFwdIatStd "0" :
                                          ex:hasBwdIatStd "0" :
                                          ex:hasFinFlagCount "0";
                                          ex:hasSynFlagCount "0";
                                          ex:hasRstFlagCount "0" :
                                          ex:hasAckFlagCount "0";
                                          ex:hasDownUpRatio "1";
                                          ex:hasActiveMean "0";
                                          ex:hasPackets <http://example.com/CSE-CIC-IDS2018/package 1>;
                                          ex:hasLink <http://example.com/CSE-CIC-IDS2018/link 1>;
                                          ex:hasOccurrence <http://example.com/CSE-CIC-IDS2018/occurrence 1>
```

Exemplo de uma instância da classe **Flow**.

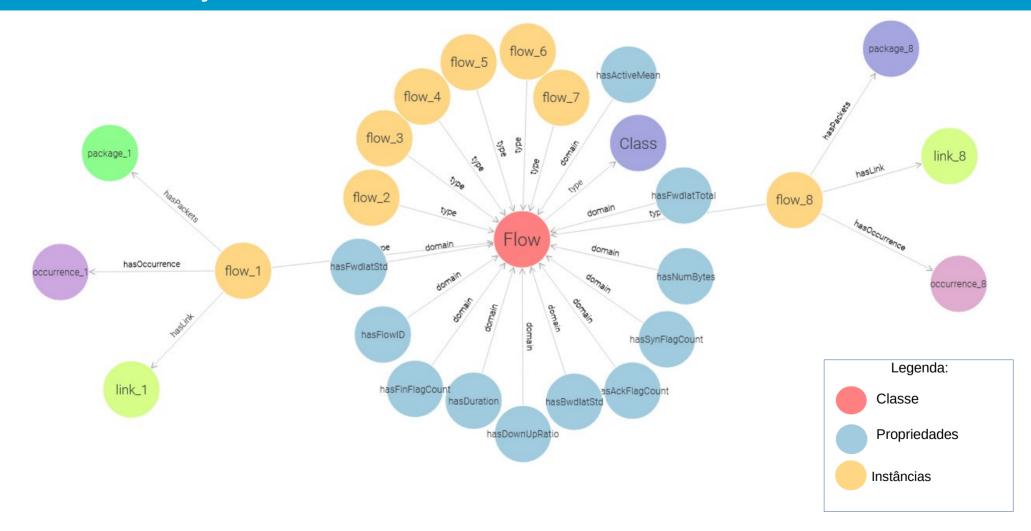
exemplo de uma de duas instâncias da classe **Device** e uma da classe Link.

# IMPLEMENTAÇÃO DO MODELO DE DADOS NO GRAPHDB

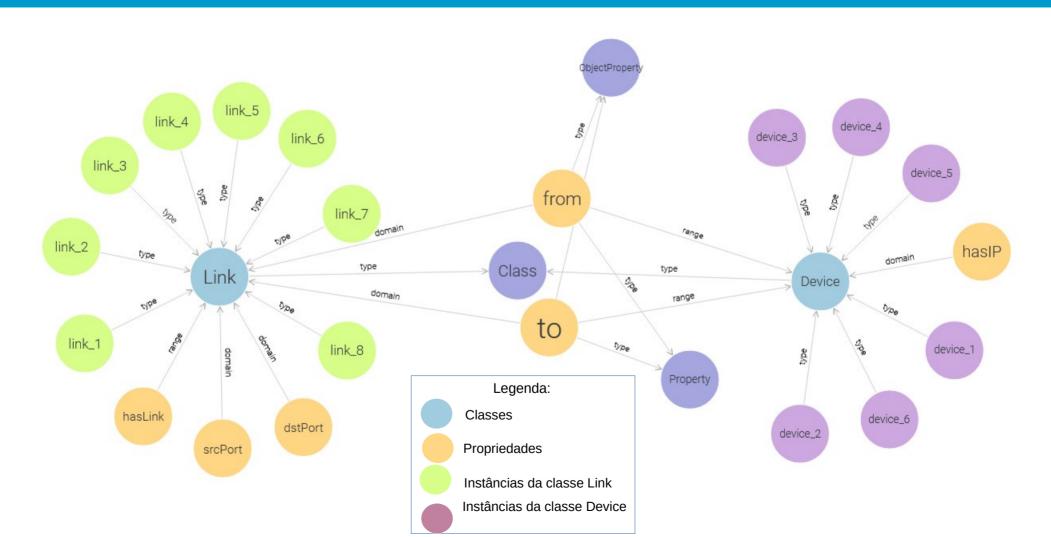
- O GraphDB é um banco de dados em grafo compatível com especificações RDF e SPARQL.
- O GraphDB foi instalado localmente na máquina da autora com acesso através da URL localhost:7200/.
- O modelo de dados no formato TTL foi importado no GraphDB.



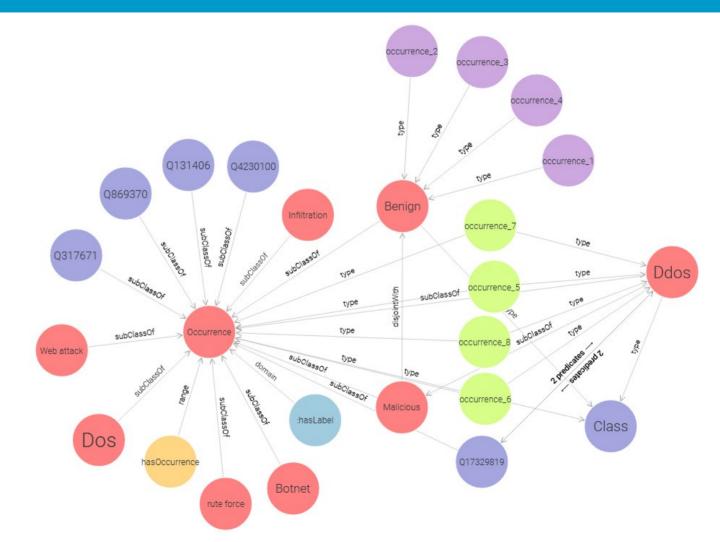
## IMPLEMENTAÇÃO NO GRAPHDB - CLASSE FLOW



## IMPLEMENTAÇÃO NO GRAPHDB - CLASSES LINK E DEVICE



## IMPLEMENTAÇÃO NO GRAPHDB - CLASSE OCCURRENCE E SUBCLASSES



Legenda:
Classes
Propriedades
Instância das classes Ddos e Maligno
Instâncias da classe Benigno

1) Selecionar o data/hora de todos os fluxos:

```
PREFIX ex: <http://example.com/CSE-CIC-IDS2018/>
select ?timestamp where {
    ?fluxo ex:hasTimestamp ?timestamp .
    }
```

	timestamp
1	"20/02/2018 08:50:51"
2	"20/02/2018 08:32:55"
3	"20/02/2018 09:29:47"
4	"20/02/2018 09:48:07"
5	"20/02/2018 10:13:54"
6	"20/02/2018 10:13:54"
7	"20/02/2018 10:13:54"
8	"20/02/2018 10:13:54"

2) Selecionar os endereços IP de origem e o tipo de ataque dos fluxos com ocorrências maligna:

```
PREFIX ex: <http://example.com/CSE-CIC-IDS2018/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX net: <http://purl.org/toco/>
PREFIX: <http://example.com/CSE-CIC-IDS2018>
select ?flow ?src_ip ?label where {
    ?flow ex:hasLink ?link .
    ?link net:from ?src_device .
    ?src_device ex:hasIP ?src_ip .
    ?flow ex:hasOccurrence ?occurrence .
    ?occurrence rdf:type ex:Malicious .
    ?occurrence ex:hasLabel ?label .
```

	flow	src_ip	label
1	ex:flow_6	"52.14.136.135"	"DDoS"
	ex:flow_5	"52.14.136.135"	"DDoS"
	ex:flow_7	"52.14.136.135"	"DDoS"
4	ex:flow_8	"52.14.136.135"	"DDoS"

3) Selecionar os endereços IP atacados e seus respectivos IP atacantes:

```
PREFIX ex: <a href="http://example.com/cse-cic-ids2018/">http://example.com/cse-cic-ids2018/</a>
PREFIX rdf: <a href="http://www.w3.org/1999/02/22-rdf-syntax-ns#">http://www.w3.org/2000/01/rdf-schema#</a>
PREFIX rdfs: <a href="http://purl.org/toco/">http://www.w3.org/2000/01/rdf-schema#</a>
PREFIX net: <a href="http://purl.org/toco/">http://purl.org/toco/</a>
PREFIX: <a href="http://example.com/cse-cic-ids2018">http://example.com/cse-cic-ids2018</a>
select ?src_ip ?dst_ip ?label where {
    ?flow ex:hasLink ?link .
    ?link net:from ?src_device .
    ?link net:to ?dst_device .
    ?src_device ex:hasIP ?src_ip .
    ?dst_device ex:hasIP ?dst_ip .
    ?flow ex:hasOccurrence ?occurrence .
    ?occurrence rdf:type ex:Malicious .
    ?occurrence ex:hasLabel ?label .
```

	src_ip \$	dst_ip	label
1	"52.14.136.135"	"172.31.69.25"	"DDoS"
2	"52.14.136.135"	"172.31.69.25"	"DDoS"
3	"52.14.136.135"	"172.31.69.25"	"DDoS"
4	"52.14.136.135"	"172.31.69.25"	"DDoS"

4) Informar a quantidade de ocorrências malignas.

```
PREFIX ex: <http://example.com/CSE-CIC-IDS2018/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX net: <http://purl.org/toco/>
select (COUNT(?occurrence)AS ?count)
where{
     ?flow ex:hasOccurrence ?occurrence .
     ?occurrence rdf:type ex:Malicious .
}
```

```
count

1 "4"^^xsd:integer
```

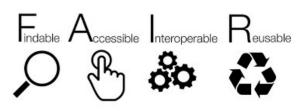
5) Execute uma consulta federada que retorne um resumo sobre os tipos de ataques ocorridos.

```
PREFIX ex: <http://example.com/CSE-CIC-IDS2018/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
select DISTINCT ?type ?description
where {
     ?flow ex:hasOccurrence ?occurrence .
     ?occurrence rdf:type ?type .
     ?type owl:equivalentClass ?x .
     SERVICE <https://query.wikidata.org/sparql>
     {?x <http://schema.org/description> ?description .
     FILTER (langMatches(lang(?description), "en")) .}
}
```

	label \$	description
1	"Ddos"@iri-based	"cyber attack" <sup>@en</sup>
2	"Ddos" <sup>@en</sup>	"cyber attack" <sup>@en</sup>

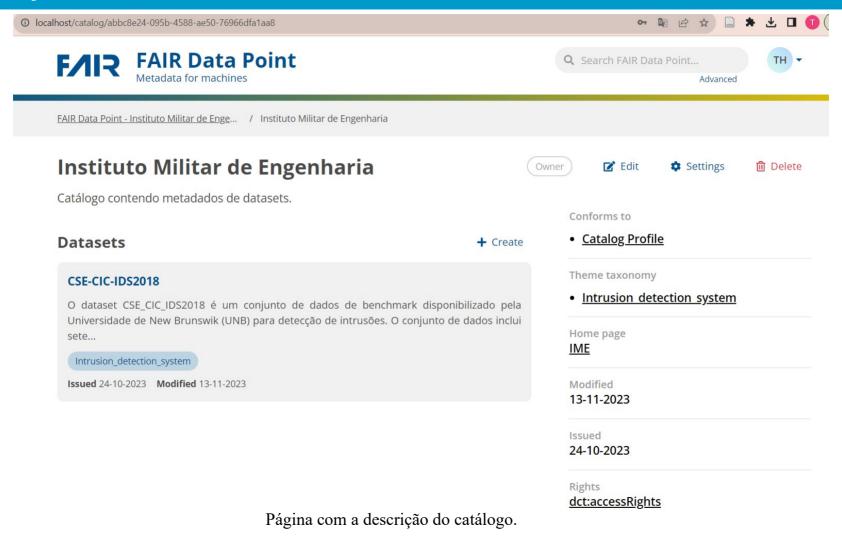
O **FAIR Data Point** é um servidor de aplicação que expõe metadados na internet seguindo os **princípios FAIR** (*Findable*, *Accessible*, *Reusable* and *Interoperable*) (WILKINSON et al., 2016).

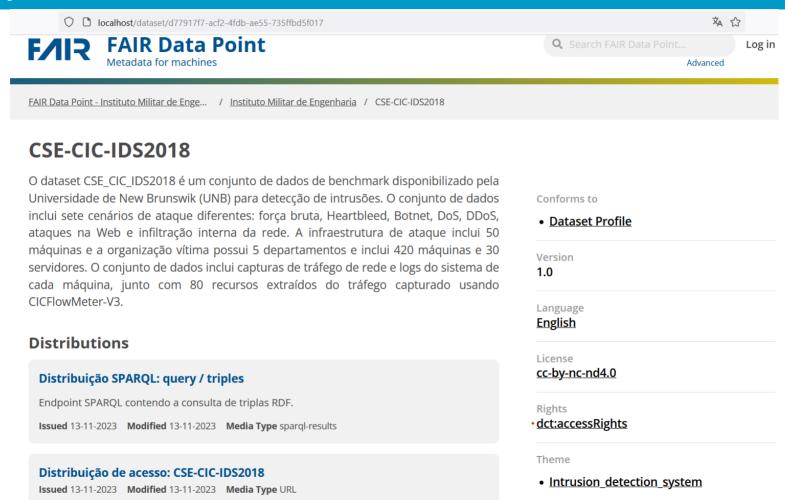
- Possui três objetivos principais:
- (i) Permitir que criadores exponham os metadados de seus objetos digitais de uma forma que segue os princípios FAIR.
- (ii) Permitir que os consumidores descubram informações sobre objetos digitais de interesse; e
- (iii) Forneça esses metadados de forma acionável por máquina.
- O FDP usa o modelo Data Catalog Vocabulaire (DCAT) versão 2 como base para seus metadados





Página inicial com a descrição do FAIR Data Point.





Página com a descrição do dataset e suas distribuições.

"Devidamente projetada, a **Web Semântica** pode ajudar na evolução do conhecimento humano como um todo" (BERNERS-LEE et al., 2001).

## **REFERÊNCIAS**

- BERNERS-LEE, Tim; HENDLER, James; LASSILA, Ora. The semantic web. Scientific american, v. 284, n. 5, p. 34-43, 2001.
- BREITMAN, K. K. Web semântica: a internet do futuro. [s.l.] Grupo Gen-LTC, 2000.
- Data Catalog Vocabulary. Disponível em: <a href="https://www.w3.org/TR/vocab-dcat-2/#classifying-datasets">https://www.w3.org/TR/vocab-dcat-2/#classifying-datasets</a>. Acesso em: 13 nov. 2023.
- DBPedia. Disponível em: <a href="https://dbpedia.org/page/">https://dbpedia.org/page/</a>. Acesso em: 20 nov. 2023.
- IDS 2018 | UNB. Disponível em: <a href="https://www.unb.ca/cic/datasets/ids-2018.html">https://www.unb.ca/cic/datasets/ids-2018.html</a>. Acesso em: 6 out. 2023.
- Introdução ao FAIR Data Point., 8 jun. 2020. Disponível em: <a href="https://www.youtube.com/watch?v=PtS">https://www.youtube.com/watch?v=PtS</a> ek7BXSA>. Acesso em: 20 out. 2023
- KENYON, A.; DEKA, L.; ELIZONDO, D. Are public intrusion datasets fit for purpose characterising the state of the art in intrusion event datasets. Computers & Security, v. 99, p. 102022, 1 dez. 2020.
- LAUFER, C. Guia da Web Semântica., 2015.
- MITTAL, N. K. A survey on Wireless Sensor Network for Community Intrusion Detection Systems. 2016 3rd International Conference on Recent Advances in Information Technology (RAIT). Anais... Em: 2016 3RD INTERNATIONAL CONFERENCE ON RECENT ADVANCES IN INFORMATION TECHNOLOGY (RAIT). mar. 2016. Disponível em: <a href="https://ieeexplore.ieee.org/document/7507884">https://ieeexplore.ieee.org/document/7507884</a>. Acesso em: 27 nov. 2023
- NIST. Disponível em <a href="https://www.nist.gov/news-events/news/2023/08/building-cybersecurity-and-privacy-learning-program-nist-releases-draft-sp">https://www.nist.gov/news-events/news/2023/08/building-cybersecurity-and-privacy-learning-program-nist-releases-draft-sp</a>. Acessado em 30 nov. 2023.

# **REFERÊNCIAS**

- SANTOS, L. O. B. DA S. et al. FAIR Data Point: A FAIR-Oriented Approach for Metadata Publication. Data Intelligence, v. 5, n. 1, p. 163–183, mar. 2023.
- SILVA, M. L. SEC4ML: Anonimização de Dados de Incidentes de Segurança da Informação para tarefas de Aprendizado de Máquina. 2022. Dissertação (Mestrado). Programa de Pós Graduação em Ciências em Sistemas e Computação. Instituto Militar de Engenharia, Rio de Janeiro, 2022.
- TOCO. Disponível em: <a href="https://qianruzhou333.github.io/toco">https://qianruzhou333.github.io/toco</a> ontology/#d4e1810>. Acesso em: 31 out. 2023.
- specs.fairdatapoint.org. Disponível em: <a href="https://specs.fairdatapoint.org/fdp-specs-v1.2.html">https://specs.fairdatapoint.org/fdp-specs-v1.2.html</a>. Acesso em: 13 nov. 2023.
- TORINO, E.; VIDOTTI, S. A. B. G.; CONEGLIAN, C. S. #SejaJUSTOeCUIDADOSO: princípios FAIR e CARE na gestão de dados de pesquisa. [s.l.] IBICT, 2021.
- WILKINSON, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. **Scientific Data**, v. 3, p. 160018, 1 mar. 2016.
- W3C. Disponível em: <a href="https://www.w3.org/TR/rdf11-primer/">https://www.w3.org/TR/rdf11-primer/</a>. Acesso em 30 nov. 2023.
- OWL Web Ontology Language. Disponível em: <a href="https://www.w3.org/TR/2004/REC-owl-ref-20040210/">https://www.w3.org/TR/2004/REC-owl-ref-20040210/</a>. Acesso em: 20 nov. 2023.
- ŽÁČEK, Martin; MIARKA, Rostislav; SÝKORA, Ondřej. Visualization of semantic data. In: Artificial Intelligence Perspectives and Applications: Proceedings of the 4th Computer Science On-line Conference 2015 (CSOC2015), Vol 1: Artificial Intelligence Perspectives and Applications. Springer International Publishing, 2015. p. 277-285.

# Obrigada!