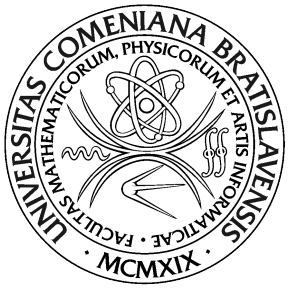
UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY



SÉMANTICKÉ PUBLIKOVANIE SPRAVODAJSKÝCH DÁT

Diplomová práca

2021 Bc. Matej Rychtárik

UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

SÉMANTICKÉ PUBLIKOVANIE SPRAVODAJSKÝCH DÁT

Diplomová práca

Študijný program:	Aplikovaná informatika
Študijný odbor:	2511 Aplikovaná informatika
Školiace pracovisko:	Katedra aplikovanej informatiky
Školiteľ:	doc. RNDr. Martin Homola, PhD.

Bratislava, 2021 Bc. Matej Rychtárik

Čestne prehlasujem, že túto diplomovú prácu som vypracoval samostatne len s použitím uvedenej literatúry a za pomoci konzultácií u môjho školiteľa.

. . . . . .

Bratislava, 2021 Bc. Matej Rychtárik

Poďakovanie

Touto cestou by som sa chcel v prvom rade poďakovať môjmu školiteľovi doc. RNDr. Martinovi Homolovi, PhD. za jeho cenné rady a usmernenia, ktoré mi veľmi pomohli pri riešení tejto diplomovej práce.

Abstrakt

Abstract

Obsah

Kapitola 1

Úvod

Nejaky strucny uvod do problematiky

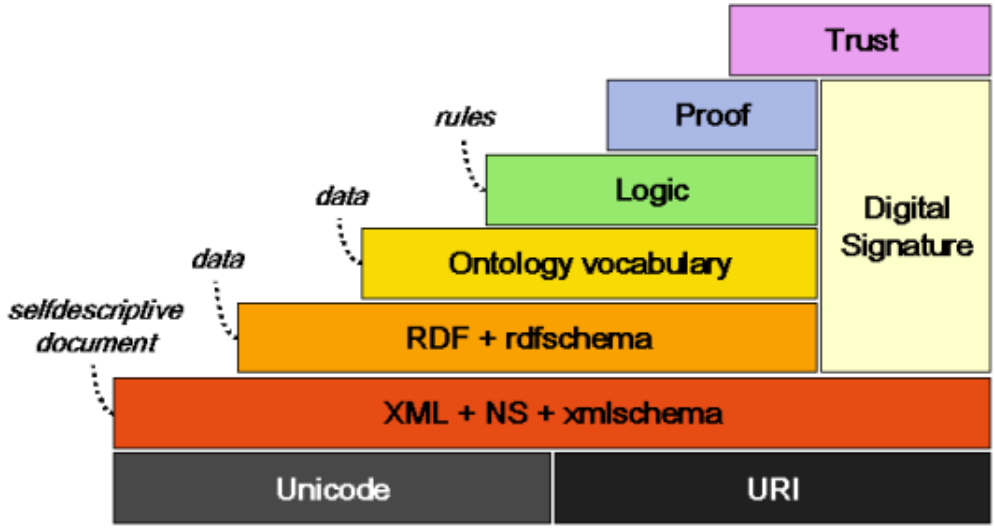
Časť I

Prehľad problematiky

Kapitola 2

Sémantický web

Semantický web [?] poskytuje spoločný framework, ktorý umožňuje zdieľanie a opätovné použitie údajov v rámci aplikácií. Štandardy podporujú spoločné dátové formáty a protokoly, kde najpodstatnejším je Resource Description Framework (RDF). Prvýkrát pojem Semantický web zaviedol Tim Berners-Lee a popisoval "dátový web", ktorý môže byť strojovo čitateľný. Zámerom je zvýšiť použiteľnosť webu a jeho prepojených zdrojov vytvorením sémantického webu. Semantický web má vrstvovú štruktúru ako si môžeme všimnúť na obrázku 2.1. Jednotlivé údaje sú potrebné až vo vyšších vrstvách. XML vrstva zaručuje, že môžeme spájať Semantický web s inými normami založenými napríklad na XML, ktorá je rozšírená a podporovaná a RDF dáta sa v nej dajú dobre prenášať, spracovávať a uchovávať. RDF a RDFS vrstva definuje typ zdrojov. Ontologická vrstva podporuje vývoj ontológií, vďaka ktorým môžeme definovať vzťahy medzi rôznymi pojmami.



Obr. 2.1: Semantic Web - vrstvy.
Zdroj: [?]

MH: ↑Toto trochu povrchne: (1) Ucelom SW nie je vyssia pouzitelnost webu (to je nepresne), ale je to lepsia pristupnost informacii publikovanych na webe pre strojove spracovanie. (2) Ak chces popisovat vrstvy SW podla tohto diagramu, bolo by dobre keby si popisal vsetky vrstvy – U XML by som sa obmedzil na to, ze je to proste dobry format pre textovu reprezentaciu dat v suboroch a pre ich vymenu medzi softvermi – toto vsak uz je dnes prekonane, uz vymiename SW data aj ako JSON, embedujeme ich do HTML5 (chcelo by to poznamku) – O RDF a RDFS si vlastne nic uzitocne (z coho citatel nieco vyrozumie) nepovedal – no a ostatne vrstvy si uplne preskocil
MH: Toto si menil? Zda sa mi, ze trochu asi aj ano, ale o tych dalsich vrstvach si nic nenapisal

Text uvedený nižšie popisuje niekoľko technológií, ktoré sú potrebné pre tvorbu sémantického webu.

2.1 Linked Data

MH: ↓Tato sekcia je celkom fajn ale chybalo mi trochu premostenie od SW – linked data bola iniciativa, ze ked uz SW formaty mame, podme v nich aj data zverenjnovat

MH: Inak ked uz si sa rozhodol pisat po slovensky, mal by si pouzivat aj slovensku terminologiu (co je celkom peklo) – ale teda Linked Data su *prepojene data*, LD network je *siet prepojenych dat* – ludia to takto pouzivaju

Linked Data [**?**] je metóda zverjňovania štrukturovaných dát. Ich hlavným cieľom je poprepájať existujúce databázy (primárne písané v RDF formáte), medzi rôznymi údajmi a umožniť ľuďom zdielať štrukturované dáta na webe pomocou HTML. Časť vízie do budúcna je, aby sa Internet stal globálnou databázou. Princípy Linked Data prvýkrát načrtol Tim Berners-Lee. Popísal 4 pravidlá pre zverejňovanie dát na webe:

1. používať URI ako názvy objektov, ktoré sú identifikátormi informácie, jej umiestnenia a ďalších vlastností,
2. používať HTTP URI, aby si ich ľudia vedeli pozrieť,
3. uvádzať informácie o tom, čo názov identifikuje pri vyhľadávaní pomocou otvorených štandardov, ako sú napríklad RDF alebo SPARQL,
4. pri publikovaní údajov na webe, zahrnúť odkazy aj na iné URI, aby sa dalo objavovať viac vecí.

Sú známe aj ako Linked Data princípy.

MH: Tu mi chyba informacia, ze sa tato inciativa ujala, a ze vdaka tomu vznikla na webe tzv. siet prepojenych dat, ktora obsahuje obrovske mnozstvo datovych zdrojov a nieco viac o tej sieti.

2.2 Resource Description Framework (RDF)

RDF [**?**] je štandartný model na zakódovanie metadát a ďalších informácií. Je to taktiež formát, ktorý bol navrhnutý a štandardizovaný na reprezentáciu dát pre sémantický web. Zdroje týchto dát sú väčšinou webové zdroje, ktoré môžu byť čokoľvek, napríklad dokumenty, ľudia, fyzické objekty, atď. Taktiež poskytuje spoločný framework na vyjadrenie informácií a možnosť zdieľať ich medzi softvérmi, bez straty ich hodnoty. Dáta sa uchovávajú v Triple Store databázach, ktorých formát je striktne daný. Výhodou je, že dáta môžu byť spracované aj softvérmi, pre ktoré dané dáta neboli vytvorené.

MH: ↑Na co RDF sluzi sa uz citatel dozvedel v skorsich castiach (ked to tam lepsie ozrejmis). Niektore veci, ktore tu ↑pises su nepresne (napr. to o tych metadatach a "dalsich informaciac" alebo o zdrojoch. Tiez o Triple Stores predbiehas, budes o tom pisat neskor... Asi by som to tu skratil a len by som nadviazal, ze RDF je zakladny datovy format pre SW a tu ho popiseme...

MH: ↓Zvysok ide dobrym smerom, je to presne to, co by som si predstavoval, ze tu budes pisat, len by som to chcel vidiet mozno trochu pomenej, podrobnejsie, systmatickejsie prebrate... Na vacsom priestore, mozno postupne ten priklad budovat... Vysvetlit na nom vsetky zakladne moznosti RDF

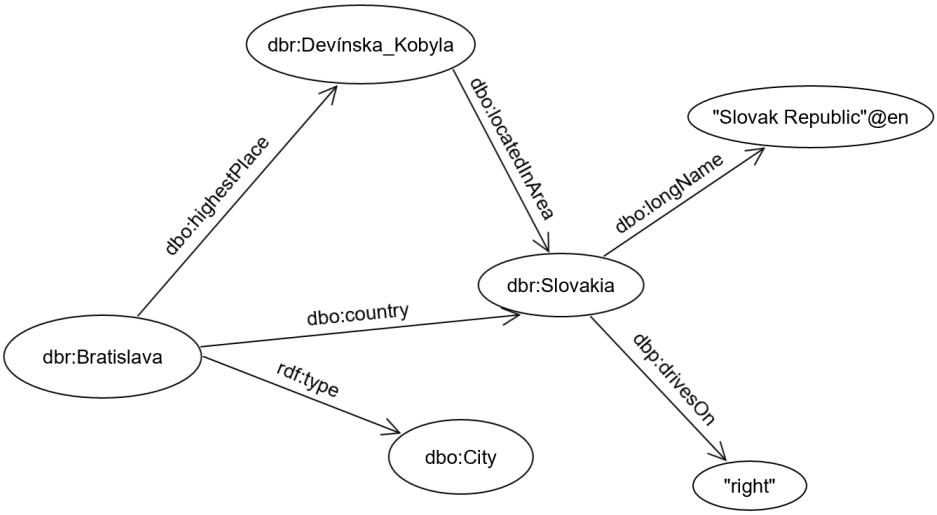
RDF súbor je taký dokument, ktorý ukladá RDF grafy do špecifického formátu serializácie pre RDF, ako sú napríklad N-Triple, TURTLE, RDF/XML a mnohé ďalšie. RDF bol postavený na myšlienke vytvárať údaje vo forme predmet-predikát-objekt, ktorý sa volá "triple", ďalej len trojica. Trojica je základná stavebná jednotka akejkoľvek množiny dát zapísaných v RDF. Tieto údaje sú reprezentované ako orientované grafy. Predmet a objekt predstavujú vrcholy a predikát je orientovaná hrana medzi nimi. Predmet môže byť použítý aj ako objekt v inej trojici. Týmto spôsobom sa trojice prepájajú a vzniká z nich grafová databáza. Predmet je vždy definovaný ako URI a popisuje zdroj informácie. Objekt môže byť taktiež nejaké URI popisujúce

zdroj, ale taktiež to môže byť primitívna hodnota ako napríklad string, integer, date, atď. Predikát popisuje, aký vzťah alebo rola medzi predmetom a objektom existuje. Predikát je vždy reprezentovaný ako URI, ktoré pochádza z ontolológií (kolekcie viacerých URI).

Na uľahčenie ukladania a čitateľnosti dát sa využívajú takzvané prefixy, ktoré sú preddefinovaním základných URI, do ktorých sa dodáva zvyšná hodnota URI pomocou dvojbodky, ako je to uvedené v nasledujúcom príklade a graficky znázornené v obrázku 2.2.

```
@prefix  rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns> .
@prefix  dbr: <http://dbpedia.org/resource/> .
@prefix   dbo: <http://dbpedia.org/ontology/> .
@prefix  dbp:<http://dbpedia.org/property/> .

dbr:Bratislava dbo:highestPlace dbr:Devínska_Kobyla .
dbr:Bratislava rdf:type dbo:City .
dbr:Bratislava dbo:country dbr:Slovakia .
dbr:Devínska_Kobyla dbo:locatedInArea dbr:Slovakia .
dbr:Slovakia dbp:drivesOn "right" .
dbr:Slovakia dbo:longName "Slovak Republic"@en .
```



Obr. 2.2: Príklad grafovej databázy.

Kapitola 3

Ontológie

MH: Predpokladam, ze o ontologiach budeme pisat o znacny kus viac a podrobnejsie... Mozes vychadzat z mojej prednasky ale tiez napr z uvodnej kapitoly *Handbook on Ontologies*

MH: Pozn. k strukture prace: Bolo by lepsie, keby *Prehlad problematiky* bola cast `\part{}`, *Semanticy web* samostatna kapitola`\chapter{}`, no a potom si myslim, ze *Ontologie* by tiez mali mat vlastnu kapitolu, ktora bude nasledovat hned po SW. Tretou kapitolou by potom mohol byt prehlad ontologii z oblasti bezpecnosti, co by uz do prehladu problematiky mohlo aj stacit

Výraz ontológia [**?**] pochádza z gréckeho slova kde '*ontos*' znamená existencia a '*logos*' znamená veda. Ontológia v informatike je uceleným popisom pojmov v určitej oblasti záujmu. Obsahuje určitú klasifikáciu údajov do hierarchicky usporiadaných kategórií a množinu odvodzovacích pravidiel, pomocou ktorých je možné z faktov odvodiť nové skutočnosti. Prostredníctvom ontológií je možné vytvárať spojenia v prirodzenom jazyku, vykonávať analýzu údajov a sprostredkovať výhody webu obohateného o sémantiku.

MH: nerozumiem co myslis pod „vytvárať spojenia v prirodzenomjazyku“ a tiez tvrdenie „sprostredkovať výhody webu obohateného o sémantiku“ je velmi abstraktne neviem celkom prist na to co si tym asi myslel

3.1 Základné pojmy

Ontológia sa skladá zo základných stavebných prvkov *Trieda*, *Entita*, *Atribút*, *Väzba*.

Triedy alebo typy definujú skupiny alebo množiny objektov. Triedy majú hierarchickú štruktúru zloženú z ich podtried. Každá podtrieba spĺňa vlastnosti nadtriedy a môže byť rozšírená o vlastné vlastnosti.

Entity sú individuálne inštancie nejakej nami zadefinovanej triedy. Ak by sme mali entitu *Jablko*, a triedy *Ovocie* a *Jedlo*, kde *Ovocie* je podtriedou *Jedlo*, tak nám z ontológie vyplýva, že ak je entita *Jabllko individuálnou inštanciou triedy Ovocie, tak je aj individuálnou inštanciou triedy Jedlo.*

MH: ↑ *Jablko* nie je velmi dobry priklad na entitu, kedze vacsinou ho uvazujeme ako triedu (a teda podtriedu triedy Ovocie)

Atribúty sú vlastnosti *Tried* a *Entít* a môžu niesť rôzne informácie o danom objekte. *Atribúty* môžu mať rôzne hodnoty, ako reťazec, číslo, dátum, pravdivostnú hodnotu alebo inú premennú. Ak by sme si zobrali predchádzajúcu entitu *Jablko*, jej číselná vlastnosť môže byť napríklad mesiac kedy sa oberá alebo jej reťazcová vlastnosť farba.

MH: ↑ Je pomerne nezvycajne aby mali atributy ako hodnotu inu premennu

Väzby sú najpodstatnejšou súčasťou ontológie. Poskytujú prepájanie jednotlivých objektov tried. Je to jednosmerné spojenie, ktoré určuje vzťah, v akom sú dve dané triedy. Tým vznikne trojica *trieda:vzťah:trieda*, ktorá sa nazýva triplet. Medzi triplet sa radí aj trojica *trieda:atribút:hodnota*. Väzby sa zvyknú definovať aj inverzne.

MH: ↑ nazov *vazby* je velmi exoticky. Ak ti ide o object properties, pouzil by som slovo *vztahy*.

MH: ↑ to co je *triplet* by si mal zadefinovat vyssie, kde pises o RDF.

MH: ↑ Co su „objekty tried“? Kusok vyssie si si pre objekty zvolil nazov *entity*, mal by si ho teda pouzivat. Ak ti ide o vztahy medzi triedami, ako napr. vztah podtriedy a nadtriedy, v RDF je to sice vyjadrene pomocou vlasntosti, ale z logickeho hladiska to chapeme ako *axiom*

Ontológia má veľa vlastností, ktoré musia byť dodržané. Každý prvok musí byť jasne indetifikovateľný. Taktiež zakazuje zapisovanie duplicitných dát, čo nám zaobstará vlastnosť efektívneho ukladania informácií, kde to môže nie len uľahčiť vyhľadávanie ale aj obsah pamäti na disku.

MH: ↑ Neviem, ci zrovna tieto vlastnosi ontologii su tie najpodstatnjsie, a teda treba ich spominat na tomto mieste.

3.2 Využitie ontológií

MH: ↓ Asi by som sa neodvazoval tvrdit to co pises v prvej vete. Ak zoberieme ako ranne vyuzitie napr. SNOMED, nie je to pravda, kedze sa vyuzival v medicinskej praxi. S druhou vetou mozno nesuhlasit. Kto je bezny pouzivatel? Nikomu takemu sa nerozsirila do pocitaca nejaka ontologia (ze by o tom vedel)...

Ontológie sa začali využívať najmä v organizáciách, ktoré sa šoecializovali na umelú inteligenciu. Neskôr sa to rozšírilo aj do počítačov bežných používateľov. Napríklad firma Amazon používa ontológie na kategorizovanie tovaru v ich elektronickom obchode.

Na zápis týchto ontológií sa používa niekoľko jazykov, kde najznámejším je asi Resource Description Framework (RDF), ktorý je rpimárne určený na využitie vo webových stránkach, pre hľadanie informácií strojmi. Ontológie si našli uplatnenie aj v medicínskej oblasti a to napríklad SNOMED, čo je najväčším viacjazyčným medicínskym slovníkom na svete.

MH: ↑ Tu chces asi pisat o vyuziti ontologii, takze zmienka o jazyku RDF sem nepatri, navyse RDF nie je jazyk na zapis ontologii (tym je RDFS) a o RDF su uz pisal vyssie

Taktiež sa s ním stretávame každodenne pri vyhľadávaní na stránke Google, kde ako bočný panel sú zobrazené informácie o vyhľadávanom objekte (obrázok nižšie). Tieto dáta je možné zobraziť preto, lebo výsledkom takéhoto panelu je vyhľadávanie informácií na webovej stránke, ktorá obsahuje sémantické dáta.

MH: ↑ SNOMED a Google spominas dobre, a v druhom pripade si nespomenuk ziadnu ontologiu a pritom ju dobre pozname (Schema.org)

Na získavanie dát zo sémantických webov a z RDF úložísk sú využívane SPARQL dopyty. Syntax jazyku SPARQL je veľmi podobná klasickému SQL jazyku, kde aj SPARQL umožňuje okrem dopytovania aj vkladanie, editáciu a vymazávanie dát.

MH: ↑ toto mi sem opat nesedi, zrejme tomu chces venovat nejaku kratku podsekciu skor v predchadzajucej kapitole (?)

3.3 Ciele ontológie

Zadefinovanie a zdieľanie jednotného zápisu informácií pre danú doménu. Ak napríklad viac stránok využíva na popis pojmov takúto zadefinovanú ontológiu, vedia boti získať a vyhľadávať viac dát o hľadanej informácii.

Taktiež je jej cieľom opätovné použitie ontológie, napríklad ak máme dobre zadefinovanú ontológiu, môžu ontologický inžinieri doplniť do našej ontológie ďalšie vlastnosti a tým by základ ontológie bol rovnaký ale bol by rozšírený o určité dáta, podľa potreby ontologických inžinierov.

MH: ↑ Tuto cast mozno lespie najko pouzit v uvode predch. casti (?)

3.4 Syntax ontológií

MAM SPRAVIT AJ TOTO?

MH: Kedze cela Tvoja praca je o ontologiach a budes ich zrejme nejako zapisovat mozno by si mohol nieco povedat aj o jazykoch na zaspis ontologii, minimalen apson o jednom, s ktorym budes pracovat dalej (cize zrejme OWL)

Kapitola 4

Existujúce ontologické riešenia v oblasti bezpečnosti

STRUCNY POPIS KAPITOLY

4.1 CTI model

TEN BY SOM MOZNO ZAHRNUL SEM CI?

4.2 UCO

MH: Lepsie ked v nadpisoch nebudes pouzivat skratky

MH: Chybaju tu akekolvek referencie

Unified Cybersecurity Ontology alebo skrátene UCO je rozšírením pôvodného projektu Intrusion Detection System (IDS), ktorého tvorcom je rovnaká skupina. Spája viaceré bežne dostupné bezpečnostné štandardy používané v kybernetickej bezpečnosti. Prevažne pokrýva STIX, ktorý je najväčším a najkomplexnejším štandardom, pokrývajúcim najväčšiu časť kybernetickej bez-

pečnostnej domény ale taktiež poskytuje iné relevatné štandardy ako CVE4, CCE5, CVSS6, CAPEC7, CYBOX8, KillChain9 a STUCCO10.

MH: ↑ Zrejme muslis *pokryva* ine relevantne standardy? Ak nie, nerozumiem, v akom zmysle ich poskytuje?

Aj keď je STIX najkomplexnejším štandardom a zjednocuje všetky informácie o kybernetických hrozbách, má tieto dáta uložené v XML súboroch, takže nepodporuje výhody uvažovania v ontológiách, čo UCO poskytuje.

MH: ↑ *uvazovanie v ontologiach* nie je spravny slovensky vyraz pre reasoning – skus napr. *inferencia*

Okrem týchto štandardov obsahuje aj mapovanie na všeobecné ~~svetové~~ databázy ako sú Google Knowledge Graph, DBPedia a Yago. Vďaka týmto mapovaniam je možné mať prístup k verejným databázam r rôznych domén záujmu.

Základnými triedami, využívanými v UCO sú:

- *Means* – Čo je zamýšľané daným útokom.
- *Consequences* – Dôsledky útoku.
- *Attack* – Typ útoku.
- *Attacker* – Kto je inicitorom daného útoku.
- *Attack-Pattern* – Vzorec útoku, podľa ktorého je útok riadený.
- *Exploit* – K čomu útok slúži.
- *Exploit Target* – K čomu slúži cieľ alebo výsledok útoku.
- *Indicators* – Indikátor útoku.

Každá z týchto tried je mapovaná na už reálne existujúcu triedu v niektorom z vyššie uvedených štandardov, prevažne na STIX schému.

Ontológia UCO umožňuje analytikom zachytávať špecifické vedomosti o kybernetickej bezpečnosti pomocov termínov a tried z ontológie a taktiež umožňuje písať pravidlá, ktoré sa môžu použiť na odvodenie nových poznatkov.

Vývojári extrahovali dáta z National Vulnerability Database (NVD), ktorá je uložená v XML súboroch. Potom boli namapované na triple store DBPedia a dáta boli uložené na FUSEKI server, ktorý podporuje dopytovanie z rôznych zdrojov rovnako ako ich odvodzovanie.

4.3 ICAS

4.4 •

Literatúra

[MB17] Vasileios Mavroeidis and Siri Bromander. Cyber threat intelligence model: An evaluation of taxonomies, sharing standards, and ontologies within cyber threat intelligence. In Joel Brynielsson, editor, *European Intelligence and Security Informatics Conference, EISIC 2017, Athens, Greece, September 11-13, 2017*, pages 91–98. IEEE Computer Society, 2017.

[OCM12] Leo Obrst, Penny Chase, and Richard Markeloff. Developing an ontology of the cyber security domain. In *STIDS*, pages 49–56, 2012.

[OCWM14] Alessandro Oltramari, Lorrie Faith Cranor, Robert J Walls, and Patrick D McDaniel. Building an ontology of cyber security. In *STIDS*, pages 54–61. Citeseer, 2014.

[PUJF03] John Pinkston, Jeffrey Undercoffer, Anupam Joshi, and Timothy Finin. A target-centric ontology for intrusion detection. In *Procs. of the IJCAI-03 Workshop on Ontologies and Distributed Systems*, 2003.

[SPF+16] Zareen Syed, Ankur Padia, Tim Finin, M. Lisa Mathews, and Anupam Joshi. UCO: A unified cybersecurity ontology. In Da-

vid R. Martinez, William W. Streilein, Kevin M. Carter, and Arunesh Sinha, editors, *Artificial Intelligence for Cyber Security, Papers from the 2016 AAAI Workshop, Phoenix, Arizona, USA, February 12, 2016*, volume WS-16-03 of *AAAI Workshops*. AAAI Press, 2016.

[TPKN18] Takeshi Takahashi, Bhola Panta, Youki Kadobayashi, and Koji Nakao. Web of cybersecurity: Linking, locating, and discovering structured cybersecurity information. *Int. J. Communication Systems*, 31(3), 2018.

Zoznam obrázkov