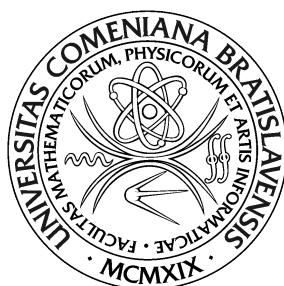


UNIVERZITA KOMENSKÉHO V BRATISLAVE  
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY



# 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



Univerzita Komenského v Bratislave  
Fakulta matematiky, fyziky a informatiky

## ZADANIE ZÁVEREČNEJ PRÁCE

**Meno a priezvisko študenta:** Bc. Matej Rychtárik  
**Študijný program:** aplikovaná informatika (Jednoodborové štúdium, magisterský II. st., denná forma)  
**Študijný odbor:** informatika  
**Typ záverečnej práce:** diplomová  
**Jazyk záverečnej práce:** slovenský  
**Sekundárny jazyk:** anglický

**Názov:** Sémantické publikovanie spravodajských dát o bezpečnostných hrozbách  
*Semantic punishing of security threat intelligence data*

**Anotácia:** V súčasnosti zaznamenávame veľké množstvo nových spravodajských dát o rôznych bezpečnostných hrozbách. Pre popis a publikovanie týchto dát vznikli v minulosti viaceré štandardy. Nový trend v oblasti však ukazuje potrebu sémantickej anotácie týchto dát za účelom zvýšenia ich dosahu a interoperability.

**Cieľ:** Cieľom je navrhnúť vhodnú ontológiu pre publikovanie spravodajských dát o bezpečnostných hrozbách a vytvorenie repozitára za týmto účelom v sieti prepojených dát.

**Literatúra:** [1] Allemang, D. and Hendler, J., 2011. Semantic web for the working ontologist: effective modeling in RDFS and OWL. Elsevier.  
[2] Heath, T. and Bizer, C., 2011. Linked data: Evolving the web into a global data space. Synthesis lectures on the semantic web: theory and technology, 1(1), pp.1-136.  
[3] Mavroeidis, V. and Bromander, S., 2017. Cyber threat intelligence model: an evaluation of taxonomies, sharing standards, and ontologies within cyber threat intelligence. In EISIC 2017 (pp. 91-98). IEEE.

**Vedúci:** doc. RNDr. Martin Homola, PhD.  
**Rektorát, dekanát:** FMFI.Dek - Dekanát  
**Dátum zadania:** 02.10.2019

**Dátum schválenia:** 14.10.2019  
prof. RNDr. Roman Ďurikovič, PhD.  
garant študijného programu

.....  
študent

.....  
vedúci práce

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

# Pod'akovanie

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

<b>1</b>	<b>Úvod</b>	<b>1</b>
<b>I</b>	<b>Prehľad problematiky</b>	<b>2</b>
<b>2</b>	<b>Sémantický web</b>	<b>3</b>
2.1	Linked Data . . . . .	4
2.2	Resource Description Framework (RDF) . . . . .	5
<b>3</b>	<b>Ontológie</b>	<b>8</b>
3.1	Základné pojmy . . . . .	9
3.2	Využitie ontológií . . . . .	10
3.3	Ciele ontológie . . . . .	10
3.4	Syntax ontológií . . . . .	11
<b>4</b>	<b>Existujúce ontologické riešenia v oblasti bezpečnosti</b>	<b>12</b>
4.1	CTI model . . . . .	12
4.2	UCO . . . . .	12
4.3	ICAS . . . . .	14
4.4	• . . . . .	14



# Kapitola 1

## Úvod

Nejaky strucny uvod do problematiky

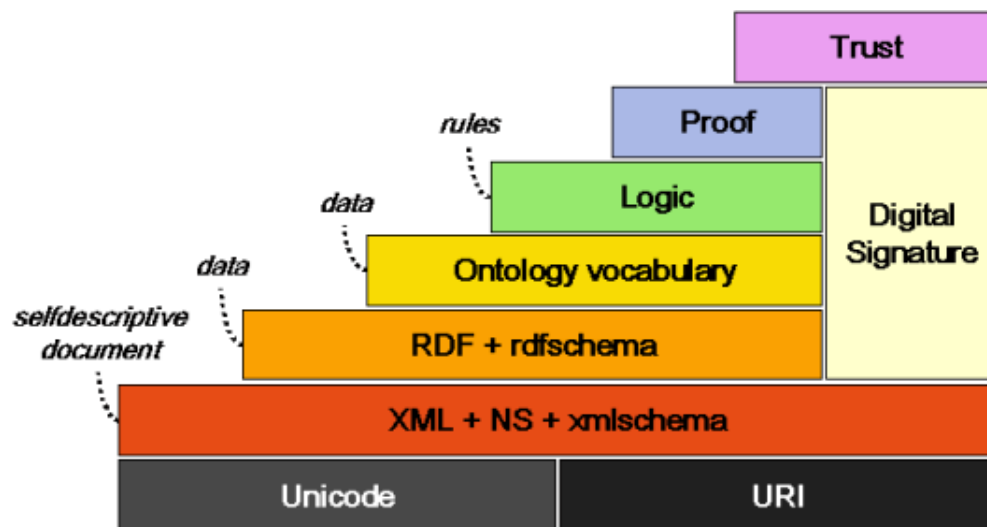
# Časť I

## 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 povrchné: (1) Ucelom SW nie je vyššia použiteľnosť webu (to je nepresne), ale je to lepšia prístupnosť informácií publikovaných na webe pre strojové spracovanie. (2) Ak chceš popisovať vrstvy SW podľa tohto diagramu, bolo by dobre keby si popísal všetky vrstvy – U XML by som sa obmedzil na to, že je to prostý dobrý formát pre textovú reprezentáciu dát v súboroch a pre ich výmenu medzi softvermi – toto však už je dnes prekonané, už vymieňame SW dáta aj ako JSON, embedujeme ich do HTML5 (chcelo by to poznámku) – O RDF a RDFS si vlastne nič užitočné (z čoho čitateľ niečo vyrozumie) nepovedal – no a ostatné vrstvy si úplne preskocil

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

## 2.1 Linked Data

MH: ↓Tato sekcia je celkom fajn ale chýbalo mi trochu premostenie od SW – linked data bola iniciatíva, že keď už SW formáty máme, podme v nich aj dáta zverejňovať

MH: Inak keď už si sa rozhodol písať po slovensky, mal by si používať aj slovenskú terminológiu (čo je celkom peklo) – ale teda Linked Data sú *prepojené dáta*, LD network je *sieť prepojených dát* – ľudia to takto používajú

Linked Data [?] je metóda zverejňovania štrukturovaných dát. Ich hlavným cieľom je poprepájať existujúce databázy (primárne písané v RDF formáte), medzi rôznymi údajmi a umožniť ľuďom zdieľať štrukturované dáta na webe pomocou HTML. Časť vízie do budúcnosti je, aby sa Internet stal globálnou databázou. Princípy Linked Data prvýkrát načrtol Tim Berners-Lee. Popísal 4 pravidiel 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 chýba informácia, že sa táto iniciatíva ujala, a že vďaka tomu vznikla na webe tzv. sieť prepojených dát, ktorá obsahuje obrovské množstvo dátových zdrojov a niečo viac o tej sieti.

## 2.2 Resource Description Framework (RDF)

RDF [?] je štandardný 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 softvérmí, bez straty ich hodnoty. Dáta sa uchovávajú v Triple Store databázach, ktorých formát je striktne daný. Výhodou je, že dáta môžu byť spracované aj softvérmí, pre ktoré dané dáta neboli vytvorené.

MH: ↑Na čo RDF sluzi sa už citateľ dozvedel v skorsich castiach (ked to tam lepsie ozrejmis). Niektore veci, ktore tu ↑pises su nepresne (napr. to o tych metadatach a “dalsich informaciach” alebo o zdrojoch. Tiez o Triple Stores predbiehas, budeš 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 dobrým smerom, je to presne to, čo by som si predstavoval, že tu budeš písať, len by som to chcel vidieť možno trochu pomenej, podrobnejšie, systematickejšie prebrať. . . Na vacsom priestore, možno postupne ten príklad budovať. . . Vysvetliť na nom všetky základné možnosti 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 použitý 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 ontológií (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 .
```

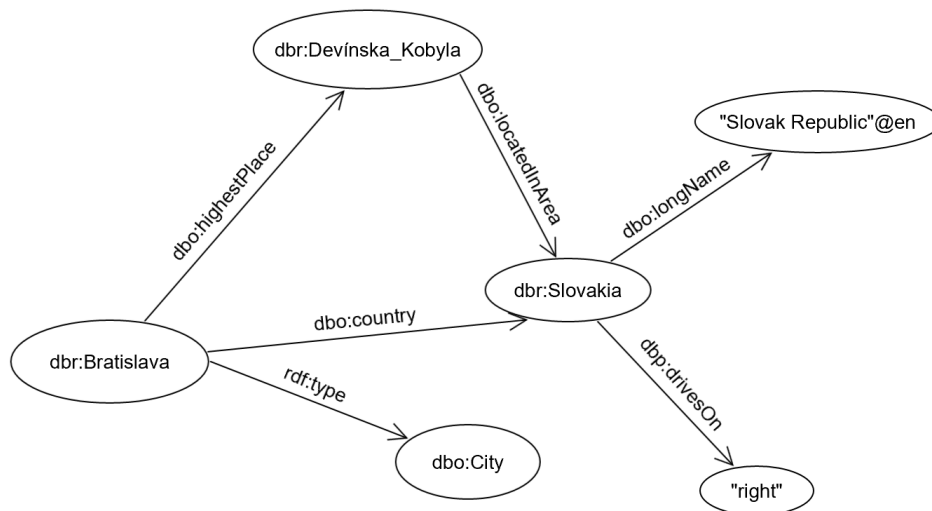
```
dbr:Bratislava rdf:type dbo:City .
```

```
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 podrobnejšie... Mozes vychadzat z mojej prednasky ale tiez napr z uvodnej kapitoly *Handbook on Ontologies*

MH: Pozn. k strukture prace: Bolo by lepsie, keby *Prehľad problematiky* bola cast \part{}, *Semantic web* samostatna kapitola \chapter{}, no a potom si myslim, ze *Ontologie* by tiez mali mat vlastnu kapitolu, ktora bude nasledovat hned po SW. Treťou kapitolou by potom mohol byt prehľad ontologii z oblasti bezpecnosti, co by uz do prehľadu 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.



### 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 nadtrieby a môže byť rozšírená o vlastné vlastnosti.

**Entity** sú individuálne inštancie nejakej nami zadanovej 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 *Jablko* *individuálnou inštanciou triedy Ovocie*, tak je aj *individuálnou inštanciou triedy Jedlo*.

**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.

**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.

Ontológia má veľa vlastností, ktoré musia byť dodržané. Každý prvok musí byť jasne identifikovateľný. 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.

## 3.2 Využitie ontológií

Ontológie sa začali využívať najmä v organizáciách, ktoré sa špecializovali 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 primárne 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.

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 zobrazíť preto, lebo výsledkom takéhoto panelu je vyhľadávanie informácií na webovej stránke, ktorá obsahuje sémantické dáta.

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

## 3.3 Ciele ontológie

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

Taktiež je jej cieľom opätovné použitie ontológie, napríklad ak máme dobre zadanú ontológiu, môžu ontologickí 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.

## **3.4 Syntax ontológií**

**MAM SPRAVIT AJ TOTO?**

# 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

Unified Cybersecurity Ontology alebo skrátené 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 bezpečnostnej domény ale taktiež poskytuje iné relevantné štandardy ako CVE4, CCE5, CVSS6, CAPEC7, CYBOX8, KillChain9 a STUCCO10.

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.

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 v rôznych doménach záujmu.

Základnými triedami, využívanými v UCO sú:

- *Means* – Čo je zamýšľané daným útokom.
- *Consequences* – Dôsledky útoku.
- *Attack* – Typ útoku.
- *Attacker* – Kto je iniciátorom 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 pomocou 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 Brynielson, 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<sup>+</sup>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

2.1	Semantic Web - vrstvy. Zdroj: [?] . . . . .	4
2.2	Príklad grafovej databázy. . . . .	7