Vysoká škola ekonomická v Praze
Fakulta informatiky a statistiky
Katedra informačního a znalostního inženýrství

Studijní program: Aplikovaná informatika
Obor: Aplikovaná Informatika

Analýza využívanosti modelu Schema.org

BAKALÁRSKÁ PRÁCA

Študent : Ján Káva
Vedúci : doc. Ing. Vojtěch Svátek, Dr.

2018

Prehlásenie:

Prehlasujem, že som bakalársku prácu spracoval samostatne a že som uviedol všetky použité pramene a literatúru, z ktorej som čerpal.

V Prahe dňa 1. Mája 2018 . . . . . .

Ján Káva

Poďakovanie

Ďakujem každému, kto mi prejavil dôveru a podporu počas môjho štúdia a počas písania tejto práce, mojej rodine, priateľom a môjmu vedúcemu doc. Ing Vojtěchovi Svátkovi, Dr. za cenné rady a povzbudivé slová.

Abstrakt

Bakalárska práca sa zaoberá využívaním sémantického modelu schema.org. Teoretická časť pojednáva o štruktúre sémantického webu a o využití modelu schema.org spolu s dátovými formátmi Microdata, RDFa a JSON-LD. V praktickej časti je na základe dát extraktovaných z webového crawlu z novembra 2017 analyzovaná miera využitia modelu schema.org v porovnaní s inými sémantickymi modelmi ako aj využívanosť samotných schema.org typov a rozšírení v rámci dátových formátov Microdata, RDFa a JSON-LD.

Kľúčové slová

schema.org, sémantický model, RDFa, Microdata, JSON-LD, web data commons

Abstract

This bachelor thesis focuses on utilization of semantic model schema.org on the web. The theoretical part is describing structure of semantic web and implementation of schema.org within the data formats Microdata, RDFa and JSON-LD. Practical part provides analysis of utilization of schema.org in comparison to other semantic models as well as usage of schema.org's types and extensions within data formats Microdata, RDFa and JSON-LD, based on data extracted from a web crawl from November 2017.

Keywords

schema.org, semantic model, RDFa, Microdata, JSON-LD, web data commons

Obsah

Obsah ii

1 Úvod 4

1.1 Vymedzenie témy 4
1.2 Dôvod výberu témy 4
1.3 Ciele práce 4
1.4 Spôsob dosiahnutia cieľa 5
1.5 Predpoklady a obmedzenia práce 5
1.6 Štruktúra práce 5
1.7 Výstupy a očakávané prínosy 6

2 Komentovaná rešerš informačných zdrojov 7

2.1 Rešerš použitých informačných zdrojov 7
2.2 Rešerš predchádzajúcich kvalifikačných prací 7

3 Sémantický web 9

3.1 Komponenty sémantického webu 9
3.1.1 Štruktúra sémantického webu 10
Syntax, identifikátor a znaková sada 10
Resource Description Framework 11
Taxonómie 11
RDF Schema 11
Ontológie, OWL 12

4 Sémantické modely 13

4.1 Model ogp.me 13
4.2 Dublin Core Metadata Initiative 14
4.3 Model data-vocabulary.org 14
4.4 Friend of a Friend 14

5 Model schema.org 15

5.1 Dátový model schema.org 15
5.2 Hostované schema.org rozšírenia 15
5.3 Schema.org a SEO 16
5.3.1 Search Engine Optimization 16
5.3.2 Využitie schema.org pri SEO 17

6 Dátové formáty a schema.org 18

6.1 RDFa 20
6.2 Microdata 22

6.3 JSON-LD 23

7 Využívanosť modelu schema.org 25

7.1 Dáta k analýze 25

7.1.1 Data Extraction Framework 26

7.1.2 Vyextraktované dáta 27

7.2 Analýza sémantických modelov 30

7.2.1 Využitie sémantických modelov vo formáte Microdata 30

7.2.2 Využitie sémantických modelov vo formáte JSON-LD 31

7.2.3 Využitie sémantických modelov vo formáte RDFa 32

7.3 Analýza schema.org 33

7.3.1 Využívanosť schema.org typov 34

7.3.2 Využívanosť schema.org rozšírení 36

7.4 Zhrnutie 37

8 Záver 39

8.1 Ciele a ich naplnenie 39

8.2 Prínosy práce 40

8.3 Motivácia pre ďalšie práce 40

Terminologický slovník 41

Použité informačné zdroje 42

Zoznam obrázkov 44

Zoznam výpisov kódov 45

Zoznam tabuliek 46

1 Úvod

V tejto kapitole budú definované základne piliere tejto bakalárskej práce, a to vymedzenie témy, dôvod výberu témy, cieľe tejto práce, spôsob ich dosiahnutia, predpoklady a obmedzenia práce, štruktúra práce, výstupy a očakávané prínosy.

1.1 Vymedzenie témy

Témou tejto bakalárskej práce je analýza využívanosti sémantického modelu schema.org. Práca bude pojednávať o teoretických princípoch sémantických modelov na webe, a hlavne teda modelu schema.org, ďalej o možnostiach spracovávania dát pomocu tohto modelu v správne definovanom kontexte v rámci dátových formátov RDFa, Microdata a JSON-LD.

Ďalej sa bude práca zaoberať analýzou využívanosti modelu schema.org na webe v rámci vyššie spomínaných dátových formátov a jeho rozšírení a typov.

1.2 Dôvod výberu témy

Oblasť sémantického webu je jedným z hlavných trendov v smerovaní vývoja webu. V súčasnej dobe sa veľké spoločnosti orientujú na vývoj inteligentných technológií ako sú samojazdiace autá, inteligentné montážne linky, Internet of Things a pod. všetky tieto inteligentné technológie ale k svojmu fungovaniu potrebujú vhodne štruktúrované dáta.

Myšlienka Webu 3.0, teda webu, kde všetky informácie majú svoj význam, ktorému rozumejú aj stroje, bola prvýkrát spomenutá už v roku 2006 v blogu Jeffreyeho Zeldmana (Zeldman, 2006), a odvtedy sa začalo hovoriť o databázovej transformácií webu, kde sú všetky dáta uložené na webe vhodne reprezentované v danom kontexte a kde takýmto dátam rozumejú aj stroje nielen ľudia. Táto myšlienka je skvelá a môže pomôcť ľudstvu k lepšiemu využitiu strojov k stereotypným činnostiam, ktoré ale vyžadujú presné a správne spracovanie informácií.

1.3 Ciele práce

Cieľom teoretickej časti práce je popísať základne princípy sémantického webu, sémantických modelov, dátových formátov vhodných pre spracovanie dát na sémantickom webe a vzťahy medzi nimi, a týmto vytvoriť vhodný základ pre riešenie praktickej časti.

Cieľom praktickej časti je analýza využívanosti sémantického modelu schema.org na webe. V práci bude analyzované ako je schema.org využívaná na webe v rámci dátových formátov RDFa, Microdata a JSON-LD vo vzťahu k iným sémantickým modelom, ale aj to ako sú v schema.org využívané jej rozšírenia a typy.

1.4 Spôsob dosiahnutia cieľa

Pre ciele definované v predchádzajúcej podkapitole je potrebné stanoviť metódy a postupy a ich dosiahnutie. Aby mohli byť splnené ciele teoretickej časti, je potrebné naštudovať si princípy sémantického webu, sémantických modelov a spôsob zápisu štruktúrovaných dát vo využití s modelom schema.org.

Pre dosiahnutie cieľov praktickej časti je nutné získať vhodné dáta z webcrawlingových projektov pre analýzu, porozumieť im, vhodne ich spracovať a prezentovať.

1.5 Predpoklady a obmedzenia práce

Pri každej práci takéhoto rozsahu sa dá predpokladať, že sa riešiteľ stretne s viacerými obmedzeniami. Pre mňa je najväčším obmedzením to, že pred písaním práce nemám žiadne výrazne skúsenosti v oblasti sémantickeho webu a preto je v mojom prípade o to väčšia potreba kvalitne naštudovať túto problematiku, čo môže byť ale na druhej strane aj motiváciou.

1.6 Štruktúra práce

Práca je rozdelená do niekoľkých tématicky usporiadaných kapitol. Takto usporiadané kapitoly postupne budujú podklad pre praktickú časť práce. Rozdelenie práce do kapitol je následovné:

1. Úvod – kapitola obsahuje základné body o bakalárskej práci. V tejto kapitole je vymedzená téma, dôvod jej výberu, ciele práce a očakávané prínosy.
2. Komentovaná rešerše informačných zdrojov – v tejto kapitole je prehľad literatúry využitej pri vypracovávaní práce a taktiež prehľad predchádzajúcich prací, ktoré riešia podobnú problematiku.
3. Sémantický web – v tejto kapitole sú rozobraté hlavné princípy a štruktúra sémantickeho webu.

4. Sémantické modely – v tejto kapitole sú popísané najčastejšie používané sémantické modely na sémantickom webe.
5. Model schema.org – v tejto kapitole je popísaný sémanticky model schema.org, jeho rozšírenia, dátový model a jeho využitie pri SEO.
6. Dátové formáty a schema.org – v tejto kapitole je popísané a demonštrované ako zapísať informáciu v štruktúrovanom strojovo čitateľnom formáte pomocou formátov RDFa, Microdata a JSON-LD.
7. Využívanosť modelu schema.org – v tejto kapitole je analyzovaná miera využívanosti modelu schema.org naprieč formátmi Microdata, RDFa a JSON-LD.
8. Záver – v tejto poslednej kapitole je zhodnotené naplnenie cieľov, prínosy tejto práce a motivácia pre ďalšie práce.

1.7 Výstupy a očakávané prínosy

Hlavným cieľom tejto práce je analyzovať mieru využívanosti modelu schema.org. Výstupom je teda zistená miera využívanosti tohto modelu na sémantickom webe v porovnaní s inými bežne používanými sémantickými modelmi. Model schema.org vznikol v roku 2011 a preto môže byť zaujímavé vidieť ako je v súčasnej dobe využívaný. Teoretická časť slúži ako úvod do sémantického spracovania informácií na webe, sémantických modelov, ktoré tieto informácie popisujú a dátových formátov, v ktorých sú tieto dáta uložené.

2 Komentovaná rešerš informačných zdrojov

Táto kapitola obsahuje rešerš predchádzajúcich kvalifikačných prací aby táto práca neriešila problématiku, ktorá už bola spracovaná v inej práci a taktiež rešerš použitých informačných zdrojov.

2.1 Rešerš použitých informačných zdrojov

Pri písaní teoretickej časti som najviac čerpal z knihy Leslieho F. Sikosa „*Mastering Structured Data on the Semantic Web*“ (Sikos, 2015) a taktiež z kapitoly *2 Ontologie a sémanticky web* z knihy „*Umělá inteligence (6)*“, ktorú napísali M. Obitko, O. Zamazal a V. Svátek (M. Obitko, O. Zamazal a V. Svátek, 2013) a takisto z pracovných listov vydávaných konzorciom W3C. V knihe Leslieho Sikosa sú veľmi pekne a zrozumiteľne vysvetlené, okrem iného aj náležitosti sémantického webu a reprezentácia štruktúrovaných znalostí na internete. V spomínanej kapitole knihy *„Umělá inteligence (6)“* autori jasne a zrozumiteľne vysvetľujú princíp sémantickeho webu tak, že dôkladne rozoberajú každú jeho komponentu. V pracovných listoch konzorcia W3C sú vecne a stručne popísané všetky náležitosti štruktúrovaných dát, ktoré sú pre túto prácu skvele využiteľné. Ďalej pri praktickej časti som hlavne čerpal informácie z webu projektu Web Data Commons a Common Crawl, kde som z Web Data Commons (Web Data Commons, 2018) získal štatistiky o využívaní štruktúrovaných dát vo formátoch RDFa, Microdata a JSON-LD.

2.2 Rešerš predchádzajúcich kvalifikačných prací

Rešerš prací týkajúcich sa sémantického modelu schema.org a analýzy dát na sémantickom webe vypracovaných na českých vysokých školách:

M. KAVALEC, *Ontology Learning and Information Extraction for the Semantic Web*, Dizertačná práca, 2006.

- Táto práca je rozdelená do troch tématických celkov a to sémantický web, extrakcia informácií a učenie ontológií. Autor v práci rozoberá možnosti extrakcie informácií zo sémantického webu a učenie ontológií pre kvalitnejšiu extrakciu dát.

J. PAŘENICA, *Sémantický web a propojená data*, Bakalárska práca, 2013.

- Táto práca pojednáva o technológiách a využitiu sémantického webu a vytváraním prepojených dát na internete. V práci sú popísané vrstvy, ktoré tvoria sémantický web, ďalej práca obsahuje metodický popis tvorby prepojených dát a následne na praktickom príklade ukázaný metodický popis.

T. LANG, *Sémantický web*, Diplomová práca, 2011.

— Autor vo svojej práci analyzuje vývoj sémantického webu, od počiatkov WWW a hypertextových dokumentov až po víziu Webu 3.0. Ďalej autor v práci vytvára projekt sémantického webu vo WordPresse pomocou pluginu WP-RDFa.

M. DRLÍK, *Mikroformáty pro sémanticky web*, Bakalárska práca, 2011.

— V práci sú rozoberané mikroformáty, význam mikroformátov pri tvorbe sémantického webu, nástroje pre prácu s mikroformátmi a taktiež význam mikroformátov u SEO.

L. KOŘISTKOVÁ, *Sémantický web a mikroformáty*, Bakalárska práca, 2012.

— V práci sú popísané základné informácie o sémantickom webe a jeho využití v praxi a taktiež sú definované mikroformáty, ich druhy a nevýhody.

T. HANZAL, *Modeling Events on the Semantic Web*, Diplomová práca, 2015.

— V tejto práci sa autor venuje tomu ako čo najlepšie modelovať události na sémantickom webe, rozoberá udalosti z filozofického hľadiska a popisuje ontológie pre modelovanie událostí.

M. BRABEC, *Microdata v HTML5*, Bakalárska práca, 2014.

— Autor vo svojej práci popisuje HTML5 Microdata a ich využitie na sémantickom webe. Autor taktiež popisuje využitie Microdata so sémantickým modelom schema.org.

J. HANUŠ, *Porovnání přístupu k ukládání otevřených propojených dat*, Diplomová práca, 2017.

— Táto práca rozoberá problématiku ukládania otvorených dát. V práci je okrem iného popísaná aj RDF databáza a jazyk pre dotazovanie sa do tejto databáze SPARQL.

3 Sémantický web

Súčasný web sa dá chápať ako veľká databáza dokumentov prepojených pomocou odkazov. Pomocou bežne používaných webových technológií môžu stroje prenášať a prezentovať užívateľom informácie reprezentované najčastejšie formou HTML dokumentov. Problém ale nastáva v momente keď chceme aby nám počítač zobrazil dáta na základe človekom špecifikovaných požiadavkov, podmienok a obmedzení. Človek je totižto pri čítaní dát prezentovaných v danom dokumente schopný vnímať ich v kontexte a správne ich pochopiť, pre počítač sú však tieto dáta napríklad len série textových reťazcov.

Sémantický web, je rozšírením súčasného webu, tak aby informácie v ňom prezentované boli strojovo čitateľné, a teda aby boli stroje schopné lepšie spracovávať úlohy zadané človekom kvalitnejšie a samostatnejšie.

Definícia sémantického webu podľa konzorcia W3C: „Sémantickým webom sa rozumie reprezentácia informácií na webe. Je to spoločné úsilie vedené W3C s veľkým počtom účastníkov z výskumných a priemyslových partnerov. Sémantický web je založený na Resource Description Framework (RDF), ktorý integruje rôzne aplikácie za použitia XML syntaxe a URI pre pomenovávanie zdrojov.“ (W3C, 2017)

Ešte je vhodné upozorniť na dva rozdielne výklady slova sémantika na webe resp. sémantický web, konflikt týchto významov sa môže objavovať hlavne v anglických textoch, na ktorý upozorňuje Leslie F. Sikos (Sikos, 2015) kde sa slovom sémantika môže okrem samotného významu informácií a znalostí na webe rozumieť aj sémantika značkovacích elementov. Jedná sa o vhodné a správne využívanie elementov, s ktorými prišla HTML5, pri tvorbe webov. Napríklad využitie elementu `section` namiesto nesémantického `div` a pod.

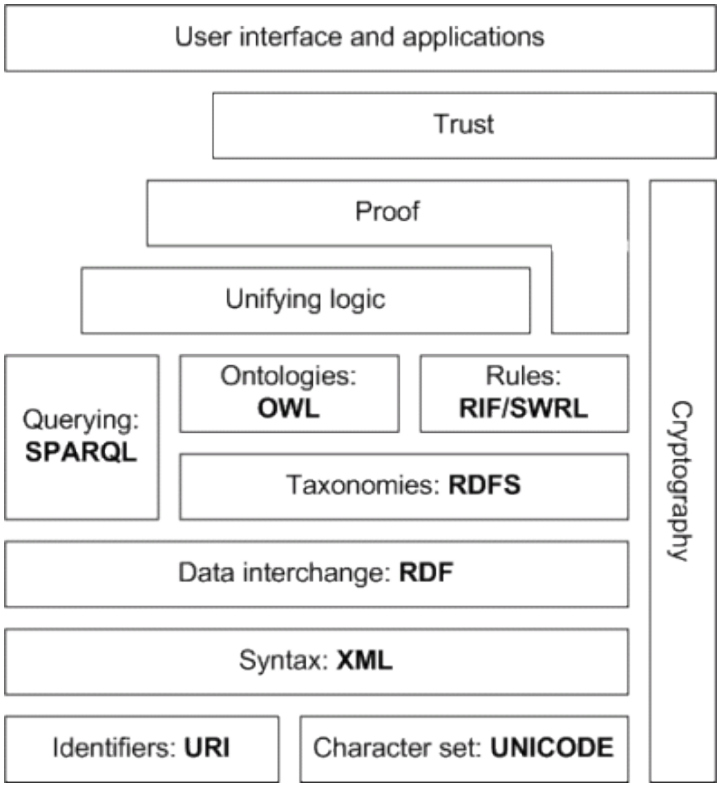
3.1 Komponenty sémantického webu

Web ako taký je postavený na princípe zdrojov, to znamená, že všetko čomu môžeme priradiť vlastnú indetitu považujeme za zdroj. Teda zdroj konceptuálne vzniká v momente, keď mu je priradená vlastná identita. Ideálnym príkladom je HTML dokument uložený na webovom serveri, ktorému je priradená jeho vlastná URL. (M. Obitko, O. Zamazal a V. Svátek, 2013)

Je zrejmé, že sémantický web ako taký sa skladá z viacerých vrstiev, kde je definované, čo a ako slúži k identifikovaniu zdrojov na sémantickom webe, ako sú informácie ukládané a prezentované a ako sú prezentované a zapisované vzťahy medzi nimi.

3.1.1 Štruktúra sémantického webu

Ako už bolo spomenuté vyššie, aby všetky komponenty sémantického webu správne a efektívne fungovali, je nutná istá hierarchia a interoperabilita vo viacerých vrstvách. Následujúca schéma znázorňuje architektúru vrstiev sémantickeho webu.



Obrázok 1:
Architektúra sémantického webu (zdroj: M. Obitko, O. Zamazal a V. Svátek, 2013)

Táto schéma znázorňuje ako je sémantický web vystavaný od použitia jednoznačných identifikátorov v znakovej sade UNICODE pomocou XML syntaxe až po samotný výstup a prezentáciu dát v uživateľskom rozhraní, teda vo výslednej aplikácií.

Pre potreby tejto práce a pochopeniu sémantického modelu schema.org je vhodné sa na niektoré z týchto komponentov pozrieť bližšie.

Syntax, identifikátor a znaková sada

Ako znaková sada sa pre identifikáciu zdrojov používa UNICODE, to umožňuje používať texty obsahujúce väčšinu znakov, ktoré v súčasnosti existujú.

Pre jednoznačnú identifikáciu zdroja sa používa URI (Uniform Resource Identifier), ktorý má jasne definovanú štruktúru a presne a jednoznačne identifikuje zdroj na webe. Konečná štruktúra URI je definovaná v RFC 3986 z januára 2005 (Network Working Group, 2005) a vyzerá následovne:

```
URI         = scheme ":" hier-part [ "?" query ] [ "#" fragment ]
```

Syntaxou pre popis dokumentov definovaných pomocou URI v znakovej sade UNICODE je značkovací jazyk XML, ale je možné používať aj inú syntax ako je napríklad HTML.

Resource Description Framework

Máme identifikovaný a syntakticky vhodne popísaný zdroj. V ďalšej vrstve sémantického webu teda prichádza na rad RDF (Resource Description Framework), ktorý taký zdroj ďalej obohacuje sémanticky.

RDF je rodina špecifikácií vypracovaná konzorciom W3C a pôvodne zamýšľaná ako model metadát. Popisuje zdroje tak, že tento popis je zrozumiteľný ako ľuďom tak aj strojom. Princípom je priradiť zdroju výraz v tvare subjekt – predikát – objekt, takéto priradenie sa nazýva RDF triplet (RDF Working Group, 2014). Všetky elementy tripletu sú taktiež identifikované webové zdroje a teda môžme tvrdiť, že triplet je pomenované spojenie dvoch zdrojov (M. Obitko, O. Zamazal a V. Svátek, 2013). Objekt ale môže tvoriť výnimku a môže byť reprezentovaný aj literálom alebo prázdnym uzlom. V prípade literálu chceme subjektu priradiť nejakú explicitnú hodnotu napr. „Ján Káva“ – má priezvisko – „Káva“, v prípade prázdneho uzlu chceme vyjadriť len to, že subjekt má nejakú vlastnosť napr. „Ján Káva“ – má priezvisko – *blank node*, tento triplet teda vyjadruje, že Ján Káva priezvisko má ale nešpecifikuje ho.

Taxonómie

Taxonómia je vedný odbor, ktorý sa zaoberá systematickým triedením a popisom v rámci danej problematiky (Peterka, 2015).

V informatike sa pojmom taxonómia rozumie hierarchia (strom) pojmov. Väčšinou popisuje len jeden druh vzťahov (napríklad vzťah „je podtriedou“), ale môže v nej implicitne existovať viacero druhov vzťahov, tak ako to je napríklad pri adresároch v súborových systémoch. Napríklad podadresár P (nemecký ovčiak) môže byť v nadradenom adresári N (pes), pretože P patrí do N. Takisto iný podadresár Q (hlava) môže byť v nadradenom N, pretože Q je časťou P.

RDF Schema

RDF Schema (Resource Description Framework Schema) je rozšírením RDF. Poskytuje mechanizmy pre popis skupín súvisiacich zdrojov a vzťahov medzi nimi.

Ontológie, OWL

Pojem ontológia (z gréckeho *on, ontos* = súci a *logos* = slovo, reč) vo filozofii pomenúva náuku o bytí (Stanford Encyclopedia of Philosophy, 2004).

V informatike sa pojem ontológia využíva v súvislosti so sémantikou webu a organizáciou poznania. Ontológie na webe sú chápané ako definície pojmov a vzťahov medzi nimi. Popisuje tzv. domény ľudského záujmu, takáto doména potom obsahuje jednotlivé triedy, ktoré sú prepojené vzťahmi. Objekty v tejto doméne a ich prepojenie ontológia popisuje pomocou štyroch prvkov a to jedinec, trieda, atribút a väzba. Niekedy sa uvádza aj piaty prvok a to udalosť.

Cieľom ontológie je teda definovať spoločné, jednotné chápanie určitej triedy pojmov. Ontológiou dokážeme znázorniť komplexnosť vzťahov medzi znalosťami. Ontológia predstavuje sieť hieriarchií, v ktorých informácie sú alebo môžu byť vzájomne prepojené logickými vzťahmi, tieto vzťahy sú založené na vlastnostiach, ktoré musia byť informáciám špecificky priradené. Aby počítače boli schopné takýmto slovníkom hierarchií rozumieť musia byť ontológie zapísané v strojovo zrozumiteľnom jazyku ako je OWL.

OWL (Web Ontology Lnaguage) je jazyk pre bohatý popis ontologií na sémantickom webe. Jedná sa teda o zápis deskripčnej logiky v RDF, s tým, že sa snaží staviať nad konštrukty RDFS (M. Obitko, O. Zamazal a V. Svátek, 2013).

4 Sémantické modely

V tejto práci bude často používaný pojem sémanticky model, preto je vhodné vysvetliť ako je tento pojem vnímaný v informatike a v akom kontexte je používaný v tejto práci. Pojem model v informatike chápeme ako abstrakciu časti reality, to znamená, že sa pomocou vhodnej a zrozumiteľnej sady symbolov snažíme zachytiť časť reality. Keďže sa na sémantickom webe snažíme vhodne zachytiť informácie resp. dáta tak aby im porozumel aj stroj, používame dátovy model, ktorý taktiež obsahuje aj to, že vysvetľuje význam svojich entít, a teda sémantický dátovy model.

V kontexte tejto práce teda môžme chápať sémanticky model ako synonymum k pojmu slovník, resp. sémantický slovník, kde je zachytená aj hierarchia a vzťahy medzi obsiahnutými pojmami.

Pre potreby tejto práce budú ďalej v krátkosti popísané následujúce modely aby bolo zrejmé ako sa odlišujú od modelu schema.org, ktorý je popísaný v kapitole 5, a poprípade, či majú špecifické využitie.

- ogp.me
- Dublin Core Metada Initiative
- data-vocabulary.org
- Friend of a Friend

4.1 Model ogp.me

Open Graph protocol (OGP) umožňuje webovej stránke to, aby sa z nej stal obohatený objekt v sociálnom grafe a určil postavenie webovej stránky v takomto grafe. Sociálny graf je graf, ktorý zobrazuje vzťahy medzi užívateľmi internetu, teda reprezentuje sociálnu sieť. V skratke sa jedná o globalné mapovanie každého a ich vzťahov navzájom.

OGP vyžaduje štyri základné vlastnosti a to:

`title` – titulok objektu, ktorý sa v grafe zobrazí

`type` – typ objektu, napr. person

`image` – URL obrázku, ktorý reprezentuje daný objekt

`url` – URL objektu, ktorá je použitá ako permanentné ID

Na základe typu objektu, môžu byť ďalej vyžadované ďalšie vlastnosti ako `audio`, `description`, `determiner`, `locale` a pod. (The Open Graph protocol, 2017).

Hlavnou úlohou modelu OGP je teda premena obsahu danej stránky na sociálny objekt, ktorý potom môže byť využitý v rámci sociálneho grafu. Na druhú stranu úlohou schema.org je obohacovanie a optimalizácia výsledkov vyhľadávania.

4.2 Dublin Core Metadata Initiative

Model Dublin Core Metadata Initiative vznikol v polovici 90. rokov minulého storočia za účelom popisovania elektronických zdrojov. V súčasnoti sa tento model rozšíril do podoby, kde je pomocou neho možné popísať hociaký zdroj, ktorý vyhovuje vlastnostiam z tohto modelu. (MediaWiki, 2011)

Model Dublin Core Metadata Initiative je hostovaný na doméne purl.org a definície jeho pojmou sú dostupné pomocou URL:

```
http://purl.org/dc/terms/
```

Tento model pre definíciu svojich pojmov využíva Persistent URL (PURL), ktorá poskytuje permanentnú adresu pre zdroj uložený na webe. Výhodou tohto je, že zdroje na webe sa môžu s plynutím času meniť, napríklad presun zdroja z jedného webového serveru (napr. example.org/zdroje/zdroj) na iný (napr. mynewexample.org/zdroj), to môže spôsobiť to, že zdroj nebude dostupný z pôvodnej URL. V tomto prípade s využitím PURL, kde by bol pôvodný zdroj odkazovaný z purl.org/jkava/zdroje stačí len zmeniť odkaz v PURLe, ktorý bude odkazovať na nový zdroj (PURL, 2017).

Tento model má hlavne využitie pri popisovaní fyzických dokumentov ako sú knihy, vedecké práce, videá a podobne, kdežto hlavný prínos schema.org je zlepšovanie SEO.

4.3 Model data-vocabulary.org

Model data-vocabulary.org je predchodcom modelu schema.org a od júna 2011 je nepodporovaný a ďalej nerozvíjaný a kompletne nahradený modelom schema.org. Každopádne na webe stále nájdeme veľa odkazov aj na tento model.

4.4 Friend of a Friend

Friend of a Friend resp. FOAF je projekt zameraný na prepájanie ľudí a informácií na webe. Hlavným cieľom modelu FOAF je teda popisovať charakteristiky ľudí, sociálnych skupín a vzťahov medzi nimi tak, aby bol tento popis nezávislý na čase a technológiách.

5 Model schema.org

Schema.org je kolaboratívny komunitný projekt, ktorého úlohou je vytvoriť, udržiavať a propagovať schémy pre štrukturované dáta na internete. Tento projekt bol založený spoločnosťami Google, Microsoft, Yahoo a Yandex (tieto spoločnosti prevádzkujú jedny z najväčších vyhľadávacích nástrojov na internete), avšak do projektu prispievajú aj ľudia z webových komunít (Schema.org Community Group, 2015).

Princípom schema.org je využívať model termínov v nej definovaných s využitím formátov RDFa, Microdata a JSON-LD pre označovanie obsahu webových stránok metadátovými informáciami. Takto označený obsah je strojovo čitateľný,a teda môžme hovoriť o sémantickom webe.

5.1 Dátový model schema.org

Dátový model schema.org je veľmi generický a odvodený od RDF schémy. Obsahuje (Schema.org Community Group, 2015):

- Sadu typov, usporiadaných do niekoľkonásobnej dedičnej hierarchie, kde každý typ môže byť podtriedov viacerých typov
- Sadu vlastností:
    - Každá vlastnosť môže mať jeden alebo viac typov ako svoju doménu. Taká vlastnosť môže byť použitá pre inštancie hociktorého z takých typov.
    - Každá vlastnosť môže mať jeden alebo viac typov ako svoj rozsah. Hodnoty tejto vlastnosti by mali byť inštancie aspoň jedného z takých typov.

Účelom typovej hierarchie schema.org nie je reprezentovať globálnu ontológiu sveta, pri jej vzniku v roku 2011 bolo zámerom usporiadať typy entít, pre ktoré mali zakladajúce spoločnosti využitie v ich vyhľadávacích nástrojoch. Počas doby existencie schema.org sa s prispením webovej komunity pomocou rozšírení výrazne zväčšil jej rozsah, avšak stále jej zámerom nie je to aby bola univerzálnou ontológiou.

5.2 Hostované schema.org rozšírenia

Projekt schema.org taktiež spravuje a vyhodnocuje rozšírenia pôvodného sémantického modelu. Každé takéto rozšírenie dostáva kúsok z menného priestoru schema.org, napríklad rozšírenie *auto* dostane menný priestor *auto.schema.org*. Termíny z rozšírenia môžu byť použité spolu s pôvodnými termínmi vo vzájomnej kompatibilite.

Každé rozšírenie je prepojené s pôvodným schema.org modelom pomocou uzlov, ktoré sú vzájomne vo vzťahu, napríklad rozšírenie Legal vytvára termín LegalPerson, ktorý je podtriedou pojmu Person z pôvodného modelu, Person je teda prepojená s LegalPerson.

Medzi súčasné schválené oficiálne rozšírenia patria:

auto.schema.org – Toto rozšírenie obohacuje základnú schema.org o 4 typy, 20 vlastností a 3 výčtové hodnoty špecifické pre automobilový priemysel. Napríklad typ Vehicle je z tohto rozšírenia obohatený o podtypy BusOrCoach, Motorcycle a MotorizedBicycle.

bib.schema.org – Toto rozšírenie obohacuje základnú schema.org o 11 typov, 17 vlastností a 1 výčtovú hodnotu, ktoré majú využitie pri spracovávaní bibliografických údajov. Napríklad typ CreativeWork je z tohto rozšírenia obohatený o podtyp ComicStory.

health-lifesci.schema.org – Toto rozšírenie je najrobustnejšie a obohacuje základnú schema.org o 99 typov, 178 vlastností a 149 výčtových hodnôt. Napríklad typ Thing je z tohto rozšírenia obohatený o podtyp MedicalThing, ktorý má mnoho ďalších podtypov z tohto rozšírenia.

iot.schema.org – Toto rozšírenie je momentálne v štádiu diskusie o tom ako ho vlastne zahrnúť do modelu schema.org.

meta.schema.org – Účelom tohto rozšírenia je podorovať implementáciu schema.org samotnej.

5.3 Schema.org a SEO

5.3.1 Search Engine Optimization

Search Engine Optimization (skratka SEO) je v informatike pojem pre označenie metodiky vytvárania a upravovania webových stránok takým spôsobom, aby ich forma a obsah boli vhodné pre automatizované spracovanie v internetových vyhľadávačoch. Cieľom SEO je získať vo výsledkoch vyhľadávania lepšiu pozíciu (odkaz na stránky bude zobrazený medzi prvými), čo môže pomôcť pri zlepšení návštevnosti danej stránky. Hodnotenie relevantnosti webových stránok vykonáva počítač, preto je hodnotenie založené na nepriamych znakoch (štruktúra, metadáta, počty odkazov a podobne). Čiastkové strategie SEO zahrňujú okrem zlepšenia postavenia webových stránok vo výsledkoch vyhľadávania taktiež odstránenie technických nedostatkov brániacich v prehliadaní webovej stránky vyhľadávacím robotom, crawlerom, užívateľom internetu a zvyšovanie koverzného pomeru stránok. Konverzný pomer stránok je pravdepodobnosť, že sa z návštevníka stránky stane zákazník, a teda že návštevník na stránke vykoná akciu, ktorá je prevádzkovateľom pôvodne zamýšľaná.

Je nutné poznamenať, že SEO ovplyvňuje len výsledky organického vyhľadávania, platené výsledky predložené napríklad službou Google AdWords budú vždy prvé vo výsledkoch vyhľadávania.

5.3.2 Využitie schema.org pri SEO

Keď webové vyhliadávače prechádzajú obsah webovej stránky, ich primárnym cieľom je zistiť o čom vlastne daná webová stránka je. Avšak ako je zrejmé vyhliadávače (stroje vo všeobecnosti) nevedia obsah danej webovej stránky spracovať tak ako človek, namiesto toho spracúvajú HTML danej stránky aby rozlíšili náležitosti ako názov stránky, obrázky, havný obsah, dátum publikovania, autora a mnohé ďalšie náležitosti, ktoré im počas procesu indexovania prídu užitočné.

Aby bolo prechádzanie webových stránok vyhliadavačmi jednoduchšie a rýchlejšie, spoločnosti prevádzkujúce jedny z najpoužívanejších vyhľadávacích nástrojov sa zhodli na spoločných štandardoch pre popisovanie obsahu danej stránky v strojovo zrozumiteľnej forme.

Prítomnosť štruktúrovaných dát zlepšuje informatívnosť a atraktívnosť výsledku vyhľadávania stránky, ktorá ich využíva. Ďalšou výhodou prítomnosti štruktúrovaných dát je možnosť zobrazovať relevantný obsah v znalostnom grafe spoločnosti Google (z ang. Google knowledge graph). Pomocou štruktúrovaných dát je taktiež možné poskytnúť vyhľadávacím nástrojom viac informácií o miestnom podnikaní, teda adresu a kontaktné informácie, čo pomáha hlavne pri hľadaní výrazov spojených s lokáciou, napríklad reštaurácie, športoviská a pod. Taktiež sú na štruktúrovaných dátach postavené nové technológie ako akcelerované mobilné stránky (z ang. accerelated mobile pages alebo AMP), ktorá umožňuje vytvárať webové stránky a reklamy na webe, ktoré sú rýchle, uživateľsky priateľské a výkonné naprieč všetkými platformami (AMP, 2017). Ďalej pri použití štruktúrovaných dát je väčšia šanca, že sa daná stránka zobrazí ako Google odpoveď na hľadaný výraz (Chris, 2016).

6 Dátové formáty a schema.org

V predchádzajúcich kapitolách bola popísaná štruktúra sémantickeho webu, sémantickych modelov a hlavne teda modelu schema.org. Hlavnou myšlienkou sémantickeho webu je teda spracovanie a reprezentácia informácií a znalostí tak, aby im porozumel nielen človek ale aj stroj a tým pádom bol stroj schopný automatizovať vybrané úlohy s využitím týchto znalostí. Tieto informácie sú potom odkazované na sémanticky model, v ktorom sú jasne definované ich vlastnosti.

Aby dáta uložené na webe boli strojovo čitateľné je nutné aby boli uložené vo vhodnej štruktúre, ktorá je ale aj zrozumiteľná pre človeka, a hlavne teda pre tvorcov obsahu na sémantickom webe.

Na základe dát z extraktu sú na sémantickom webe najviac využíváné dátové formáty RDFa, Microdata a JSON-LD. Výrazne je takisto aj využívaný formát vCard, ktorý je ale špecificky zameraný na zápis vizitiek v strojovo čitateľej forme, a preto nie je tak univerzálny ako ostatné spomínané formáty. Formáty RDFa a Microdata rozšírujú klasickú HTML štruktúru (v prípade RDFa aj XML štruktúru) o sadu rozšírení a atribútov, ktoré sú vhodné pre reprezentáciu informácií na sémantickom webe. JSON-LD, ako už názov napovedá, rozširuje formát JSON o možnosť zápisu štruktúrovaných, a teda strojovo spracovateľných dát.

Majme informácie o filme Good Will Hunting z česko-slovenskej filmovej databáze zaznamenané v následujúcej tabuľke:

Názov	Good Will Hunting
Žáner	Drama / Psychologický
Režisér	Gus Van Sant
Hrajú	Robin Williams, Matt Damon, Ben Affleck
Popis	Psychologická dráma, zobrazujúca osudy dvadsaťročného matematického génia, ale zároveň aj krčmového povaľača Willa Huntinga. Ten pochádza zo skromných pomerov a svoj život strávil prevažne v sirotčincoch. Mladíkove neuveriteľné matematické nadanie vyjde najavo, keď ako upratovač na bostonskej univerzite náhodne vyrieši zložitý matematický príklad, ktorý zostal na tabuli. Citovo vyprahlý a zablokovaný Will však odmieta využiť svoj obrovský talent a zmeniť svoj život. Nakoniec sa však podvolí tlaku okolia.

Tabuľka 1: *Tabuľka reprezentujúca záznam o filme Good Will Hunting*

Ak by sme tieto informácie chceli zapísať do HTML dokumentu, tento zápis by vyzeral následovne:

```
<div>
    <h1>Good Will Hunting</h1>
    <p>Žáner: Dráma / Psychologický </p>
    <p>Režisér: Gus Van Sant </p>
    <p>Hrajú: Robin Williams, Matt Damon, Ben Affleck </p>
    <p>Popis: Psychologická dráma, zobrazujúca osudy dvadsaťročného
matematického génia, ale zároveň aj krčmového povaľača Willa
        Huntinga. Ten pochádza zo skromných pomerov a svoj život strávil
prevažne v sirotčincoch. Mladíkove neuveriteľné
        matematické nadanie vyjde najavo, keď ako upratovač na bostonskej
univerzite náhodne vyrieši zložitý matematický
        príklad, ktorý zostal na tabuli. Citovo vyprahlý a zablokovaný Will však
odmieta využiť svoj obrovský talent a zmeniť
        svoj život. Nakoniec sa však podvolí tlaku okolia.</p>
</div>
```

Výpis 1: *Zápis záznamu o filme Good Will Hunting v klasickej HTML štruktúre*

Takto štruktúrovaný dokument je človeku jasný, avšak stroj by v ňom videl len div ktorý obsahuje názov a štyri paragrafy. V ďalších kapitolách bude demonštrované ako v rámci formátov RDFa, Microdata a JSON-LD, tieto informácie zapísať tak, aby aj stroj rozumel tomu, že sa jedná o film s priradeným žánrom, režisérom, hercami a popisom.

6.1 RDFa

RDFa alebo Resource Description Framework in Attributes je sémantickým rozšírením súčasneho HTML5 o atribúty, ktoré umožňujú vyjadrovať strojovo čitateľné dáta na webe.

V následujúcom kóde bude vyjadrený záznam o filme Good Will Hunting v strojovo čitateľnej forme pomocou RDFa.

```
<div vocab="http://schema.org/" typeof="Movie">
    <h1 property="name">Good Will Hunting</h1>
    Žáner:
    <span property="genre">Drama</span>
    <span property="genre">Psychologický</span>
    Režisér:
    <div property="director" typeof="Person">
        <span property="name">Gus Van Sant</span>
    </div>
    Hrajú:
    <div property="actor" typeof="Person">
        <span property="name">Robin Williams</span>
    </div>
    <div property="actor" typeof="Person">
        <span property="name">Matt Damon</span>
    </div>
    <div property="actor" typeof="Person">
        <span property="name">Ben Affleck</span>
    </div>
    Popis:
    <span property="description">Psychologická dráma, zobrazujúca osudy
dvadsaťročného matematického génia, ale zároveň aj krčmového povaľača Willa
Huntinga.
        Ten pochádza zo skromných pomerov a svoj život strávil prevažne v
sirotčincoch. Mladíkove neuveriteľné matematické
        nadanie vyjde najavo, keď ako upratovač na bostonskej univerzite náhodne
vyrieši zložitý matematický príklad, ktorý
        zostal na tabuli. Citovo vyprahlý a zablokovaný Will však odmieta využiť
svoj obrovský talent a zmenit svoj život.
        Nakoniec sa však podvolí tlaku okolia.
    </span>
</div>
```

Výpis 2: *Zápis záznamu o filme Good Will Hunting pomcou RDFa v strojovo čitateľnej forme*

Vidíme, že najvyššie v hierarchii je div, v ktorom je definovaný použitý sémanticky model, ďalej používame vlastnosti (property), ktoré pojem typu Movie môže obsahovať, a ktoré

ďalej vyjadrujeme pomocou iného typu a jeho vlastností (napr. vlastnosť actor, je typu Person, a tento typ má definované ďalšie vlastnosti) alebo pomocou literálu (napr. vlastnosť description je jednoduchý text). Takto štruktúrovaný záznam je sémanticky a strojovo čitateľný.

6.2 Microdata

Microdata veľmi podobne ako RDFa rozširujú pôvodnú HTML5 syntax o sadu atribútov ktoré hodnotám v elementoch správne označeným pridávajú sémantiku.

Demonštrácia využitia Microdata pre sémantický zápis záznamu o filme Good Will Hunting nižšie.

```
<div itemscope itemtype= "http://schema.org/Movie">
    <h1 itemprop="name">Good Will Hunting</h1>
    Žáner:
    <span itemprop="genre">Drama</span>
    <span itemprop="genre">Psychologický</span>
    Režisér:
    <div itemprop="director" itemscope itemtype="http://schema.org/Person">
        <span itemprop="name">Gus Van Sant</span>
    </div>
    Hrajú:
    <div itemprop="actor" itemscope itemtype="http://schema.org/Person">
        <span itemprop="name">Robin Williams</span>
    </div>
    <div itemprop="actor" itemscope itemtype="http://schema.org/Person">
        <span itemprop="name">Matt Damon</span>
    </div>
    <div itemprop="actor" itemscope itemtype="http://schema.org/Person">
        <span itemprop="name">Ben Affleck</span>
    </div>
    Popis:
    <span itemprop="description">Psychologická dráma, zobrazujúca osudy
dvadsaťročného matematického génia, ale zároveň aj krčmového povaľača Willa
Huntinga.
        Ten pochádza zo skromných pomerov a svoj život strávil prevažne v
sirotčincoch. Mladíkove neuveriteľné matematické
        nadanie vyjde najavo, keď ako upratovač na bostonskej univerzite náhodne
vyrieši zložitý matematický príklad, ktorý
        zostal na tabuli. Citovo vyprahlý a zablokovaný Will však odmieta využiť
svoj obrovský talent a zmenit svoj život.
        Nakoniec sa však podvolí tlaku okolia.
    </span>
</div>
```

Výpis 3: *Zápis záznamu o filme Good Will Hunting pomcou Microdata v strojovo čitateľnej forme*

Vidíme, že narozdiel od RDFa nedefinujeme slovník explicitne vhodným atribútom ale elementom pridávame typ pomocou atribútu itemtype pomocou absolútnej URL, atribútom

itemscope označujeme, že sa jedná o typovaný element a atribútom itemprop definujeme vlasnosť elementu.

6.3 JSON-LD

JSON (JavaScript Object Notation) je spôsob zápisu dát vo formáte podobnému zápisu javascriptového objektu.

JSON-LD (JSON for linking data) je rozšírením pôvodného JSONu, ktorý umožňuje interpretovať JSON ako prepojené dáta s minimálnymi zmenami v pôvodnom JSON dokumente. JSON-LD je 100% kompatibilný s JSON, takže JSON parsery a knižnice môžu byť využité aj pre JSON-LD.

Na kóde nižšie je demonštrované využitie formátu JSON-LD pre sémanticky zápis záznamu o filme Good Will Hunting.

```
{
    "@context": "http://schema.org",
    "@type": "Movie",
    "name": "Good Will Hunting",
    "genre": [
        "Drama",
        "Psychologický"
    ],
    "director": {
        "@type": "Person",
        "name": "Gus Van Sant"
    },
    "actor": [
        {
            "@type": "Person",
            "name": "Robin Williams"
        },
        {
            "@type": "Person",
            "name": "Matt Damon"
        },
        {
            "@type": "Person",
            "name": "Ben Affleck"
        }
    ],
    "description": "Psychologická dráma, zobrazujúca osudy dvadsaťročného
matematického génia, ale zároveň aj krčmového povaľača Willa Huntinga.
```

```
Ten pochádza zo skromných pomerov a svoj život strávil prevažne v
sirotčincoch.Mladíkove neuveriteľné matematické
nadanie vyjde najavo, keď ako upratovač na bostonskej univerzite náhodne vyrieši
zložitý matematický príklad, ktorý
zostal na tabuli.Citovo vyprahlý a zablokovaný Will však odmieta využiť svoj
obrovský talent a zmenit svoj život.
Nakoniec sa však podvolí tlaku okolia."
}
```

Výpis 4: *Zápis záznamu o filme Good Will Hunting pomcou JSON-LD v strojovo čitateľnej forme*

Vidíme, že najprv je definovaný slovník pomocou hodnoty @context a typ objektu pomocou @type a ďalej sú vyjadrované vlastnosti tohto objektu definovaním typu a následne vyjadrením hodnoty napr. vlastnosť Director je typu Person a tá ma hodnotu Gus Van Sant alebo je vlastnosť literálom napr. vlastnosť description.

7 Využívanosť modelu schema.org

V tejto kapitole budú analyzované dáta, ktoré boli získane pomocou webového crawleru Common Crawl a z nich ďalej extraktované dáta projektom Web Data Commons.

Projekt Common Crawl vznikol v Kalifornii ako nezisková organizácia založená Gil Elbaz, ktorého cieľom je demokratizovať prístup k informáciam na webe vytvorením a spravovaním dátového úložiska pre dáta získané pomocou svojho webového crawleru, ktoré je všeobecne dostupné a analyzovateľné (Common Crawl Foundation, 2017).

Projekt Web Data Commons vznikol na Univerzite v Mannheime, jeho autormi sú Christian Bizer, Robert Meusel a Anna Primpeli. Cieľom tohto projektu je extraktovať štruktúrované dáta z crawlov projektu Common Crawl (Web Data Commons, 2012a).

7.1 Dáta k analýze

Pre potreby analýzy využívanosti schema.org budú použité dáta, ktoré projekt Web Data Commons vyextraktoval z pôvodného dátového setu Common Crawl z novembra 2017 pomocou svojho Extraction frameworku.

Crawl z ktorého boli dáta extraktované je popísaný tabuľkou nižšie.

Dátum crawlu	November 2017
Veľkosť crawlu	66 Terabyte (zkompresované)
Parsované HTML URL	3 155 601 774
URL obsahujúce triplety	1 228 129 002
Počet domén v crawle	26 271 491
Domény obsahujúce triplety	7 422 886
Entity s definovaným typom	9 430 164 323
Triplety	38 721 044 133
Veľkosť extraktovaných dát	855 GB (zkompresované)

Tabuľka 2: *Tabuľka popisujúca crawl z ktorého boli extraktované štruktúrované dáta (zdroj: Web Data Commons, 2018)*

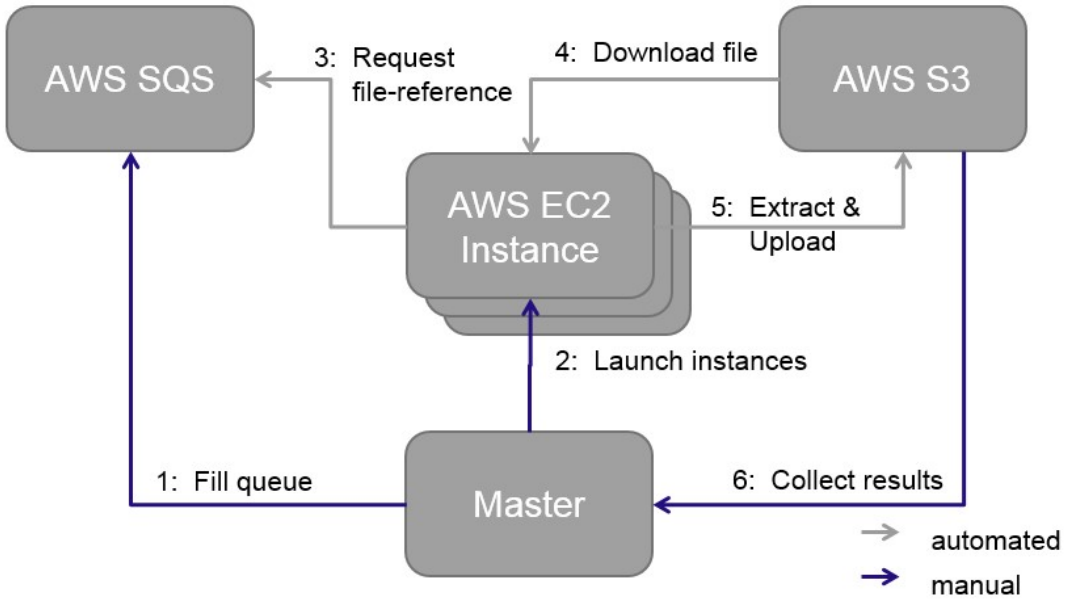
Už z tejto tabuľky môžme vyčítať zaujimavé fakty o prítomnosti RDF tripletov. V crawle bolo celkovo 3 155 milionov HTML stránok, z ktorých 1 228 miliónov obsahovalo

triplety, čo predstavuje 38,92 %. Tieto stránky sú hostované na 7,4 miliónoch doménach, z celkových 26 miliónov v crawle, čo predstavuje 28,25 %.

Projekt Web Data Commons ďalej tento extrakt spracúva a publikuje štatistiky o nasadzovaní rôznych dátových formátov ako aj sémantických modelov, ktoré sú s nimi použité.

7.1.1 Data Extraction Framework

Data Extraction Framework bolo vyvinutý v rámci projektu Web Data Commons aby umožnil spracovanie veľkého množstva súborov získavaných projektom Common Crawl s využitím Amazon cloudových služieb. Framework bol pôvodne navrhnutý Hannesom Mühleisenom pre extrakciu Microdata, Microformat a RDFa formátov z crawlu Common Crawl z roku 2012. Neskôr bol rozšírený o možnosti extrakcie hyperlinkových grafov a taktiež webových tabuliek. Obrázok nižšie znázorňuje základný princíp fungovania tohto frameworku.



Obrázok 2:
Data Extraction Framework princíp (zdroj: Web Data Commons, 2012)

1. Z master uzlu je naplnená AWS Simple Queue Service (SQS) so všetkými súbormi, ktoré majú byť spracované. Tieto súbory reprezentujú úlohu pre neskoršie pracovné inštancie a sú v podstate odkazmi na súbory.
2. Z master uzlu je spustených niekoľko inštancií EC2 v Amazon Web Services prostredí. Každá spustená inštancia si autotamticky nainštaluje framework a spustí ho.

3. Každá inštancia, po spustení frameworku, si automaticky zažiadá o úlohu z SQS a začne spracovávať daný súbor.
4. Súbor je prvotne stiahnutý z Amazon Simple Storage Service (S3) úložiska, ktorý je potom následne spracovaný EC2 inštanciou.
5. Po skončení spracovania jedného súboru sa výsledky ukladajú do užívateľovho S3 úložiska, ktoré je definované unikátnym názvom a takto uvoľnená EC2 inštancia si opäť žiadá o úlohu z SQS.
6. Po vyprázdnení celého poradovníka môže master uzol začať zbierať vyextraktované dáta a šatistiky.

Výstupom sú potom dátové sady, ktoré sú reprezentované N-Quad formou, ktorou sa rozumie sekvencia RDF termínov reprezentujúcich subjekt, predikát, objekt a označenie RDF tripletu a grafu ktorého časťou je v dátovej sade.

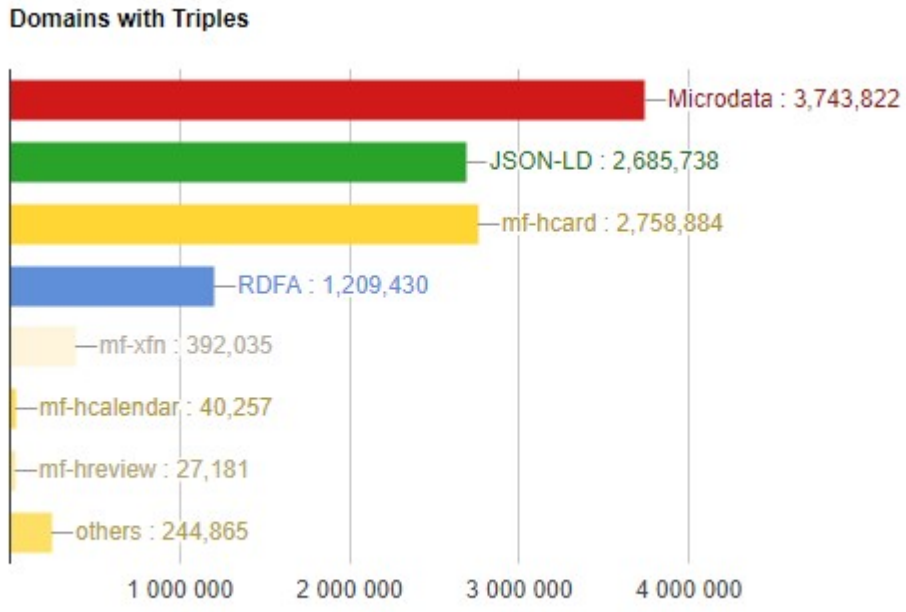
7.1.2 Vyextraktované dáta

Z výstupov Data Extraction Frameworku sú ďalej vytvárané štatistiky o prítomnosti štruktúrovaných dát vo formátoch RDFa, Microdata a JSON-LD, a práve tieto štatistiky budú použité pre zistenie využívanosti schema.org v porovnaní s inými najpoužívanejšími sémantickými modelmi.

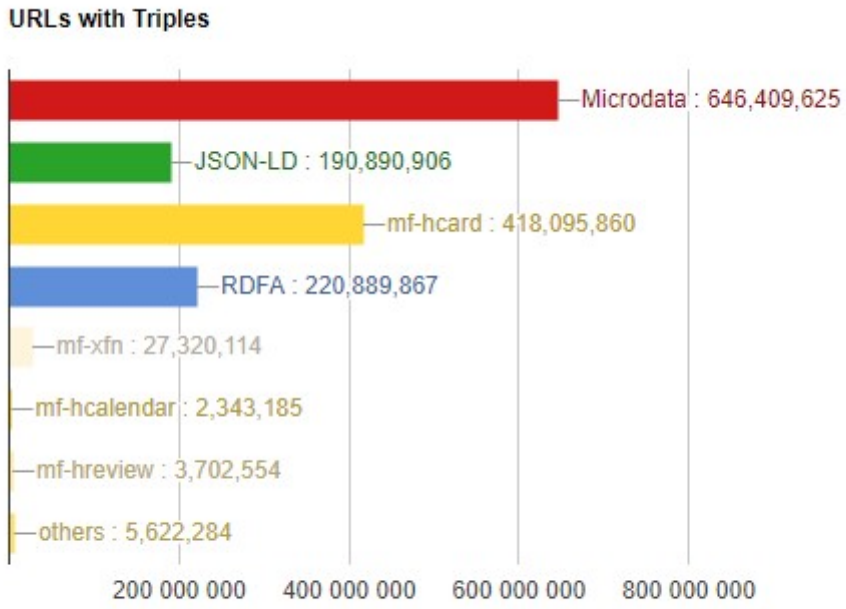
Detailné štatistiky sú verejne dostupné zo stránok projektu Web Data Commons a sú v troch variáciách podľa príslušného dátového formátu, v ktorom sú dáta reprezentované, a teda Microdata, RDFa alebo JSON-LD.

Ako už bolo spomenuté v úvode podkapitoly 7.1, štruktúrované dáta sú pritomné na 38,92 % webových stránok a na 28,25 % doménach nachádzajúcich sa v crawle.

Na následujúcich grafoch je znázornené zastúpenie dátových formátov na webových stránkach a doménach obsahujúcich štruktúrované dáta.



Obrázok 3:
Graf reprezentujúci počet domén obsahujúcich triplety naprieč dátovými formátmi (zdroj: Web Data Commons, 2018)



Obrázok 4:
Graf reprezentujúci počet URL obsahujúcich triplety naprieč dátovými formátmi (zdroj: Web Data Commons, 2018)

Z týchto dvoch grafov je zrejmé, že najvýraznejšie zastúpené sú dátové formáty Microdata, JSON-LD, hCard a RDFa.

Formát hCard je HTML mikroformát, ktorý sa využíva pre sémantický zápis vizitiek vo formáte VCard do HTML, poprípade XHTML. VCard je zase súborový formát pre výmenu osobných dát v elektornickej podobe.

Čo je na týchto dvoch grafoch ale najzaujímavejšie je to, že formát Microdata je výrazne najviac prítomný na samotných webových stránkach v porovnaní s RDFa a JSON-LD, kdežto na doménach tento rozdiel nie je až tak výrazný. To je v podstate spôsobené tým, že veľké percento web stránok je hostovaných na doménach ako WordPress, BlogSpot a pod. Podľa portálu W3 Techs tvoria WordPress stránky až 29 % všetkých webových stránok (W3Techs, 2017). V týchto dátach tvoria dve najpočetnejšie domény blogspot.com a wordpress.com, ktoré predstavujú až 17,5 % všetkých domén s nájdenými tripletmi a blogspot.com samotný predstavuje 14,88 % domén (Web Data Commons, 2018).

Ďalej keď sa pozrieme na najpopulárnejšie nástroje a pluginy pre rozšírenie WordPress stránok o štruktúrované dáta ako sú napríklad Schema Creator by Raven alebo SchemaApp môžme vidieť, že tieto pluginy pridávaju sémantiku informáciam pomocou formátu Microdata a modelu schema.org. Ďalej môžme predpokladať, že autori WordPress stránok sa budú snažiť svoje stránky optimalizovať pre vyhliadavacie nástroje a teda, že štruktúrované dáta na WordPress doménach sú prítomné vo výraznej miere. Väčšina SEO best practice blogov a tutoriálov práve odporúča využívanie formátu Microdata spolu s modelom schema.org.

7.2 Analýza sémantických modelov

V predchádzajúcej kapitole bolo popísané ako sú samotné dátové formáty zastúpené na webových stránkach a doménach. V tejto kapitole bude analyzované ako sú v týchto dátových formátoch informácie definované resp. aké sémantické modely sú na webe najviac využívané.

Konkrétne teda budú analýzované dátové formáty RDFa, Microdata a JSON-LD. Pre každý z týchto formátov sú dostupné štatistiky, ktoré obsahujú tri tabuľky:

- počet použitých tried (class)
- počet použitých vlastností (property)
- počet použitých slovníkov (vocabulary)

V každej z týchto tabuliek sú ďalej štatistiky, ktoré konkrétne vyjadrujú počet výskytov entít, počet výskytov URL a počet výskytov pay-level domén (ďalej len domén).

Pre potreby následujúcej analýzy budú použité tabuľky, ktoré vyjadrujú počet výskytov menných priestorov, v ktorých sú definované jednotlivé slovníky, ktoré budú ďalej pomocou excelovskej PowerPivot funkcie prevedené do upravovateľnej tabuľky. Potom pomocou filtrov a zoskupovania budú záznamy ukazujúce na rovnaký slovník agregované do skupín a následne prezentované.

7.2.1 Využitie sémantických modelov vo formáte Microdata

Po spracovaní štatistík o sémantických modeloch používaných v rámci dátoveho formátu Microdata do vhodne prezentovateľnej formy sa dostávame k tabuľke nižšie.

Menný priestor modelu	Počet domén	Podiel domén	Počet URL	Podiel URL
schema.org	3733702	99,73%	578119783	89,44%
purl.org/dc/terms/	3549370	94,81%	606541635	93,83%
Iné	111161	2,97%	18222024	2,82%
data-vocabulary.org	1159110	30,96%	421255983	65,17%
Počet unikátnych domén/URL	3743822		646409625	

Tabuľka 3: *Tabuľka reprezentujúca využitie menných priestorov jednotlivých sémantických modelov vo formáte Microdata*

Celkovo bolo v extrakte nájdených 3 743 822 domén a 646 409 625 URL obsahujúcich štruktúrované dáta vo formáte Microdata.

Vidíme, že sa najviac vyskytujú menné priestory sémantických modelov schema.org, Dublin Core Metadata Initiative a data-vocabulary.org. Model schema.org má takmer 100% zastúpenie na doménach a takmer 89,44% zastúpenie na URL a podobne je na tom

aj model Dublin Core Metadata Initiative, ktorý sa vyskytuje na 94,81 % doménach a 93,83 % URL so štruktúrovanými dátami vo formáte Microdata.

Čo je zaujímavé, je to, že model data-vocabulary.org sa vyskytuje na na 65,17 % URL, ktoré obsahujú štruktúrované dáta vo formáte Microdata. Tento model je totižto od roku 2011 nepodporovaný a bol nahradený modelom schema.org. Bohužiaľ k tomuto modelu neexistuje ani dokumentácia, a teda nie je jasné či je kompatibilný s modelom schema.org. Takisto sa tento model vyskytuje na 30,96 % doménach so štrutkúrovanými dátami vo formáte Microdata. Tieto domény môžu byť už viac dodávateľmi nepodporované resp. stránky na nich sú viac nepodporované a upadli do zabudnutia.

7.2.2 Využitie sémantických modelov vo formáte JSON-LD

Po spracovaní štatistík o sémantických modeloch používaných v rámci dátoveho formátu JSON-LD do vhodne prezentovateľnej formy sa dostávame k tabuľke nižšie.

Menný priestor modelu	Počet domén	Podiel domén	Počet URL	Podiel URL
schema.org	2689746	100,15%	191777921	100,46%
iné	207	0,01%	208572	0,11%
Počet unikátnych domén/URL	2685738		190890906	

Tabuľka 4: *Tabuľka reprezentujúca využitie menných priestorov jednotlivých sémantických modelov vo formáte JSON-LD*

Celkovo bolo v extrakte nájdených 2 685 738 domén a 190 890 906 URL obsahujúcich štruktúrované dáta vo formáte JSON-LD.

Z kontigenčnej tabuľky je zrejmé, že miera výskytu sémantického modelu schema.org u dátového formátu JSON-LD je takmer výlučná. Čo je však zaujímavé je to, že samotný počet domén a URL s výskytom modelu schema.org je vyšší ako počet unikátnych domén resp. URL, v čase odovzdávania tejto práce som však nedostal žiadnu odpoveď od projektu Web Data Commons, ktorá by túto odchýlku vysvetľovala.

Výlučnosť modelu schema.org vyplýva z toho, že JSON-LD sa hlavne využíva pri tvorbe webových aplikácií pre prenos dát z backendu na frontend v rámci RESTového rozhrania, kde sú následne tieto dáta vykreslené do DOMu (Document Object Model). Ak sa teda pozrieme na oficiálnu dokumentáciu JSON-LD napr. v Node.JS prostredí, vidíme, že v celej dokumentácií je pre demonštráciu JSON-LD funkcionality používaný model schema.org. Dá sa teda očakávať, že tvorcovia webov budú využívať práve model schema.org spoločne s formátom JSON-LD.

7.2.3 Využitie sémantických modelov vo formáte RDFa

Po spracovaní štatistík o sémantických modeloch používaných v rámci dátoveho formátu RDFa do vhodne prezentovateľnej formy sa dostávame k tabuľke nižšie.

Menný priestor modelu	Počet domén	Podiel domén	Počet URL	Podiel URL
purl.org/dc	105196	8,70%	13282368	6,01%
purl.org/rss/1.0/modules/content/	84476	6,98%	6245850	2,83%
rdf.data-vocabulary.org/#	308378	25,50%	51553650	23,34%
rdfs.org/sioc/ns#	82359	6,81%	7487427	3,39%
xmlns.com/foaf/0.1/	102064	8,44%	11397220	5,16%
schema.org	518471	42,87%	30786205	13,94%
ogp.me	832359	68,82%	202797381	91,81%
Iné	114315	9,45%	20849538	9,44%
Počet unikátnych domén/URL	1209430		220889867	

Tabuľka 5: *Tabuľka reprezentujúca využitie menných priestorov jednotlivých sémantických modelov vo formáte RDFa*

Celkovo bolo v extrakte nájdených 1 209 430 domén 220 889 867 URL obsahujúcich štruktúrované dáta vo formáte RDFa.

V tabuľke môžme vidieť, že formát RDFa je najrôznorodejší, čo sa týka odkazovaných menných priestorov. Vidíme, že najviac výskytov v rámci unikátnych domén majú dva menné priestory a to ogp.me (kde je definovaný model Open Graph Protocol) a model schema.org, kde Open Graph Protocol sa vyskytuje na 68,82 % doménach a 91,81 % stránkach a schema.org na 42,87 % doménach a len na 13,94 % stránkach. Zvyšok tvoria menné priestory rdf.data-vocabulary.org (kde je definovaný model data-vocabulary), purl.org/dc (kde je definovaný model Dublin Core Metadata Initiative), xlmns.com/foaf/0.1 (kde je definovaný model Friend of a Friend), purl.org/rss/1.0/modules/content (kde je definovaný modul Content formátu RDF Site Summary) a rdfs.org/sioc (kde je definovaný model Semantically-Interlinked Online Communities).

Ak sa pozrieme na podiel výskytov menných priestorov na URL, tak vidíme, že najviac zastúpenými sú menné priestory ogp.me, schema.org a rdf.data-vocabulary.org. Opäť je zaujímavé vidieť, že na 23,34 % stránkach a 25,50 % doménach, ktoré obsahujú štruktúrované dáta vo formáte RDFa, sa vyskytuje model, ktorému skončila podpora v roku 2011. Ďalšou zaujímavosťou je to, že model ogp.me, ktorý umožňuje stránke aby sa z nej stal obohatený objekt v sociálnom grafe (využívaný napr. Facebookom), sa vyskytuje na takmer 91,81 % stránkach so štruktúrovanými dátami vo formáte RDFa, a teda, že vyskyt sémantických modelov v rámci RDFa je výrazne ovplyvnený pridávaním funkcionality sociálnych objektov na webové stránky.

7.3 Analýza schema.org

V predchádzajúcej kapitole boli porovnané miery využívanosti sémantických modelov na doménach a stránkach obsahujúcich štruktúrované dáta vo formátoch RDFa, Microdata a JSON-LD. V tejto kapitole bude analyzované ako je využívaný samotný model schema.org. Model schema.org je okrem svojho jadra pojmov tvorený aj rozšíreniami, ktoré sú špecifické pre dané odvetvie, preto môže byť zaujímavé ako sú tieto rozšírenia využívané. Ďalej je samotný dátový model schema.org tvorenými typmi, určite bude zaujímavé pozorovať v akej miere sú tieto typy využívané.

Keďže typy môžu byť definované viacerými možnými URI zápismi, či už sa jedná o samotný http protokol, ktorý je pri definovaní využitý alebo o pojmy zo schema.org oficiálnych rozšírení, ktoré môžu byť definované pomocou URI samotného rozšírenia ale takisto aj pomocou kanonického URI je nutné záznamy očistiť o časť obsahujúcu doménu, tak aby zostali len samotné názvy typov. Pre očistenie záznamov som teda použil následujúci regulárny výraz:

```
(https?:\/?\/?)?(health-
lifesci.)?(bib.)?(auto.)?(iot.)?(meta.)?(www.)?schema.org\/?
```

S modifikátorom g pre globálne vyhľadávanie a modifikátorom i pre case insensitive vyhľadávanie.

7.3.1 Využívanosť schema.org typov

Naprieč všetkými analýzovanými formátmi bolo nájdených celkovo 2672 rôznych definícií schema.org typov a následne po očistení pomocou regulárneho výrazu a využitím excelovskej power pivot funkcie, pomocou ktorej boli zlúčené záznamy s rovnakou hodnotou v názve typu, bolo následne nájdených 1179 unikátnych schema.org typov.

Naprieč samotnými URL má viac ako 1% zastúpenie 22 unikátnych typov, ktoré zároveň predstavujú 86,22 % všetkých schema.org typov nájdených v extrakte. Najvyužívanejšie schema.org typy v rámci URL sú reprezentované tabuľkou nižšie.

schema.org typ	Počet URL	Podiel URL
Organization	183122993	8,86%
WebSite	138546186	6,70%
Product	131854737	6,38%
BreadcrumbList	121828719	5,89%
Person	121211378	5,86%
ListItem	119573685	5,78%
SearchAction	114822276	5,55%
Offer	111290200	5,38%
WebPage	107495494	5,20%
ImageObject	91362574	4,42%
Article	81609950	3,95%
SiteNavigationElement	64467539	3,12%
PostalAddress	62913180	3,04%
AggregateRating	55359759	2,68%
WPHeader	46977244	2,27%
WPFooter	43293657	2,09%
NewsArticle	42792833	2,07%
WPSidebar	32227062	1,56%
Blog	31671940	1,53%
BlogPosting	30408394	1,47%
LocalBusiness	27130580	1,31%
CreativeWork	22504775	1,09%

Tabuľka 6: *Najvyužívanejšie schema.org typy v rámci URL*

Čo sa týka typov nájdených v rámci domén, tak viac ako 1% zastúpenie má 21 unikátnych typov, ktoré predstavujú 91,18 % všetkých schema.org typov nájdených v extrakte. Najvyužívanejšie schema.org typy v rámci domén sú reprezentované tabuľkou nižšie.

schema.org typ	Počet domén	Podiel domén
WebSite	2747652	15,53%
SearchAction	2299016	12,99%
Organization	1403919	7,93%
WebPage	1138147	6,43%
Person	753564	4,26%
SiteNavigationElement	701890	3,97%
Article	670790	3,79%
WPHeader	651120	3,68%
Product	644208	3,64%
WPFooter	600846	3,40%
Offer	557804	3,15%
PostalAddress	542043	3,06%
Blog	499289	2,82%
ImageObject	482116	2,72%
BlogPosting	447485	2,53%
WPSidebar	400294	2,26%
LocalBusiness	396020	2,24%
CreativeWork	370569	2,09%
BreadcrumbList	303450	1,71%
ListItem	295057	1,67%
AggregateRating	228494	1,29%

Tabuľka 7: *Najvyužívanejšie schema.org typy v rámci pay-level domén*

Z vyššie uvedených tabuliek je zrejmé, že štruktúrované dáta, ktoré sú na webe definované pomocou slovníka schema.org, najviac definujú objekty typu WebSite a Organization. Tieto typy umožňujú jednoznačne určiť webovú stránku alebo organizáciu webovým vyhliadačom a tak sa môžu vo výsledkoch vyhľadávania dostať vyššie a taktiež sa u výsledkov vyhľadávania zobrazia ďalšie obsiahnuté informácie ako napríklad sídlo u organizácie alebo ďalšie odkazy na danej webovej stránke, poprípade sa môžu tieto výsledky pri vyhľadávaní pomocou Google vyhliadavača zobraziť aj v Google znalostnom grafe.

Na základe najvyužívanejších typov môžme vidieť, že schema.org sa využíva hlavne kvôli SEO, čo je vlastne aj jej hlavným poslaním, ale určite má aj prínos pre sémantický web ako taký, kde pomocou schema.org môžme dáta vhodne štruktúrovať pre strojovú čitatelnosť.

7.3.2 Využívanosť schema.org rozšírení

Pre potrebu analýzy samotných schema.org rozšírení boli použité záznamy o schema.org typoch, ktoré boli očistené pomocou regulárneho výrazu z úvodu kapitoly 7.3. Ďalej bolo nutné správne priradiť tieto typy k ich prípadnemu rozšireniu, k tomu bola použitá excelovská funkcia VLOOKUP, ktorá vyzerá následovne:

```
=IFERROR(VLOOKUP(A2;extensions!$A$2:$B$489;2;0);"schema.org")
```

Kde A2 je relatívna hodnota odkazjúca na konkrétny schema.org typ a tabuľka zo záložky extensions obsahuje všetky typy schema.org rozšírení a konkrétne ju tvoria 2 stĺpce a to typ a jemu príslušiace rozšírenie.

V prípade, že vzorec nenájde typ z tabuľky extensions vráti hodnotu schema.org a teda daný typ je súčasťou základného schema.org modelu.

Týmto priradíme každému typu jeho rozšírenie a pre nájdené rozšírenie získame počet ich výskytov resp. sumu počtu výskytov typov patriacich do daného rozšírenia.

Ďalej pomocou PowerPivot funckie boli tieto záznamy zagregované do tabuľky, ktorá vyzerá následovne.

Rozšírenie	Počet domén	Počet URL
auto.schema.org	19	1591
bib.schema.org	211	116692
health-lifesci.schema.org	19242	2987872
meta.schema.org	195	9066
schema.org	17675120	2064169989
Celkom	**17694787**	**2067285210**

Tabuľka 8: *Absolútne vyjadrené zastúpenie schema.org rozšírení*

Vidíme, že základné schema.org typy tvoria podstatnú časť všetkých schema.org typov, konkrétne v rámci domén sa jedná o 99,89 % a v rámci URL 99,85 %.

Aby bolo využitie schema.org rozšírení lepšie prezentovateľné bol základný schema.org model odfiltrovaný od rozšírení a následujúcou tabuľkou je reprezentované využitie schema.org rozšírení.

Rozšírenie	Počet domén	Podiel domén	Počet URL	Podiel URL
auto.schema.org	0,10%	19	0,05%	1591
bib.schema.org	1,07%	211	3,75%	116692
health-lifesci.schema.org	97,84%	19242	95,91%	2987872
meta.schema.org	0,99%	195	0,29%	9066
Celkom	**100,00%**	**19667**	**100,00%**	**3115221**

Tabuľka 9: *Zastúpenie schema.org rozšírení bez základneho modelu*

Vidíme, že výrazne najväčsie zastúpenie má rozšírenie health-lifesci ktoré predstavuje 97,84 % rozšírení v rámci domén a 95,91 % rozšírení v rámci webových stránok. Toto rozšírenie obsahuje termíny z medicínskeho prostredia. Je určite výhodné mať medicínske termíny v strojovo čitateľnej forme, pretože to môže umožniť prácu lekárom napríklad pri určovaní diagnóz a predpisovaní liekov na základe zadaných príznakov.

Zvyšné tri rozšírenia predstavujú dokopy 2,16% zastúpenie na doménach a 4,09% zastúpenie na URL, to sa dá vysvetliť tým, že tieto rozširenia niesú tak robustné a detailné ako rozšírenie health-lifesci, resp. základ schema.org obsahuje už veľa termínov, ktoré by inak tieto rozšírenia mohli využívať.

Rozšírenie, ktoré sa v extrakte nevyskytlo ale je spomínané v oficiálnej dokuemntácií je rozšírenie iot.schema.org. Toto rozširenie je momentálne v prípravnej fázi a schema.org komunita rieši či ho vôbec vytvoriť alebo zlúčiť s projektom W3C Web of Things.

7.4 Zhrnutie

Z dostupných dát bola analyzovaná miera využívanosti schema.org na sémantickom webe. Najprv bola analyzovaná jej miera využívanosti v porovnaní s inými sémantickými modelmi, potom jej typy a nakoniec využívanosť jej rozšírení.

Z crawlu uskutočneného projektom Common Crawl boli extraktované štruktúrované dáta projektom Web Data Commons, ktorý taktiež zverejňuje štatistiky o vyextraktovaných dátach z formátov RDFa, Microdata a JSON-LD. Práve tieto štatistiky boli ďalej upravené pomocou nástroja Microsoft Excel, ktorý je vhodný pre dátovú analýzu malých data setov.

V extrakte boli najviac prítomné štruktúrované dáta vo formáte Microdata, a hlavne teda naprieč URL, narozdiel od domén, kde bol podiel formátov vyrovnanejší. Toto je spôsobené tým, že výrazny počet stránok z extraktu je hostovaný na doménach ako WordPress a BlogSpot.

Pri analýze sémantických modelov použitých vo formátoch RDFa, Microdata a JSON-LD bolo zistené, že sémantický model data-vocabulary.org, nepodporovaný od roku 2011 je stále využívaný na výraznom počte URL, hlavne vo formátoch Microdata a RDFa.

Vo formáte Microdata sú hlavne využívané modely schema.org a Dublin Core Metadata Initiative.

Ďalej, že využívanie sémantických modelov u formátu RDFa je výrazne ovplyvnené stránkami, ktoré obohacujú svoj obsah tak aby mohol byť súčasťou sociálneho grafu .

U formátu JSON-LD bolo zistené, že je v ňom sémantický model schema.org využívaný takmer vylúčne. To je spôsobené tým, že oficiálna dokumentácia JSON-LD pre všetky svoje demonštrácie využíva model schema.org.

Čo sa týka samotnej analýzy typov schema.org, bolo zistené, že najviac sú využívané typy, ktoré vylepšujú SEO danej stránke a obohacujú týmito štruktúrovanými dátami to čo sa zobrazí u ich výsledku pri vyhľadávaní.

Ďalej bolo zistené, že typy zo schema.org rozšírení predstavujú 0,11 % schema.org typov v rámci domén a 0,15 % v rámci URL nájdených v extrakte. Z nich drvivú väčšinu tvoria typy z rozšírenia health-lifesci.

8 Záver

V záverečnej kapitole tejto práce je vhodné zhodnotiť ako sa podarilo naplniť stanovené ciele a aké sú očakávané prínosy tejto práce.

8.1 Ciele a ich naplnenie

Cieľom teoretickej časti bolo vytvoriť základ pre časť praktickú, a teda je popísaná štruktúra a architektúra sémantickeho webu a význam jeho komponentov, dátové modely využívané na sémantickom webe a konkrétne teda model schema.org, ďalej je názorne demonštrované ako sa zapisujú štruktúrované dáta na webe tak aby boli strojovo čitateľné, konkrétne teda vo formátoch RDFa, Microdata a JSON-LD.

Teoretická časť je štruktúrovaná tak, aby každá kapitola tvorila čitateľovi podklad pre čítanie následujúcej kapitoly. Preto je najskôr popísaná štruktúra sémantického webu, následne je problematika zúžená na taxonómie, a teda sémantické modely využívané v rámci sémantickeho webu, ďalej je problematika zúžená konkrétne na model schema.org a na záver sú popísané možnosti zápisu štruktúrovaných dát pomocou formátov RDFa, Microdata a JSON-LD vo vzťahu k schema.org.

Praktická časť je zameraná na samotnú analýzu modelu schema.org na sémantickom webe. Pre analýzu boli využité dáta získané extraktovaním štrukturovaných dáta z web crawlu projektu Common Crawl projektom Web Data Commons.

Je analyzované to aké má využitie model schema.org v rámci štruktúrovaných dátach vo formátoch RDFa, Microdata a JSON-LD v porovnaní s ostatnými využívanými sémantickými modelmi.

Ďalej je analyzované využitie typov schema.org naprieč doménami a URL a taktiež využitie schema.org rozšírení.

Z výsledkov je zrejmé, že model schema.org má výrazné zastúpenie v štruktúrovaných dátach na sémantickom webe.

Stále veľa stránok využíva napríklad model data-vocabulary.org, ktorý je od roku 2011 nepodporovaný, a ktorý je predchodcom modelu schema.org, avšak vôbec nie je jasné či sú kompatibilné.

8.2 Prínosy práce

Teoretická časť práce tvorí úvod do sémantického webu a konkrétne do vytvárania a spracovania štruktúrovaných dát v tomto prostredí za využitia modelu schema.org.

Praktická časť prezentuje sémanticky web z pohľadu využitia týchto štrukturovaných dát vo formátoch RDFa, Microdata a JSON-LD spolu s jednotlivými sémantickymi modelmi a taktiež znázorňuje najvyužívanejšie typy a rozšírenia modelu schema.org.

8.3 Motivácia pre ďalšie práce

Zaujímavé je, že každý zdroj s ktorým som pri tvorbe práce pracoval rieši to, ako spracovať obyčajné dáta do strojovo čitateľnej podoby pomocou klasickej HTML syntaxe, resp. jej rozšírenia. Tento spôsob je vhodný ale hlavne pre statické stránky, poprípade dynamické stránky využívajúce zastaralé technológie. V súčasnosti však máme javascript frameworky, ktoré sú rýchle a hlavne nevyužívajú klasickú DOM štruktúru ale virtuálny DOM, ktorý pri zmenách renderuje len svoju určitú časť. Medzi tieto frameworky patria napríklad ReactJS alebo AngularJS, ktoré začínajú získavať výrazny podiel na webe, avšak neexistuje pre nich žiadne rozšírenie pomocou, ktorého je možné pridať im sémantiku, resp. je nutné odkazovať sa na vyrenderovanú časť HTML a takto im pridávať sémantiku, čo je často zbytočne komplikované a mohlo by to byť vyriešené pomcou nejakého vhodného balíčku.

Ďalej je vhodné spomenúť to, že existuje veľa modelov, ktoré vo svojom princípe riešia to isté, a to, že definujú pojmy používané v reálnom svete konkrétne teda schema.org, Dublin Core Metadata Initiative a FOAF. Určite by bolo zaujímavé zistiť nakoľko sú komaptibilné a porovnať to, ako definujú podobné entity zo skutočného sveta.

Terminologický slovník

Termín	Význam
URL	Uniform resource locator – adresa určujúca umiestnenie dokumentu na internete
URI	Uniform resource identifier – textový reťazec slúžiaci k presnej špecifikácií zdroja
IRI	Internationalized Resource Identifier – rozšírenie URI o sadu znakov Unicode
RDF	Resource Description Framework – model pre výmenu dát na internete
RDFa	Resource Description Framework in Attributes – technika pre priame pridávanie štruktúrovaných dát do HTML dokumentov
JSON-LD	JavaScript Object Notation for Linking Data – formát pre zaznamenávanie štruktúrovaných dát na internete
SEO	Search Engine Optimization – optimalizácia pre vyhliadavače, sada odporúčaní pre tvorbu webov, ktoré sa vo výsledkoch vyhľadávaní objavujú medzi prvými a obsahujú vhodne štruktúrované dáta
HTML	HyperText Markup Language – značkovací jazyk pre tvorbu webových stránok
DOM	Document Object Model – objektovo orientovaná reprezentácia XML alebo HTML dokumentu

Použité informačné zdroje

ALLEMANG, Dean a HENDLER, James A. 2008, Semantic web for the working ontologist modelling in RDF, RDFS and OWL. Burlington: Morgan Kaufmann Publishers. 330 s. ISBN 978-0-12-373556-0.

AMP. 2017, Building the future web, together [online], AMP, 2017 [cit. 25.11.2017]. Dostupné z WWW: https://www.ampproject.org/

BRICKLEY, Dan a MILLER, Libby. 2014, FOAF Vocabulary Specification 0.99 [online], Creative Commons, 14.1.2014 [cit. 30.11.2017]. Dostupné z WWW: http://xmlns.com/foaf/spec/

CHRIS, Alex. 2016, How to Boost your SEO with Schema Markup [online], Digital Marketing Pro, 2016 [cit. 25.11.2017]. Dostupné z WWW: https://www.digitalmarketingpro.net/how-to-boost-your-seo-with-schema-markup-751/

Common Crawl Foundation. 2017, Common Crawl [online], Common Crawl Foundation, 2017 [cit. 27.11.2017]. Dostupné z WWW: http://commoncrawl.org/

Creative Commons. 2017, JSON for Linking Data [online]. Creative Commons Public Domain, dátum neznámy [cit. 25.11.2017]. Dostupné z WWW: http://json-ld.org/

DE VIRGILLIO, Roberto a i. 2012. Semantic Search over the Web, Heidelberg: Springer. 417 s. ISBN 978-3-642-25007-1.

HTML Working Group. 2013, HTML Microdata [online]. W3C, 29.10.2013 [cit. 24.11.2017]. Dostupné z WWW: https://www.w3.org/TR/microdata/

JSON for Linking Data W3C Community Group. 2017, JSON-LD 1.1 [online]. W3C, 17.11.2017 [cit. 26.4.2017]. Dostupné z WWW: http://json-ld.org/spec/latest/json-ld/

MDN. 2017, itemref [online]. MDN, naposledy upravené 16.4.2017 [cit. 22.11.2017]. Dostupné z WWW: https://developer.mozilla.org/en-US/docs/Web/HTML/Global_attributes/itemref

MEDIAWIKI. 2011, User Guide [online]. Github, 6.9.2011 [cit. 30.4.2018]. Dostupné z WWW: https://github.com/dcmi/repository/blob/master/mediawiki_wiki/User_Guide.md

NETWORK WORKING GROUP. 2005, Uniform Resource Identifier (URI): Generic Syntax [online]. Network Working Group, Január 2005 [cit. 24.11.2017]. Dostupné z WWW: https://www.ietf.org/rfc/rfc3986.txt

OBITKO, Marek a ZAMAZAL, Ondřej a SVÁTEK, Vojtěch. Ontologie a sémanticky web. In: MAŘÍK, Vladimír ai. Umělá inteligence (6). Praha: Academia 2013, s. 85-123. ISBN 978-80-200-2276-9

THE OPEN GRAPH PROTOCOL. 2017, The Open Graph protocol [online]. ogp.me, 27.7.2017 [cit. 30.11.2017]. Dostupné z WWW: ogp.me

PETERKA, Jiří. 2015, Báječný svět počítačových sítí, část II. – Taxonomie, aneb škatulkovaní [online], eArchiv.cz, 2015 [cit. 21.11.2017]. Dostupné z WWW: http://www.earchiv.cz/b05/b0300100.php3

PURL. 2017, What is PURL? [online]. Internet Archive, dátum neznámy [cit. 30.11.2017]. Dostupné z WWW: https://archive.org/services/purl/help

RDF Working Group. 2014, RDF 1.1 Concepts and Abstract Syntax [online], W3C, 25.2.2014 [cit. 28.11.2017]. Dostupné z WWW: https://www.w3.org/TR/rdf11-concepts/

RDFa Working Group. 2015, RDFa Core 1.1 – Third Edition [online]. W3C, Marec 2015 [cit. 23.11.2017]. Dostupné z WWW: https://www.w3.org/TR/rdfa-syntax/

RDFa Working Group. 2015a, RDFa 1.1 Primer – Third Edition [online]. W3C, 17.3.2015 [cit. 23.11.2017]. Dostupné z WWW: https://www.w3.org/TR/2015/NOTE-rdfa-primer-20150317/

Schema.org Community Group. 2015, schema.org [online]. schema.org, 2015 [cit. 25.11.2017]. Dostupné z WWW: schema.org

SCHWARTZ, Barry. 2015, Google: Structured Markup May Become A Ranking Factor [online], Search Engine Roundtable, 11.9.2015 [cit. 25.11.2017]. Dostupné z WWW: https://www.seroundtable.com/google-structured-markup-ranking-purposes-20885.html

SIKOS, Leslie. 2015, Mastering Structured Data on the Semantic Web, Berkeley: Apress. 233 s. ISBN 978-1-4842-1050-5.

Stanford Encyclopedia of Philosophy. 2004, Logic and Ontology [online], Stanford, naposledy upravené 30.8.2011 [cit. 20.11.2017]. Dostupné z WWW: https://plato.stanford.edu/entries/logic-ontology/

W3C. 2017, Semantic Web [online]. W3C, dátum neznámy [cit. 22.11.2017]. Dostupné z WWW: https://www.w3.org/standards/semanticweb/

W3C Working Group. 2014, RDF Schema 1.1 [online]. W3C, 25.2.2017 [cit. 30.11.2017]. Dostupné z WWW: https://www.w3.org/TR/rdf-schema/

W3Techs. 2017, Usage statistics and market share of WordPress for websites [online]. W3Techs, December 2017 [cit. 4.12.2017]. Dostupné z WWW: https://w3techs.com/technologies/details/cm-wordpress/all/all

Web Data Commons. 2018, Web Data Commons – RDFa, Microdata, Embedded JSON-LD, and Microformats Data Sets – November 2017 [online]. Web Data Comons, 2018 [cit. 20.4.2018]. Dostupné z WWW: http://webdatacommons.org/structureddata/2017-12/stats/stats.html

Web Data Commons. 2012, Web Data Commons – Extraction Framework [online], Web Data Commons, 2014 [cit. 26.11.2017]. Dostupné z WWW: http://webdatacommons.org/framework/index.html

Web Data Commons. 2012a, Extracting Structured Data from the Common Crawl [online], Web Data Commons, 2014 [cit. 26.11.2017]. Dostupné z WWW: http://webdatacommons.org/

XML Core Working Group. 2009, Namespaces in XML 1.0 (Third Edition) [online]. W3C December 2009 [cit. 22.11.2017]. Dostupné z WWW: https://www.w3.org/TR/2009/REC-xml-names-20091208/#NT-NCName

ZELDMAN, Jeffrey. 2006, Web 3.0 [online], Alistapart, 2006 [cit. 20.4.2018]. Dostupné z WWW: https://alistapart.com/article/web3point0

Zoznam obrázkov

Obrázok 1: *Architektúra sémantického webu (zdroj: M. Obitko, O. Zamazal a V. Svátek, 2013)* 10
Obrázok 2: Data Extraction Framework princíp (zdroj: Web Data Commons, 2012) 26
Obrázok 3: Graf reprezentujúci počet domén obsahujúcich triplety naprieč dátovými formátmi (zdroj: Web Data Commons, 2018) 28
Obrázok 4: Graf reprezentujúci počet URL obsahujúcich triplety naprieč dátovými formátmi (zdroj: Web Data Commons, 2018) 28

Zoznam výpisov kódov

Výpis 1: Zápis záznamu o filme Good Will Hunting v klasickej HTML štruktúre 19
Výpis 2: Zápis záznamu o filme Good Will Hunting pomcou RDFa v strojovo čitateľnej forme 20
Výpis 3: Zápis záznamu o filme Good Will Hunting pomcou Microdata v strojovo čitateľnej forme 22
Výpis 4: Zápis záznamu o filme Good Will Hunting pomcou JSON-LD v strojovo čitateľnej forme 24

Zoznam tabuliek

Tabuľka 1: Tabuľka reprezentujúca záznam o filme Good Will Hunting 18
Tabuľka 2: Tabuľka popisujúca crawl z ktorého boli extraktované štruktúrované dáta (zdroj: Web Data Commons, 2018) 25
Tabuľka 3: Tabuľka reprezentujúca využitie menných priestorov jednotlivých sémantických modelov vo formáte Microdata 30
Tabuľka 4: Tabuľka reprezentujúca využitie menných priestorov jednotlivých sémantických modelov vo formáte JSON-LD 31
Tabuľka 5: Tabuľka reprezentujúca využitie menných priestorov jednotlivých sémantických modelov vo formáte RDFa 32
Tabuľka 6: Najvyužívanejšie schema.org typy v rámci URL 34
Tabuľka 7: Najvyužívanejšie schema.org typy v rámci pay-level domén 35
Tabuľka 8: *Absolútne vyjadrené zastúpenie schema.org rozšírení* 36
Tabuľka 9: *Zastúpenie schema.org rozšírení bez základneho modelu* 36