

UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

REPOZITÁR RECEPTOV V SIETI PREPOJENÝCH
DÁT
BAKALÁRSKA PRÁCA

2020
IVETA BALINTOVÁ

UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

REPOZITÁR RECEPTOV V SIETI PREPOJENÝCH DÁT

BAKALÁRSKA PRÁCA

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

Bratislava, 2020
Iveta Balintová



Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Iveta Balintová
Študijný program: aplikovaná informatika (Jednoodborové štúdium, bakalársky I. st., denná forma)
Študijný odbor: informatika
Typ záverečnej práce: bakalárska
Jazyk záverečnej práce: slovenský
Sekundárny jazyk: anglický

Názov: Repozitár receptov v sieti prepojených dát
Recipe repository in the linked open data network

Anotácia: Sieť prepojených dát (LOD) umožňuje publikovanie štruktúrovaných dát na webe. Takéto dáta sú následne strojovo spracovateľné vďaka ich anotácii vhodnými ontológiami.

Cieľ: Cieľom práce je navrhnúť a implementovať repozitár receptov v sieti LOD. Práca sa zameria na vhodné ontológie pre tento účel, prípadne ich úpravu, či tvorbu, a následne na vytvorenie repozitára a manažment dát receptov v tomto repozitári.

Literatúra: [1] Bizer, C., Heath, T. and Berners-Lee, T., 2011. Linked data: The story so far. In Semantic services, interoperability and web applications: emerging concepts (pp. 205-227). IGI Global.
[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] Rychtárik, M. 2019. Inteligentný receptár na báze prepojených dát. Bachelor thesis, Comenius University in Bratislava.

Vedúci: doc. RNDr. Martin Homola, PhD.
Katedra: FMFI.KAI - Katedra aplikovanej informatiky
Vedúci katedry: prof. Ing. Igor Farkaš, Dr.
Dátum zadania: 04.10.2019

Dátum schválenia: 14.10.2019

doc. RNDr. Damas Gruska, PhD.
garant študijného programu

.....
študent

.....
vedúci práce

Obsah

1	Východisková kapitola	1
1.1	Sémantický web	1
1.1.1	Linked Open Data	2
1.1.2	RDF	4
1.1.3	URI	6
1.1.4	Ontológie a slovníky	6
1.1.5	RDF Schema	7
1.1.6	OWL	8
1.1.7	SPARQL	9
1.1.8	Triplestore	10
1.2	Existujúce riešenia	10
1.2.1	Varecha	10
1.2.2	Yummly	11
1.2.3	Bakalárska práca Inteligentný receptár na báze prepojených dát	12
2	Návrh riešenia	13
2.1	Návrh ontológie	13
2.1.1	Protégé	14
2.1.2	Existujúca ontológia	14
2.1.3	Navrhnutá ontológia	14

Zoznam obrázkov

1.1	Semantic Web Layer Cake, zdroj: [12]	2
1.2	Orientovaný graf, ktorý vznikol na základe nižšie uvedeného kódu . . .	5
1.3	Výsledok dopytu nad DBpedia	9
1.4	Recept na vianočku, zdroj: [11]	10
1.5	Recept na gazpacho, zdroj: [16]	11
2.1	Grafické vyjadrenie návrhu ontológie o receptoch	13

Kapitola 1

Východisková kapitola

1.1 Sémantický web

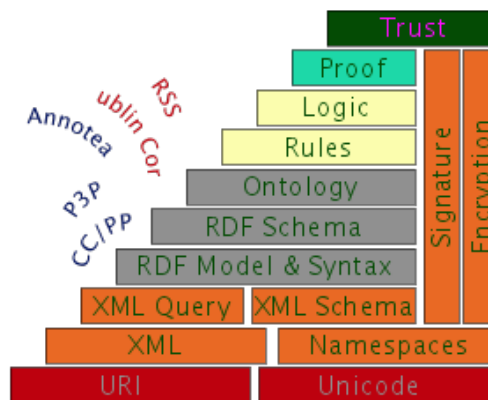
Sémantický web, podľa [15] je rozšírením súčasného webu. Väčšina obsahu, ktorý sa v súčasnosti na webe nachádza je navrhnutá pre ľudí. Programy dokážu rozpoznať, aká časť stránky je hlavička, kde sa nachádza odkaz na inú stránku, avšak nedokážu pochopiť význam toho, čo sa vlastne v jednotlivých častiach nachádza. Riešením má byť práve sémantický web, ktorý umožňuje reprezentovať a voľne publikovať dáta v nejakom vhodnom jazyku (RDF, viď. 1.1.2), ktorý im umožňuje priradiť aj význam. Tradičné systémy na reprezentáciu poznatkov boli väčšinou centralizované, čo si vyžadovalo jednotné definície bežných pojmov. Takáto centrálna kontrola sa však kvôli neustálemu rozrastaniu systému stáva nezvládnuiteľnou, a teda definície a významy priradzujeme použitím ontológií (viď. 1.1.4).

Semantic Web Stack, prípadne Semantic Web (Layer) Cake ilustruje štruktúru sémantického webu. Na obr. 1.1 vidíme ako sú organizované technológie, ktoré sú pre tento web štandardizované. Každá vrstva využíva schopnosti nižšej vrstvy. Tieto vrstvy môžeme rozdeliť do 3 skupín a na základe [8] ich definovať nasledovne:

- Pojmy zo spodnej vrstvy poznáme už z hypertextového webu. Z toho vyplýva, že sémantický web je iba rozšírením klasického webu. URI (viď. 1.1.3) umožňujú jednoznačne identifikovať zdroje. UNICODE je kódovanie slúžiace na reprezentáciu a spracovanie textov v rôznych jazykoch. XML, XML Query, XML Schema a XML Namespaces boli vyvinuté a štandardizované konzorciom W3C. XML je značkový jazyk, ktorý umožňuje vytvárať dokumenty so štruktúrovanými dátami, ktorým sémantický web dodáva význam. XML Query je štandardizovaný jazyk na kombinovanie dokumentov, databáz, či webových stránok, umožňuje robiť dopyty nad XML dátami. XML Schema je jazyk na vyjadrenie obmedzení o XML dokumentoch a špecifikuje, ako formálne opísať prvky v týchto dokumentoch. Názvy elementov si definuje autor XML dokumentu. Pri spájaní viacerých

dokumentov môže dochádzať ku konfliktu názvov v prípade, keď je rovnaký názov použitý s iným významom. Práve tento problém riešia XML Namespaces. Vytvorením prefixu pre namespace a následným používaním prefixu využívame názvy elementov v takom kontexte, v akom sú definované v nami vybranom namespace.

- Druhou skupinou, nachádzajúcou sa v strede, sú technológie štandardizované W3C. Umožňujú budovať aplikácie sémantického webu. Sú nimi RDF Model a Syntax (viď. 1.1.2), RDF Schema (viď. 1.1.5) a Ontológie (viď. 1.1.4). RDF je dátový model, v ktorom vyjadrujeme vzťah medzi hocíjakými dvoma zdrojmi vo forme trojice (subjekt, predikát, objekt). RDF Schema (RDFS) je univerzálny jazyk, ktorý umožňuje definovať nové slovníky. Ontológie, v užšom ponímaní slovníky, sú súbory pojmov a vzťahov medzi pojmami z určitej časti sveta, ktorá nás zaujíma.
- Treťou skupinou sú technológie, ktoré štandardizované zatiaľ nie sú alebo sú iba nápadmi. Na základe súborov pravidiel vieme pomocou logiky odvádzať nové informácie. Tieto odvodené informácie podporujeme dôkazmi. Najvyššie sa nachádza trust(pravda), teda mechanizmus, ktorým by sme vedeli určiť, ktorým informáciám môžeme dôverovať.



Obr. 1.1: Semantic Web Layer Cake, zdroj: [12]

1.1.1 Linked Open Data

Termín Linked Data, spracovaný podľa [3], označuje súbor všetkých dát publikovaných na webe pomocou technológií sémantického webu. Ide v podstate o sémantický web aplikovaný do praxe. Ak sú linked data, teda prepojené údaje, voľne dostupné na opätovné použitie označujeme ich termínom Linked Open Data. Linked Open Data sú silnou kombináciou prepojených a otvorených údajov. Jedným z pozoruhodných

príkladov LOD je je DBpedia – snaha o získanie štruktúrovaných informácií z Wikipédie a ich sprístupnenie na webe.

Zásady pri tvorbe prepojených údajov (Linked data principles) predstavil Tim Berners-Lee. Ide o tieto pravidlá:

1. Používajte URI ako názvy objektov.
2. Používajte identifikátory URI protokolu HTTP, aby ľudia mohli tieto mená vyhľadať.
3. Keď niekto vyhľadá URI, poskytnite užitočné informácie o tom, čo názov identifikuje pomocou otvorených štandardov, ako sú napr. RDF alebo SPARQL.
4. Zahrňte odkazy na ďalšie URIs, aby ľudia mohli dohľadať ďalšie informácie.

Štvrtý princíp prepojených údajov podporuje použitie hypertextových odkazov nie len medzi webovými dokumentami, ale medzi ľubovoľnými objektami, napr. odkaz medzi človekom a miestom, alebo medzi miestom a spoločnosťou. Oproti klasickému webu však majú tieto odkazy aj typy, teda napr. medzi dvoma ľuďmi môže mať odkaz typ priateľ.

Podľa [7], Tim Berners-Lee navrhol 5-hviezdičkovú schému rozvinutia linked open data.

- Jednu hviezdičku získajú dáta, ktoré sú sprístupnené na webe v ľubovoľnom formáte pod otvorenou licenciou. Môžeme si ich prečítať, zdieľať, meniť, je jednoduché ich poskytnúť, avšak dáta sú uzavreté v rámci dokumentu a je ťažké ich z neho dostať.
- Dve hviezdičky získajú dáta, ktoré sú sprístupnené v štruktúrovanej podobe, je možné ich spracovať a proprietárnym softvérom získať agregácie, či výpočty. Dáta sú však stále uzavreté v rámci dokumentu a ich získanie závisí od proprietárneho softvéru.
- Tri hviezdičky sú priradené dátam, ktoré sú sprístupnené v neproprietárnom otvorenom formáte.
- Štyroma hviezdičkami sú hodnotené dáta, v ktorých používame URI na určenie vecí tak, aby sa na nich mohli ľudia odkazovať. Takéto dáta je bezpečné kombinovať s inými dátami, keďže vieme, že v prípade rovnakého URI ide o rovnakú vec. Najbežnejším spôsobom reprezentácie dát je použitie RDF.
- Päť hviezdičiek, a teda najlepšie hodnotenie, získavajú dáta, ktoré sú nalinkované na iné. Vytvárame tým kontext a umožňuje nám to objavovať ďalšie súvisiace dáta.

1.1.2 RDF

Táto podkapitola je spracovaná podľa [14, 1]. RDF je skratka pre Resource Description Framework. Doteraz je to najjednoduchší a zároveň najsilnejší z formátov, ktorý poskytuje spôsob na vyjadrovanie informácií o zdrojoch. Zdrojom môže byť hocičo, teda dokument, osoba, fyzický objekt a podobne. RDF bol vyvinutý a odsúhlasený W3C. Je určený pre situácie, pri ktorých informácie na webe musia byť spracovávané prevažne aplikáciami, nie len zobrazované ľuďom.

RDF dáta sú zložené z tvrdení. Tvrdenie, v RDF nazývané aj trojica, resp. triple, umožňuje vyjadriť vzťah medzi dvoma zdrojmi a má vždy nasledujúcu štruktúru:

<subjekt> <predikát> <objekt>

Príklad trojice, vyjadrujúci informáciu, že Mount Everest je najvyšší vrch v Himalájach:

<Himaláje> <najvyšší vrch> <Mount Everest>

Pre lepšie pochopenie môžeme RDF dáta skúsiť porovnať s dátami v relačnej databáze, kde sú údaje uložené v tabuľkách. Zisťujeme, že v podstate bunka v tabuľke je prezentovaná tiež troma hodnotami. Identifikátor pre riadok sa nazýva subjekt, je to vec, o ktorej daný výrok hovorí. Identifikátor pre stĺpec sa nazýva predikát, hovorí o nejakej vlastnosti entity daného riadka(subjektu) a hodnota v bunke sa nazýva objekt. Ak by sme teda chceli dáta z relačnej databázy pretransformovať na dáta v RDF formáte, dalo by sa to pomerne jednoducho. V prípade opačného smeru, teda zmeny ľubovoľných RDF dát na relačnú databázu by bolo problémom, že predikáty nie sú pevne dané, a teda jediné, čo by sme vedeli s istotou vytvoriť by bola jedna veľká tabuľka s 3 stĺpcami, pričom každý riadok by vyjadroval jeden triple.

Trojice sa stávajú zaujímavejšími, keď viac ako jedna trojica opisuje rovnakú entitu, teda zdroj. V tom prípade je výhodné vizualizovať ich pomocou súvislého orientovaného grafu, pričom subjekt a objekt sú vrcholmi grafu a predikát je hrana medzi nimi. Skupina rovnakých tvrdení, môže byť preložená do rôznych syntaxí, avšak vždy budú predstavovať ten istý graf. Medzi formáty pre písanie RDF grafov patria Turtle family of RDF languages (N-Triples, Turtle, TriG, N-Quads), JSON-LD, RDFa, RDF/XML. Vyššie uvedená trojica (<Himaláje> <najvyšší vrch> <Mount Everest>) vyjadrená použitím Turtle syntaxe:

dbr:Himalayas dbp:highest dbr:Mount_Everest

RDF je teda predovšetkým systém na modelovanie údajov. Každý vzťah medzi hocíjakými dvoma entitami je presne reprezentovaný, čo umožňuje veľmi jednoduché

spájanie údajov z viacerých zdrojov. Dôvodom ľahkého spájania je to, že nie je potrebné usporiadať stĺpce tabuliek, obávať sa chýbajúcich údajov konkrétneho stĺpca a podobne, keďže vzťah buď existuje alebo nie.

Jednotlivé časti trojice (subjekt, predikát a objekt) môžu byť rôzne reprezentované. Prostredníctvom IRIs (viď. 1.1.3) môžeme vyjadriť subjekt, objekt aj predikát. Literály, jednoduché hodnoty ako sú reťazce, dátumy, čísla, môžu vystupovať iba ako objekt. V pozícií objektu a subjektu môžu vystupovať aj tzv. blank nodes. Sú to prázdne uzly, ktoré sa môžu použiť na označenie zdrojov bez toho, aby boli výslovne pomenované. Naznačujú teda existenciu nejakej veci, ktorá však nie je definovaná prostredníctvom IRI.

Príklad blank node (`_:someone`):

```
@prefix xsd: <http://www.w3.org/2001/XMLSchema#>.
```

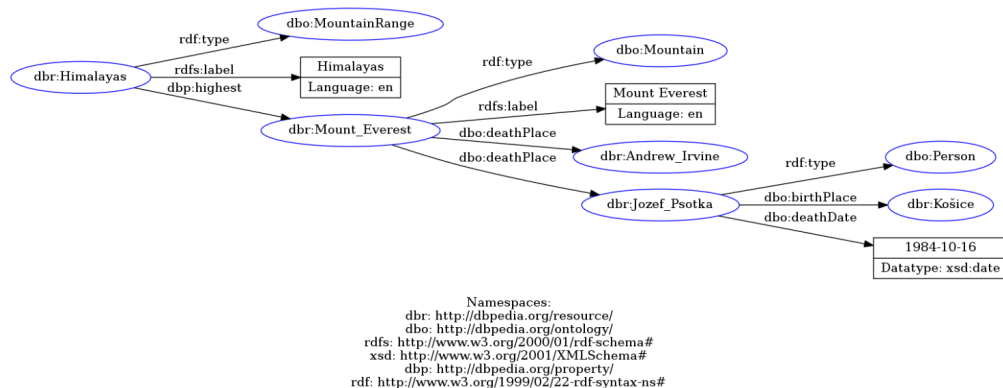
```
@prefix foaf: <http://xmlns.com/foaf/0.1/>
```

```
@prefix ex: <http://example.org/example#>.
```

```
ex:Jane foaf:knows _:someone.
```

```
_:someone foaf:name "John"^^xsd:string.
```

Na obr. 1.2 vidíme graf vyjadrujúci informácie o Mount Evereste, Himalájach a Jozefovi Psotkovi. Tento graf vznikol použitím Turtle syntaxe. Tá je najčitateľnejšia pre ľudí.



Obr. 1.2: Orientovaný graf, ktorý vznikol na základe nižšie uvedeného kódu

```
@prefix dbr: <http://dbpedia.org/resource/>.
```

```
@prefix dbo: <http://dbpedia.org/ontology/>.
```

```
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
```

```
@prefix xsd: <http://www.w3.org/2001/XMLSchema#>.
```

```
@prefix dbp: <http://dbpedia.org/property/>.
```

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
```

```
dbr:Himalayas
```

```
a dbo:MountainRange;
```

```
rdfs:label "Himalayas"@en;
```

```
dbp:highest dbr:Mount_Everest.
```

```
dbr:Mount_Everest
```

```
a dbo:Mountain;
```

```
rdfs:label "Mount Everest"@en;
```

```
dbo:deathPlace dbr:Andrew_Irvine;
```

```
dbo:deathPlace dbr:Jozef_Psotka.
```

```
dbr:Jozef_Psotka
```

```
a dbo:Person;
```

```
dbo:birthPlace dbr:Košice;
```

```
dbo:deathDate "1984-10-16"^^xsd:date.
```

1.1.3 URI

URI, podľa [1] ponúka globálnu identifikáciu pre zdroje na celom webe. Jeho syntax umožňuje „dereferenciu“ - použitie všetkých informácií v URI (ktoré špecifikujú meno servera, protokol, číslo portu, meno súboru a podobne) na to, aby lokalizovali súbor. Ak všetky časti fungujú, vtedy môžeme vyhlásiť, že URI je zároveň aj URL, teda URL je špeciálnym prípadom URI. Avšak tento rozdiel nie je podstatný z pohľadu modelovania. Vieme, že ak ukazujú dva vrcholy v RDF grafe na rovnaké URI, ide o ten istý vrchol. Na základe [14] vieme, že IRI je skratka pre International Resource Identifier, a je zovšeobecnením URI. V reťazci IRI môžu byť použité aj znaky, ktoré nie sú v ASCII.

1.1.4 Ontológie a slovníky


Ontológie

Ontológia [9] je súborom presných tvrdení o nejakej časti sveta, ktorá nás zaujíma. Zvyčajne máme k dispozícii množiny pojmov a vzťahy medzi nimi. Označujeme ich ako doménu záujmu alebo predmet ontológie. Jednotlivé tvrdenia sú prezentované ako triedy a vzťahy, a vytvárajú hierarchickú štruktúru celej ontológie. Súčasťou ontológií sú takisto aj obmedzenia a pravidlá, pomocou ktorých je možné odvodzovať nové fakty, ktoré neboli explicitne uvedené. Presnými opismi v ontológiách predchádzame nedorozumeniam, ku ktorým môže dochádzať v oblasti ľudskej komunikácie a zabezpečujeme,

aby sa softvér správal jednotne a predvídateľne.

Slovníky(vocabularies)

Podľa [14, 10], jednotlivé vyhlásenia, ktoré o zdrojoch robíme síce obsahujú IRIs, avšak dátový model nemá žiadne vedomosti o tom, čo v skutočnosti znamenajú. V praxi sa teda využívajú slovníky, ktoré na sémantickom webe definujú pojmy a vzťahy používané na opis a reprezentáciu nejakej oblasti záujmu. Tú následne využívame v nejakej konkrétnej aplikácii. Slovník je základným stavebným kameňom inferenčných techník na sémantickom webe. Príklady existujúcich slovníkov:

- FOAF, teda „Friend of a Friend“, je jedným z prvých RDF slovníkov používaných na celom svete. Popisuje ľudí, ich aktivity a interakciu s inými ľuďmi. Umožňuje rôznym skupinám ľudí popísať napr. sociálne siete bez potreby centralizovanej databázy.
- SKOS, teda Simple Knowledge Organization System, a je využívaný najmä ľuďmi zaoberajúcimi sa informatikou a knihovníkmi. 
- DC: Dublin Core, skupina termínov, ktoré sa používajú na opis digitálnych zdrojov (obrázky, webové stránky,...) alebo fyzických zdrojov (knihy, umelecké diela,...). Zahŕňajú vlastnosti ako napríklad "creator", "publisher" alebo "title".
- RDF Schema, popísaná v podkapitole 1.1.5

Porovnanie slovníkov a ontológií

Úlohou ontológií a slovníkov je teda pomáhať pri spájaní dát na sémantickom webe. Predchádzame tým napríklad prípadu, že v dvoch rôznych dátových množinách by boli použité rovnaké pojmy s rôznym významom.

Čo sa týka rozdielu medzi ontológiami a slovníkmi, podľa [10], neexistuje presný rozdiel, na základe ktorého by sme mohli tvrdiť, že niečo je určite slovník a niečo je určite ontológia. Pojmom ontológia označujeme komplexnejšiu množinu dát, a teda za hlavný rozdiel by sa dala považovať úroveň abstrakcie a vzťahov vrámci obsahu. Pri tvorbe ontológií sa využívajú existujúce slovníky na zníženie úsilia, ktoré musíme vynaložiť na budovanie ontológie od nuly.

1.1.5 RDF Schema

RDF Schema [6] je slovníkom pre modelovanie RDF dát. Poskytuje spôsob na opísanie skupín zdrojov, ktoré spolu súvisia a vzťahov medzi týmito zdrojmi. Triedy a vlastnosti v RDFS by mohli byť paralelou k triedam a atribútom používaným v objektovo-orientovaných programovacích jazykoch.

Príkladom tried z RDFS je `rdfs:Resource`, čo je trieda, do ktorej patrí každý zdroj. Ďalej trieda `rdfs:Literal`, vyjadrujúca hodnoty ako čísla, či textové reťazce alebo `rdfs:Container`, čo je trieda RDF kontajnerov, ktoré by sme mohli prirovnať k akýmisi poliam alebo množinám v programovacích jazykoch.

Príkladom vlastností z RDFS sú `rdfs:subClassOf`, teda vlastnosť byť podtriedou, `rdfs:domain`, vlastnosť vyjadrujúca doménu nejakého vzťahu, `rdfs:comment`, popis subjektu alebo `rdfs:seeAlso`, teda nejaké ďalšie informácie o subjekte.

1.1.6 OWL

Na základe [9] je OWL deklaratívny jazyk, teda logickým spôsobom popisuje stav vecí. Využíva sa na vyjadrenie ontológií a v podstate je rozšírením RDFS (viď. 1.1.5), napr. o termy, ktoré popisujú množinové operácie. Pomocou nástrojov na odvodzovanie je možné prinášať nové informácie a vytvárať závery, ktoré vyplývajú z toho, čo sme zdefinovali. Spôsob, akým sa tieto závery odvodzujú však závisí od konkrétnych implementácií a nie je súčasťou OWL dokumentu. Existujú rôzne syntaxe pre OWL, ktoré slúžia na rôzne účely, napr. RDF/XML syntax, ktorá ako jediná musí byť podporovaná všetkými nástrojmi OWL. Príklady rôznej syntaxe vidíme nižšie a je v nich vyjadrená hierarchia medzi triedami `Dog` a `Pet`, teda vlastnosť byť podtriedou.

Functional-Style Syntax

```
SubClassOf(:Dog :Pet)
```

RDF/XML Syntax

```
<owl:Class rdf:about="Dog">
<rdfs:subClassOf rdf:resource="Pet"/>
</owl:Class>
```

Turtle Syntax

```
:Dog rdfs:subClassOf :Pet .
```

Manchester Syntax

```
Class: Dog
```

```
SubClassOf: Pet
```

OWL/XML Syntax

```
<SubClassOf>
<Class IRI="Dog"/>
<Class IRI="Pet"/>
</SubClassOf>
```

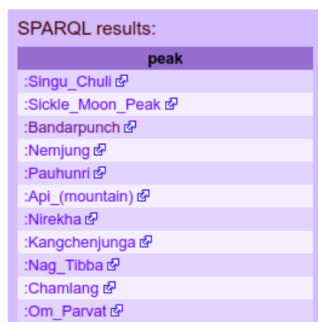
1.1.7 SPARQL

SPARQL, spracovaný na základe zdrojov [4, 2], je jazyk navrhnutý na vytváranie dopytov nad RDF dátami. Kľúčovými slovami v dopytoch sú PREFIX, SELECT, WHERE. Využívať môžeme aj FROM, ORDER BY, LIMIT a podobné výrazy, ktoré sú využívané aj v SQL dopytoch. Podmienky vo WHERE časti píšeme vo forme trojíc, ktoré sú podobné trojiciam v RDF, ale môžu obsahovať premenné. Tie zvyšujú flexibilitu pri porovnávaní s RDF dátami. Takisto sú podporované aj ASK dopyty, ktoré vracajú boolean hodnoty a CONSTRUCT dopyty, pomocou ktorých môžeme vytvárať RDF grafy z výsledkov dopytov.

Ak chceme robiť dopyty nad dátami nepotrebuje žiaden špeciálny softvér, keďže kolekcie dát sú často dostupné prostredníctvom SPARQL endpointov. Je to webová služba, ktorá akceptuje SPARQL dopyty a vracia výsledky. DBpedia je najpopulárnejším SPARQL endpointom.

Nižšie môžeme vidieť dopyt nad endpointom DBpedia (<http://dbpedia.org/snorql/>). Na základe neho dostávame zoznam vrchov, obr. 1.3, ktoré patria do triedy dbo:Mountain a zároveň sa nachádzajú v pohorí Himaláje. Prefix rdf:type, teda príslušnosť k určitej triede, sme nahradili skratkou a. Vo výslednom zozname je každý vrch popísaný prislúchajúcim URI, takže sa prostredníctvom neho môžeme dostať k ďalším informáciám o vybranom vrchu.

```
PREFIX dbo: <http://dbpedia.org/ontology/>
SELECT ?peak WHERE {
?peak a dbo:Mountain;
dbo:mountainRange :Himalayas.
}
```



SPARQL results:

peak
:Singu_Chuli ↗
:Sickle_Moon_Peak ↗
:Bandarpunch ↗
:Nemjung ↗
:Pauhunri ↗
:Api_(mountain) ↗
:Nirekha ↗
:Kangchenjunga ↗
:Nag_Tibba ↗
:Chamlang ↗
:Om_Parvat ↗

Obr. 1.3: Výsledok dopytu nad DBpedia

1.1.8 Triplestore

RDF Triplestore je podľa [5] typ grafovej databázy, v ktorej sú uložené sémantické fakty. Triplestorey sú uprednostňované na manažovanie prepojených dát, sú flexibilnejšie a lacnejšie ako relačná databáza. Zvládajú sémantické dopyty a používanie odvodzovania na zistenie nových informácií z už existujúcich vzťahov.

1.2 Existujúce riešenia

Táto kapitola sa venuje už existujúcim riešeniam daného problému. Uvádzam bakalársku prácu, ktorá sa touto témou zaoberala minulý rok a vybrala som si jednu slovenskú a jednu zahraničnú webovú aplikáciu, na ktorých je k dispozícii veľké množstvo receptov nie len ako jeden blok textu, ale ako dáta, s ktorými je možné pracovať.

1.2.1 Varecha

Varecha hľadaj jedlo QHľadaj

Suroviny

Vianočky:

- 1 kg múka hladká
- 1 kg múka polohrubá
- 50 g soľ
- 150 g cukor kryštálový
- 2 kocky drożdžie
- 1 l mlieko
- 250 g maslo
- 200 g (na lupienky posypať) mandľové
- 2 ks (na žltky potieranie)

Krém:

- 200 g lieskovce
- 560 g čokoláda
- 50 ml olej
- 100 g cukor kryštálový
- 200 g glukóza
- 50 g kakao
- 150 ml mlieko
- trocha extrakt vanilkový

Postup

1
Z droždia, trochy mlieka a cukru spravíme kvások, ktorý potom prilejeme k ostatným surovinám a vymiesime cesto, ktoré necháme na teplom mieste vykysnúť.

2
Z vykysnutého cesta spravíme dlhé valce...

3
...a postupne z nich vytvárame prepletaním vianočky, ktoré potrieme žltkami, posypeme mandľovými lupienkami a necháme znova nakysnúť. Pečieme v rúre vyhriatej na 175 °C asi 20 minút dozlata.

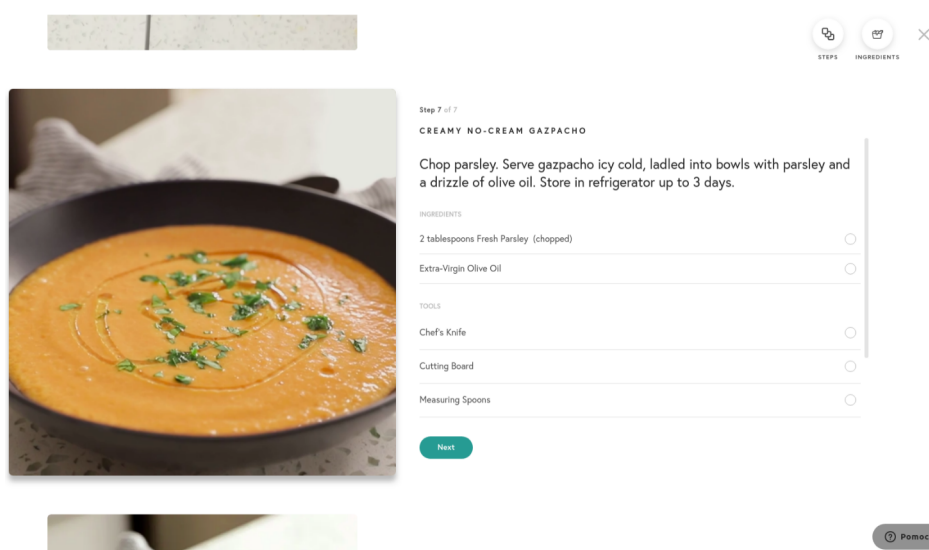
Obr. 1.4: Recept na vianočku, zdroj: [11]

Tento slovenský internetový portál [11] vznikol v roku 2009, pričom stihol zverejniť viac než 60 tisíc receptov. Pre neprihlásených používateľov je možnosť vyhľadávať recepty podľa druhov jedál, krajín, ingrediencií, atď. Pre prihlásených je dostupné aj vytváranie vlastného receptu, pridávanie komentárov, vlastných poznámok.

Vytváranie vlastného receptu, je zaujímavejšie, keďže nám vlastne predstavuje jeho jednotlivé časti. Samostatne sú vyplňané informácie ako názov, úvodný popis, čas vare-

nia, či tepelnej úpravy, množstvo porcií a podobne. Zaradzovanie do rôznych kategórií funguje na základe toho, že autor sám usúdi, kam sa recept hodí a danú kategóriu zaklikne. V prípade surovín je možnosť rozdeľovať ich na základe toho, v ktorej časti jedla sa využijú, pridávať ich množstvá a jednotky. Pri udaní názvu ingrediencie, ktorý web dokáže rozpoznať, je možné sa dostať kliknutím na ňu k jej nutričným hodnotám alebo ďalším jedlám, v ktorých sa využíva. Samotný postup je odkrokováný, avšak nikdy nie je presnejšie povedané, ako by mal daný krok vyzerieť, a teda je to na úvážení autora, čo tam uvedie. Ukážku receptu so surovinami a postupom vidíme na obr. 1.4.

1.2.2 Yummly



Obr. 1.5: Recept na gazpacho, zdroj: [16]

Podľa informácií na stránke [16] je tu k dispozícii viac ako 2 milióny receptov a má približne 26 miliónov používateľov. V prípade, že sme prihlásení, môžeme si v profile nastavovať rôzne preferencie, diéty, alergie a chute. Na základe všetkých týchto informácií nám aplikácia dokáže poskytovať stále relevantnejšie údaje bez toho, aby sme to museli vždy sami vyhľadávať.

Niektoré ingrediencie sa dajú jedným kliknutím na nich dohľadať priamo v obchode. Čo sa týka vyhľadávania, okrem bežných parametrov, ktoré sme si uviedli napr. aj pri portály Varecha [11], je možné vyhľadávať jedlá na základe spôsobu prípravy. Popisy postupov boli takisto podobné, pri videorecepte však boli aj samostatne uvedené nádoby, či iné zariadenia a zároveň aj ingrediencie potrebné na vykonanie daného kroku, tak ako vidíme na obr. 1.5. Nachádzajú sa tu však aj recepty, ktoré ako postup obsahovali odkaz na článok alebo inú stránku.

1.2.3 Bakalárska práca Inteligentný receptár na báze prepojených dát

Téme vytvárania receptov sa venovala minuloročná bakalárska práca [13]. Jej cieľom bolo najmä navrhnúť a implementovať webovú aplikáciu, ktorá spravuje a zobrazuje dáta nad štandardom sémantického webu.

Došlo k preskúmaniu existujúcich LOD zdrojov venujúcich sa tejto tématike, a zároveň k návrhu vlastnej ontológie. Jej hlavnou triedou bola trieda Recipe, ktorá popisovala samotný recept. Ten má dva hlavné atribúty, ktoré tvoria každý recept. Sú nimi ingrediencie a postup. Zoznam ingrediencií, a následne samotná ingrediencia sú reprezentované samostatnými triedami. Každá ingrediencia má nejakú jednotku merania a zároveň množstvo. Tieto vlastnosti sú popísané triedou Mass. Postup je v tejto ontológii reprezentovaný ako pole reťazcov. Ďalšími atribútmi prislúchajúcimi k triede Recipe sú autor, čas prípravy, počet porcií, kategórie, a zároveň hodnotenie, ktoré má tiež samostatnú triedu.

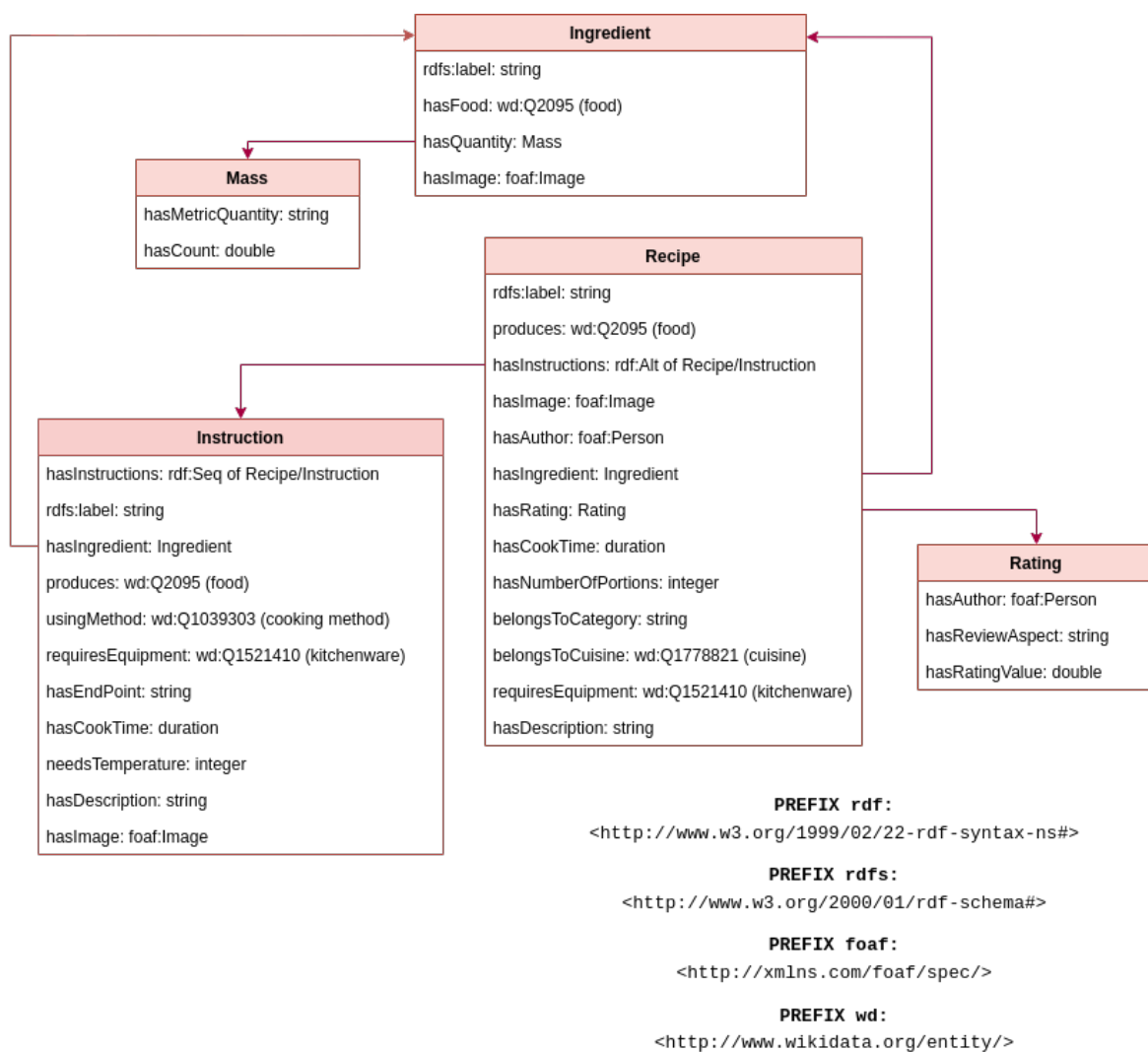
Začiatkové dáta boli čerpané zo stránky DBpedia. Na ich uskladnenie bol využitý triple store TDB, ktorý je komponentom Jena, čo je voľne dostupný Java framework pre vytváranie sémantického webu. Dopyty a aktualizácie boli vykonávané na SPARQL server Fuseki. Ten je spojený s TDB, používaného na trvalé uskladnenie dát. Pri pripojení na server a následné dopytovanie a aktualizáciu dát bola využitá knižnica EasyRdf, ktorá nám umožňuje ľahké vytváranie a prácu s RDF dátami. Okrem informáciách o receptoch a ingredienciách sa v aplikácii využívajú aj prihlasovacie údaje, či informácie o obľúbených receptoch. Z toho dôvodu boli využité dva typy databáz, na oddelenie osobných a verejných dát.

Čo sa týka výslednej aplikácie tá umožňuje používateľovi registráciu, vyhľadávanie receptu na základe rôznych kritérií ako sú napríklad ingrediencie, dĺžka prípravy, kategória alebo zásoby. Ďalej je možné pridávať ingrediencie do chladničky, vytvárať nákupný zoznam, vytvárať a upravovať recepty, a reagovať na ne, teda hodnotiť ich, exportovať do pdf, označovať ako obľúbené a podobne. Niektoré časti funkcionality sú prístupné len prihláseným používateľom.

Kapitola 2

Návrh riešenia

2.1 Návrh ontológie



Obr. 2.1: Grafické vyjadrenie návrhu ontológie o receptoch

2.1.1 Protégé

Na vytvorenie ontológie sme využili Protégé, ktorý je možné stiahnuť na stránke <https://protege.stanford.edu/>. Protégé poskytuje grafické používateľské rozhranie, je to voľne dostupný editor ontológií a framework pre budovanie inteligentných systémov. Umožňuje modelovať triedy, predikáty, či vytvárať obmedzenia na jednotlivé predikáty. Následne je možné uložiť takto vytvorenú ontológiu v niektorej zo syntaxí pre jazyk OWL. Pri tvorbe našej ontológie sme využívali túto základnú funkcionálnosť.

2.1.2 Existujúca ontológia

Naša ontológia nadväzuje na ontológiu o receptoch v minuloročnej bakalárskej práci [13]. Hlavnou úlohou bolo rozšíriť pôvodnú ontológiu o triedu, respektíve triedy, ktoré budú reprezentovať jednotlivé kroky (inštrukcie) v postupe receptu, keďže predtým bol postup iba pole reťazcov. V priebehu tvorby ontológie však došlo aj k ďalším menším zmenám v existujúcich triedach. Zároveň boli z pôvodnej ontológie odstránené triedy ingredientList a Food. Trieda ingredientList iba zapuzdrovala všetky ingrediencie patriace ku konkrétnemu receptu a trieda Food, ktorá vyjadrovala už existujúcu triedu a v našom prípade nebolo potrebné k nej pridávať žiaden ďalší predikát.

2.1.3 Navrhnutá ontológia

Nižšie je popísaná ontológia podľa obr. 2.1. Pri popisovaní ontológie je v zátvorkách uvedený presný názov predikátu tak, ako sa nachádza v ontológii a v jej grafickom vyjadrení na obr. 2.1. V pravom dolnom rohu obr. 2.1 sú uvedené využité prefixy v prípade, že trieda alebo predikát boli prevzaté z už existujúceho slovníka.

Popis tried existujúcich v predchádzajúcej ontológii

Hlavnou triedou je trieda Recipe, ktorá popisuje recept ako celok. Každý recept môže mať priradený obrázok (hasImage), názov vytváraného receptu zrozumiteľný pre používateľa (rdfs:label), autora (hasAuthor), čas potrebný na prípravu (hasCookTime), počet porcií, ktoré daný recept vytvára (hasNumberOfPortions), stručný popis, ktorý však ešte nepatrí k postupu (hasDescription) a možnosť zaradiť recept do nejakej kategórie (belongsToCategory), pričom tento predikát môže byť použitý opakovane v prípade, že recept patrí do viacerých kategórií. Ďalšími predikátmi sú príslušnosť k nejakej národnej kuchyni (belongsToCuisine), požadované kuchynské potreby (requiresEquipment), hodnotenia daného receptu inými používateľmi (hasRating). V tom prípade je objektom inštancia triedy Rating, pričom každé hodnotenie má autora (hasAuthor) a môže obsahovať nejaký slovný komentár (hasReviewAspect) alebo číselne vyjadrenú spokojnosť s receptom (hasRatingValue). Každý recept vytvára nejaké jedlo, ktoré sa

v niektorých prípadoch môže použiť ako ingrediencia v inom recepte, a preto bolo potrebné recept zaradiť aj do triedy Food (produces).

Predikát umožňujúci priradzovať k danému receptu ingrediencie (hasIngredient) spája triedu Recipe s triedou Ingredient. Každá ingrediencia môže mať názov zrozumiteľný pre používateľa (rdfs:label), obrázok (hasImage), jednotku merania (hasMetricQuantity), množstvo (hasCount), pričom posledné 2 spomenuté predikáty patria triede Mass. Každá ingrediencia je nejaké už existujúce jedlo (hasFood).

Popis triedy Instruction

Každý recept musí mať nejaký postup (hasInstructions). Objektom tohto predikátu je kontajner Alt, ktorý sa používa v prípadoch, kedy si používateľ môže vybrať jednu z možností, ktoré kontajner obsahuje. V tomto prípade je možné do kontajnera pridávať inštancie triedy Recipe, keďže k jednému jedlu existuje viacero receptov a takýmto spôsobom sa na nich odkážeme, alebo inštancie triedy Instruction, ak vytvárame náš vlastný postup.

Pri definovaní krokov v postupe receptu sme uvažovali nad viacerými možnosťami ako ich vytvárať. Plánovali sme oddeľovať triedu Description, Stage a Step, vďaka ktorým by sa dal recept pekne rozčleniť. Avšak nie každý recept má takúto štruktúru a pri iných typoch receptov by boli častokrát zbytočné triedy Stage, či Description. V prípade samotného kroku sme uvažovali nad tým, že bude obsahovať triedu Activity, ktorá presne popíše, ktorú surovinu, akým spôsobom, a za akých podmienok bude spracovávať. Toto by síce dokonale zobralo každý krok receptu, avšak praktické využitie takto podrobnej ontológie by bolo príliš náročné.

Rozhodli sme sa preto pre možnosť vytvoriť jedinú triedu Instruction. Inštrukcia môže zaoberať celú postupnosť krokov k receptu, môže vyjadrovať nejakú časť krokov, ktoré tvoria samostatnú časť jedla, napr. v prípade koláča to je inštrukcia na vytvorenie plnky, cesta, či polevy, alebo môže inštrukcia vyjadrovať jediný krok. Využitie triedy bude teda závisieť od toho, ako bude vyzeráť postup k receptu. Práve kvôli rôznorodosti postupov sme museli vytvoriť dostatočne všeobecnú triedu, aby mohla umožniť rovnako ľahko vytvoriť recept s jedným krokom, ako aj recept s pomerne zložitým postupom, v ktorom sú isté časti receptu oddelené do samostatných celkov. Používateľovi je umožnené priradiť techniku prípravy, ktorú je potrebné využiť pri danom kroku (usingMethod), napr. vyprážanie, pečenie, možnosť priradiť obrázok (hasImage), názov (rdfs:label), teplotu, pri ktorej má byť daná inštrukcia vykonávaná (needsTemperature), požadované kuchynské potreby (requiresEquipment), či potraviny spracovávané v danej časti receptu (hasIngredient). Trieda Instruction ďalej umožňuje vyjadriť čas potrebný na vykonanie nejakého kroku (hasCookTime). V niektorých prípadoch je trvanie kroku vyjadrené slovne, napr. pečieme, kým nezehnedne, vtedy použijeme

predikát, ktorý vyjadruje práve prechod ingrediencie do nejakého finálneho stavu (`hasEndPoint`). Ak ide o skupinu krokov, ktoré vytvárajú celý postup k receptu alebo časť receptu, napr. prípravu cesta pri pečení koláča, využijeme predikát `hasInstructions`. Ten ako objekt obsahuje sekvenciu ďalších inštrukcií, alebo receptov, v prípade, že by sme sa v nejakej časti receptu chceli odkázať na už existujúci recept. Trieda `rdf:Seq` nám umožňuje ukladať prvky tak, aby sme poznali poradie, v ktorom sme ich pridávali. Ak chceme popisovať už koncový krok, ktorý sa skladá iba zo slovného popisu a ďalej sa nečlení, využijeme predikát `hasDescription`. V prípade, že nejaká inštrukcia vytvára jedlo, môžeme využiť predikát `produces`.

Literatúra

- [1] *Semantic web for the working ontologist modeling in RDF, RDFS and OWL*. Denise E. M. Penrose, 2008.
- [2] *Learning SPARQL*. O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472, 2011.
- [3] *Linked Data: Evolving the Web into a Global Data Space (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology*. Morgan & Claypool, 2011.
- [4] SPARQL 1.1 overview. W3C recommendation, W3C, March 2013. <http://www.w3.org/TR/2013/REC-sparql11-overview-20130321/>.
- [5] Ontotext AD. What is rdf triplestore?, 2019.
- [6] Ramanathan Guha and Dan Brickley. RDF schema 1.1. W3C recommendation, W3C, February 2014. <http://www.w3.org/TR/2014/REC-rdf-schema-20140225/>.
- [7] Michael Hausenblas. 5 star open data, 2020.
- [8] Craig Knoblock. The semantic web, 2020.
- [9] Peter Patel-Schneider Bijan Parsia Markus Krötzsch, Sebastian Rudolph and Pascal Hitzler. Owl 2 web ontology language primer (second edition). W3C recommendation, W3C, 2012. <http://www.w3.org/TR/2013/REC-sparql11-overview-20130321/>.
- [10] Keio Beihang MIT, ERCIM. Vocabularies, 2020.
- [11] a.s. Varecha PEREX. Varecha, 2020.
- [12] Eric Prud'hommeaux. Semantic web specification at w3c, 2020.
- [13] Matej Rychtárik. Inteligentný receptár na báze prepojených dát, 2019.
- [14] Guus Schreiber and Yves Raimond. RDF 1.1 primer. W3C note, W3C, June 2014. <http://www.w3.org/TR/2014/NOTE-rdf11-primer-20140624/>.

- [15] James Hendler Tim Berners-Lee and Ora Lassila. A new form of web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific American*, 284(5):34–43, 2001.
- [16] X. Yummly, 2020.