Sem vložte zadání Vaší práce.

České vysoké učení technické v Praze
Fakulta informačních technologií
Katedra Softwarového inženýrství

Bakalářská práce

Webový server pro poskytování dynamicky generovaných objektů ve formátech RDF

Jan Řasa

Vedoucí práce: RNDr. Jakub Klímek, Ph.D.

27. dubna 2016

Poděkování

Díky všem

Prohlášení

Prohlašuji, že jsem předloženou práci vypracoval(a) samostatně a že jsem uvedl(a) veškeré použité informační zdroje v souladu s Metodickým pokynem o etické přípravě vysokoškolských závěrečných prací.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona, ve znění pozdějších předpisů. V souladu s ust. § 46 odst. 6 tohoto zákona tímto uděluji nevýhradní oprávnění (licenci) k užití této mojí práce, a to včetně všech počítačových programů, jež jsou její součástí či přílohou, a veškeré jejich dokumentace (dále souhrnně jen „Dílo“), a to všem osobám, které si přejí Dílo užít. Tyto osoby jsou oprávněny Dílo užít jakýmkoli způsobem, který nesnižuje hodnotu Díla, a za jakýmkoli účelem (včetně užití k výdělečným účelům). Toto oprávnění je časově, teritoriálně i množstevně neomezené. Každá osoba, která využije výše uvedenou licenci, se však zavazuje udělit ke každému dílu, které vznikne (byť jen zčásti) na základě Díla, úpravou Díla, spojením Díla s jiným dílem, zařazením Díla do díla souborného či zpracováním Díla (včetně překladu), licenci alespoň ve výše uvedeném rozsahu a zároveň zpřístupnit zdrojový kód takového díla alespoň srovnatelným způsobem a ve srovnatelném rozsahu, jako je zpřístupněn zdrojový kód Díla.

V Praze dne 27. dubna 2016 .....

Odkaz na tuto práci

Řasa, Jan. *Webový server pro poskytování dynamicky generovaných objektů ve formátech RDF*. Bakalářská práce. Praha: České vysoké učení technické v Praze, Fakulta informačních technologií, 2016.

Abstrakt

Řada objektů na webu dat tvoří natolik velkou skupinu (až nekonečnou), že je není možné všechny perzistentně uložit a publikovat. Je nutné nabídnout server, který na základě požadavku klienta na daný objekt příslušná data dynamicky vygeneruje a odešle klientovi. Tato práce se zabývá Cílem této práce je analýza již existujících řešení, návrh a implementace daného serveru. Server bude umožňovat uživateli definovat RDF (Resource Description Framework) objekty pro dynamické generování v několika formátech. Součástí této práce je také návrh vhodných formátů pro tyto definice. Návrh bude zohledňovat již zaběhlé technologie z oboru Linked data, jako je například dotazovací jazyk SPARQL (Protocol and RDF Query Language). Server bude implementován v jazyce Java.

Klíčová slova sémantický web, linked data, RDF, dynamické generování, SPARQL, java, webový server

Abstract

Sem doplňte ekvivalent abstraktu Vaší práce v angličtině.

Keywords Nahraďte seznamem klíčových slov v angličtině oddělených čárkou.

Obsah

Seznam obrázků

Úvod

Při publikování dat na webu dat je vždy důležité zamyslet se nad tím, jak a kde budou tyto data uložena. Způsobů je mnoho. Mezi ty nejzákladnější patří ukládání dat do různých databázových systémů, nebo pouze přímo na určené místo na disku. Oba zmíněné způsoby a jim podobné mají ovšem jeden zásadní problém, a tím je kapacita uložiště.

Pro uložení většiny informací, se kterými se lze setkat na webu, jako jsou obrázky, zprávy, informace o počasí a mnoho dalších, se tato skutečnot nemusí příliš řešit. Kapacita uložiští nám pro tyto informace stačí. Nicméně existuje řada informací(dále také objektů), které tvoří natolik velkou skupinu (až nekonečnou), že je není možné všechny perzistentně uložit a publikovat.

Dobrým příkladem jsou například časová data - konkrétní čas či časový interval. V kontextu Linked Data [**?**] se často odkazuje na časový objekt. Ať už se jedná například o časy příjezdů autobusů nebo o datum nějaké události, vždy je potřeba mít daný časový objekt nějakým způsobem uložen.

Existují ale kapacity na uložení každého takového objektu? Časových objektů je přeci nekonečně mnoho. Mohou odkazovat jak do minulosti, tak do budoucnosti. A není potřeba se omezovat pouze na časové objekty. Jako další příklad může být informace o vztahu mezi lidmi, respektive mezi kterýmikoliv subjekty. Tyto informace taktéž vyžadují ohromné kapacity uložiští.

Všechny tyto objekty ale spojuje fakt, že mají vždy stejnou strukturu a jen část informací se mění(například jen sekunda, nebo jméno člověka). A to je ideální příležitost pro to, aby se ukládání těchto objektů zaměnilo za dynamické generování. Pro vygenerování objektu stačí vždy použít stejnou strukturu a jen dosadit potřebné informace tak, aby vznikl požadovaný objekt.

Kapitola 1

Cíl práce

Cílem této práce je navrhnout, implementovat a otestovat webový server pro poskytování dynamicky generovaných objektů v RDF formátech. Server bude splňovat následující požadavky:

- Server bude umožňovat administrátorovi založit nový typ objektů včetně jejich atributů a umožní nakonfigurovat URL, pod kterými budou objekty dostupné.
- Typy objektů bude možné založit přes konfigurační soubor.
- Klient bude moci přístupem na dané URL získat data o příslušném objektu.
- Data o objektech budou klientům dostupná v RDF serializacích (RDF/XML, Turtle, N-Triples, JSON-LD) a jako webová stránka.
- Server bude využívat mechanismu Content Negotiation pro určení formátu výstupu požadavku klienta.
- Server bude implementován v jazyce Java.

Kapitola 2

Analýza

2.1 Účel aplikace

Webový server bude poskytovat uživatelům funkci pro dynamické generování objektů ve formátech RDF. Uživatel si bude moci nadefinovat vlastní typy objektů pomocí šablony včetně URL, na které budou dané objekty přístupné. Objekty budou generovány za použití informací, které bude obsahovat URL adresa při požadavku na daný objekt.

Tento způsob generování objektů ve velké míře šetří především finanční zdroje za pořizování nových uložišť. Pro velké množství stejných objektů (až nekonečně mnoho) stačí vytvořit pouze jednu definici (šablonu) objektu, která se dle specifikovaných pravidel vyplní informacemi z URL adresy a uživateli se zobrazí jako požadovaný objekt.

Další výhodou je z hlediska jednotné definice objektů i možnost lehce upravit strukturu konkrétních definic na jednom místě. Na konkrétní definici objektu může být odkazováno s různými parametry z mnoha jiných objektů a jedinou změnou v definici lze ovlivnit informace i v těchto objektech. Měnit již dříve vytvořené a uložené konkrétní objekty v takovém počtu by bylo téměř nereálné.

Účelem serveru je také chování, které co nejméně omezuje možné klienty [1]. Server využívá principu Content Negotiation [1] a podporuje nejpoužívanější typy RDF serializací, včetně možnosti zobrazit informace o objektu v HTML podobě při zobrazení prohlížečem.

2.2 Existující řešení

Pro funkce, které má tento server splňovat, neexistuje v současnosti žádné jiné řešení. Minimálně není možné dohledat žádné veřejné řešení, ani informace o nějakém soukromém řešení, které by sloužilo například pro soukromá data

[1] Zde se klientem rozumí především aplikace které by k objektům přistupovaly.

v rámci nějaké společnosti. Nicméně existuje velmi populární řešení jednoho typu objektu - časových intervalů, tzv. British Time Intervals, které je hojně používané, převážně pak také v rámci dat, které poskytuje britská vláda.

2.2.1 British Time Intervals

Pro popis těchto objektů je asi nejlepší odkázat se na konkrétní definici na webové stránce, ze které jsou tyto intervaly dostupné.[2]

> Linked data for every time interval and instant into the past and future, from years down to seconds. This is an infinite set of linked data. It includes government years and properly handles the transition to the Gregorian calendar within the UK.

Zde je vhodné všimnout si hlavně slov *infinite set*, což přesně charakterizuje typy objektů, kterými se tato práce zabývá.

2.2.1.1 Struktura URI

Pro přístup k jednotlivým generovaným objektům slouží URL adresa, která má vždy předepsanou strukturu. Informace, které jsou dostupné k tomuto datasetu, obsahují popis těchto struktur a jsou veřejně dostupné na stránkách společnosti Epimorphics [3]. Takovou URL adresou může být například *http://reference.data.gov.uk/doc/government-year/{year1}-{year2}*, kde po dosazení let míto *year1* a *year2* a přístupem na danou adresu získáme požadovaný objekt časového intervalu.

2.2.1.2 Generování objektu

Takový objekt je tedy dynamicky vygenerován s použitím parametrů z URL adresy. Konkrétní popis toho, jak jsou tyto objekty interně generovány není veřejný, ale je velmi pravděpodobné, že je použit minimálně jeden z následujících způsobů:

- Předem připravená RDF šablona (v libovolném formátu - Turtle, RDF/XML, N-TRIPLES ...) s placeholdery, do kterých se dosadí parametry z URL adresy.
- SPARQL [4][5] dotaz s placeholdery.

2.3 Požadavky

2.3.1 Funkční požadavky

- přístup k administraci objektů přes více rozhraní
- podpora více formátů pro definice a zobrazení objektů

- zobrazení existujících definic objektů
- administrace definic objektů
- konfigurace objektů bude možná několika způsoby
- vygenerování a zobrazení konkrétních objektů

2.3.1.1 Přístup k administraci objektů přes více rozhraní

Administrátori bude umožněn přístup k definicím objektů dvěma způsoby:

- přes webové rozhraní
- přes API

Oba tyto způsoby budou poskytovat stejnou funkcionalitu. Webovým rozhraním se myslí jednoduchá aplikace pro administraci objektů z webového prohlížeče. API bude sloužit k administraci z jiných potenciálních aplikací.

2.3.1.2 Podpora více formátů pro definice a zobrazení objektů

Server bude podporovat následující formáty pro definice objektů přes API:

- JSON formát [**?**]
- RDF serializace TURTLE [**?**]

Konkrétní vygenerované objekty budou klientům dostupné v těchto RDF serializacích:

- RDF/XML [**?**]
- JSON-LD [**?**]
- N-TRIPLES [**?**]
- TURTLE

Každá definice objektů bude také umožňovat definovat HTML formát objektu pro zobrazení v prohlížeči.

2.3.1.3 Zobrazení existujících definic objektů

Ve webové aplikaci budou zobrazeny všechny aktuální definice objektů v tabulce s možností konkrétní definici upravit (tedy také zobrazit definici konkrétního objektu) nebo smazat přes tlačítka vedle každé definice.

Přes API bude možné získat definice objektů ve dvou formátech:

- RDF serializace

- JSON formát

Konkrétní formát bude určen principem Content Negotiation. Získat půjde seznam všech definic a také konkrétní definice.

2.3.1.4 Administrace definic objektů

Administrátorovi bude umožněno přidávat nové definice, měnit a mazat již vytvořené definice. Tyto akce budou umožněny jak z webové aplikace, tak i přes API. Dále bude moci administrátor definovat objekt konfiguračním souborem v RDF serializaci TURTLE.

2.3.1.5 Konfigurace objektů bude možná několika způsoby

Koknfigurací objektů se zde myslí možné způsoby jak a z čeho se bude generovat výsledný RDF objekt. Server bude podporovat následující způsoby:

- generování ze SPARQL šablony - CONSTRUCT dotaz [**?**]

  Vstupem bude SPARQL šablona CONSTRUCT dotazu s placeholdery [2], za které se dosadí při požadavku na objekt hodnoty z URL adresy a provede se příkaz který vygeneruje objekt.

- generování ze SPARQL šablony - vzdálený CONSTRUCT nebo DESCRIBE dotaz [**?**]

  Vstupem bude SPARQL šablona jako v prvním případě. Navíc bude možné provést DESCRIBE dotaz. Rozdíl oproti prvnímu případu je v tom, že se dotaz přepošle na zvolenou adresu SPARQL Endpointu [**?**], zde se provede a klientovy je pak vrácen daný objekt.

- generování z RDF šablony

  Vstupem bude RDF šablona podporovaných serializací s placeholdery, za které se dosadí při požadavku na objekt hodnoty z URL adresy.

- Proxy objekt

  Server bude umožňovat roli prostředníka při generování objektů. Požadavek na objekt se přepošle na jiný server a výsledek se přeloží klientovi dle požadované serializace. Tento způsob umožňuje generování objektů přes jiné aplikace.

2.3.1.6 Vygenerování a zobrazení konkrétních objektů

Klient bude moci přístupem na konkrétní URL adresu získat data o příslušném objektu. Typ RDF serializace nebo zobrazení HTML stránky se určí přes Content Negotiation.

[2] Placeholder: část šablony, která jasně identifikuje místo, kam se dosadí parametry z URL adresy.

2.3.2 Nefunkční požadavky

- server bude implementován v jazyce Java
- celá aplikace bude uložena ve WAR souboru pro zjednodušené nasazení

Kapitola 3

Návrh

3.1 Použité technologie

3.1.1 Resource Description Framework (RDF)

Resource Description Framework je rodina specifikací, která se používá jako metoda pro modelování informací - objektů. Jedná se o model metadat, které popisují nějaké zdroje.

Příkladem může být obyčejná webová stránka obsahující nějaké informace. Webové stránky se zaměřují především na uživatele. Důležité je, aby se uživateli stránky líbily a dbá se tedy hodně na design. Pokud jsou stránky přehledné, pak člověk nemívá problémy pochopit dané informace. Nicméně stroj (počítač, program) tyto informace sice zobrazí uživateli, ale samotné informace si interpretovat nedokáže.

RDF model popisuje tedy způsoby, jakými docílit toho, že poskytované informace budou čitelné i pro stroje.

Základní kostrou RDF modelu dat jsou takzvané *trojice*. Tyto trojice se skládají ze subjektu, predikátu a objektu. Trojice se může volně přeložit i do podoby, kde subjekt má nějakou vlastnost (predikát) s konkrétní hodnotou (objekt). Tedy všechny trojice, které mají stejný subjekt tento subjekt definují.

Každý subjekt je identifikován nejčastěji přes URI. Bavíme-li se o datech na webu, tak zde může být URI klasická URL adresa, jak ji známe z běžného používání. Co se týče objektu (hodnoty), tak se může jednat o literál (řetězec, číslo apod.), ale hodnotou může být zase URI nějakého dalšího objektu. Dokonce i predikát může být objektem s vlastním URI. Tím, že se objekt skládá z dalších objektů, které jsou jednoznačně identifikovány pomocí URI, získáváme velkou výhodu tohoto modelu. Ve výsledku vzniká graf popisující tyto trojice, což je i pro stoje čitelná struktura.

Samotné RDF popisuje pouze model. Pro uložení tohoto modelu je zapotřebí informace nějakým způsobem serializovat. Pro uložení RDF objektů se používají nejčastěji tyto RDF serializace:

- RDF/XML
- Turtle
- N-Triples
- JSON-LD

Pro ukázku je zde příklad ve formátu Turtle, který je převzat z W3C specifikace [6].

```
@base <http://example.org/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix rel: <http://www.perceive.net/schemas/relationship/> .

<#green-goblin>
    rel:enemyOf <#spiderman> ;
    a foaf:Person ;  # in the context of the Marvel universe
    foaf:name "Green Goblin" .

<#spiderman>
    rel:enemyOf <#green-goblin> ;
    a foaf:Person ;
    foaf:name "Spiderman".
```

<#green−goblin> je zde subjekt identifikovaný URI (http://example.org/#green-goblin), *rel:enemyOf* je predikát - v tomto případě zase objekt identifikovaný URI (http://www.perceive.net/schemas/relationship/enemyOf) a *<#spiderman>* je objekt - zase identifikován URI. Tento model dohromady poskytuje informaci, že objekt na dané URI je nepřítelem Spidermana. *afoaf:Person* znamená, že se jedná o objekt Person a *foaf : name"GreenGoblin"* zase to, že jeho jméno je Green Goblin. Tímto je tedy definován objekt *< #green − goblin >* a podobně tomu je u objektu *<#spiderman>*.

3.1.2 Linked Data

Linked data, jak už název napovdá, popisuje metodu propojování informací na webu dat mezi sebou. Tim Berners-Lee, zakaldatel WWW, popisuje Linked Data výstižně takto: „*Semantic Web isn't just about putting data on the web. It is about making links, so that a person or machine can explore the web of data. With linked data, when you have some of it, you can find other, related, data.*“ [7]

Aby se publikovaná data mohly využívat v co největší míře, nestačí je pouze zpřístupnit na webu dat. Největší užitek přináší linkování těchto dat dohromady. Stejně, jak je tomu u HTML dokumentů, slouží pro linkování v rámci RDF URI, které jasně identifikují objekty. Nad takto prolinkovanými

daty se poté dají najít různé vztahy, které by bez prolinkování nebylo monžé nalézt.

Linked Data staví na těchto čtyřech základních principech:

- Pro názvy a identifikaci objektů se používá URI
- Aby byly data přístupné, používají se HTTP URI
- Při přístupu na konkrétní URI lze získat informace dle standardů RDF, SPARQL, ...
- Na ostatní data se odkazuje také přes HTTP URI

Při zohlednění všech těchto principů lze získat maximálně propojené informace a jak už bylo u RDF zmíněno, ve výsledku vznikne ohromný graf vzájemně propojených informací, ve kterém je možné následně najít cenné informace i o objektech, které na první pohled nemusí mít mezi sebou nic společného.

3.1.3 Java

Při vývoji je použit programovací jazyk Java. Použití javy vychází už z požadavků ze zadání. Hlavní důvody, proč je zde java vhodná a proč byla i jedním z požadavků jsou následující:

- Rozšířenost javy ve světě Linked Data
- Snadná integrace kvalitních knihoven pro práci s RDF
- Jednoduše nasaditelná aplikace přes webový archiv (war soubor)
- Výkon

Rozšířenost javy ve světě Linked Data je opravdu velká. Důvodem, proč je tento fakt zmíněn ve výhodách, je možnost případné snadné integrace dalších systémů pro vývojáře z tohoto oboru. Důležitým důvodem je zmíněný výkon aplikace. Na ten mají vliv jak knihovny pro práci s RDF, tak ale i samotný typ aplikace. Server si může uchovávat předzpracované šablony, patterny regulárních výrazů a další informace přímo v paměti, tudíž klient ve výsledku dostane výsledek rychleji, než by tomu bylo při použití například PHP nebo podobných jazyků.

3.1.4 Apache Jena

Pro práci s RDF daty byla zvolena knihovna Apache Jena. Tato knihovna podporuje práci přímo s RDF serializacemi a díky procesoru ARQ také práci se SPARQL dotazy. Další možností pro použití byla knihovna Sesame. Knihovny si jsou velice podobné. Obě mají dobrou dokumentaci i velmi dobře zpracované

ukázky, jak lze knihovny využívat. Rozhodujícím faktorem při výběru byl výkon.

Při výkonostním srovnání byl zohledněn publikovaný test *The Berlin SPARQL Benchmark* [8]. Jena a Sesame se dle něj výkonostně velmi liší. Sesame je několikanásobně rychlejší na dotazování, ale Jena naopak s velkým náskokem vede při načítání souborů v Turtle formátu.

Výkonostně má vliv načítání definic objektů, které se uchovávají v turtle formátu. Jedná se o proces, který bude spuštěn při zapnutí aplikace a při případném novém načtení (reloadu) definic (v případě nahrání definice přímo do filesystému dle požadavku). Rychlost SPARQL dotazů tedy není v tomto případě tolik důležitá. Při požadavku klienta na vygenerování objektu už se s RDF definicí nijak nepracuje - není potřeba se nad definicemi dále dotazovat. U typů SPARQL Endpoint/Construct také nedochází k dotazovaní se nad lokálně uloženými daty.

3.1.5 Další Java knihovny

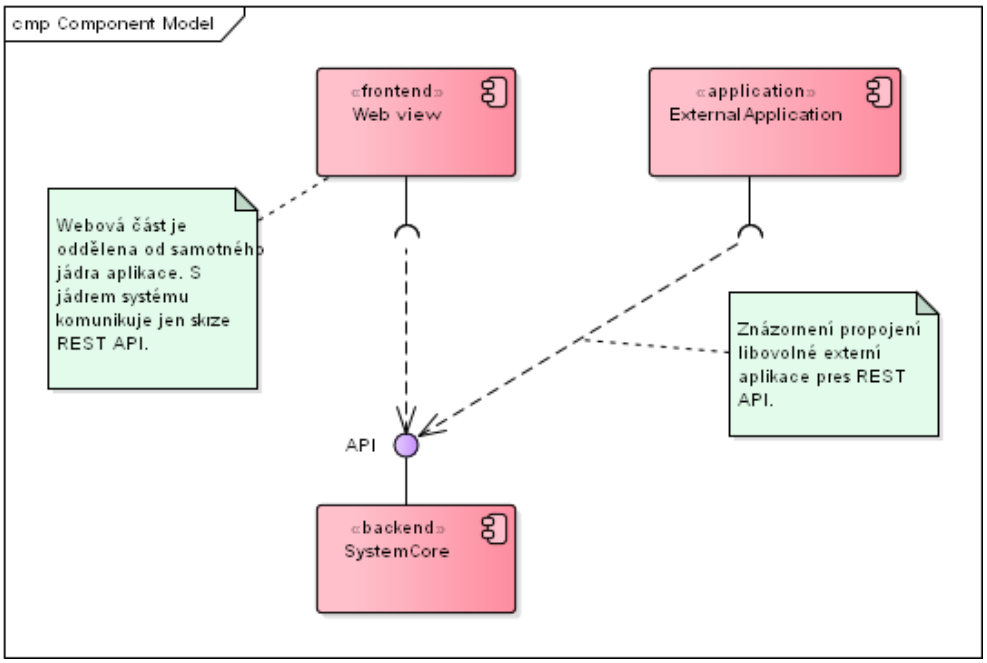
- Jersey, GSON - pro práci s definicemi objektů v rámci API
- Log4j - logovací systém
- Typesafe - konfigurace
- JUnit - testování

3.1.6 Maven

Závislosti, sestavení a testování se provádí přes build manažer Maven. Maven byl vybrán jako standard pro většinu Java aplikací. Vývoj taktéž probíhal v IDE IntelliJ IDEA, které obsahuje velmi dobrou integraci tohoto systému.

3.1.7 AngularJS

Součástí celého systému je i webová aplikace pro správu definic objektů. Zde je použit framework AngularJS, který je snadno napojitelný na REST API serveru.



Obrázek 3.1: Dvě základní části systému (jádro systému a webová, nebo jakákoliv jiná aplikace napojená na API)

3.2 Návrh architektury

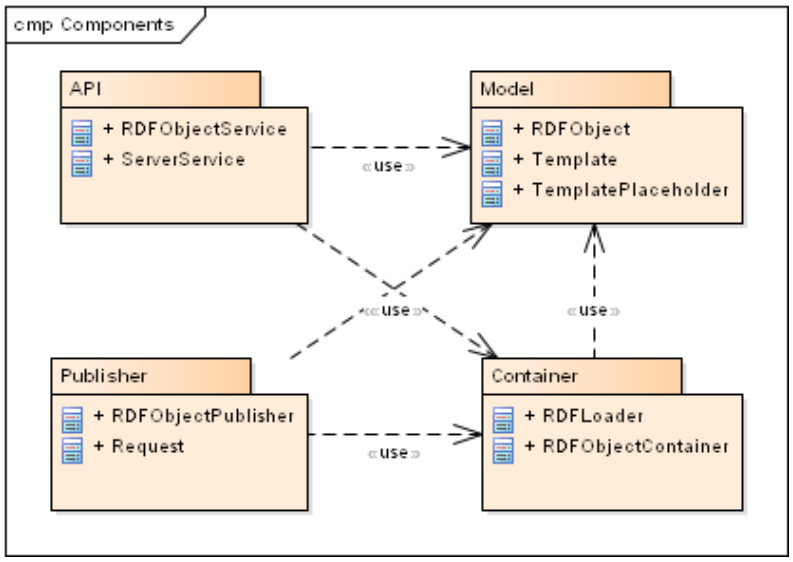
Na diagramu 3.1 lze vidět, že je systém rozdělen do 2 částí. První částí je webová aplikace, která slouží pro ulehčení administrace definic objektů. Hlavní funkcionalita tohoto serveru ale není touto částí ovlivněna, na API se lze napojit i z jiných aplikací. Tato část je napsaná ve frameworku AngularJS která využívá serverové API. U této aplikace se dá mluvit o známé MVC[3] architektuře, kde je ovšem Model (zde definice objektu) primárně součástí druhé části systému.

Druhou částí je serverová část, dále také pod názvem *Jádro systému*. Architektura této části je velice podobná MVC architektuře. O čistém MVC nelze mluvit z toho důvodu, že je zde velká provázanost komponent a není až tak striktně oddělena zodpovědnost jednotlivých částí.

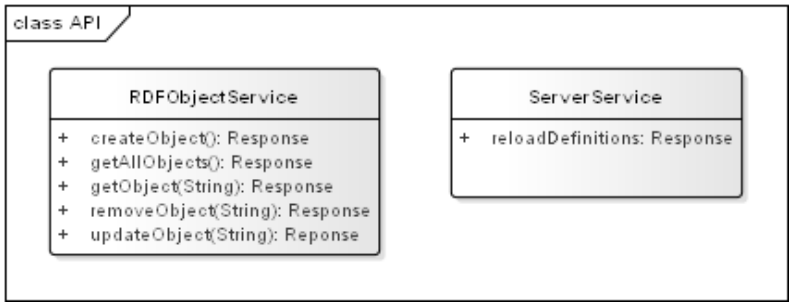
3.2.1 Jádro systému

Jádro systému se skládá ze čtyř hlavních komponent, které lze také vidět na obrázku 3.2. Pro administraci definic objektů slouží komponenta API. Další komponentou je Model, který obsahuje třídy využívané všemi komponentami. O načítání, ukládání a přístup k definicím objektů se stará komponenta Con-

[3] Model-View-Controller



Obrázek 3.2: Základní čtyři komponenty jádra systému



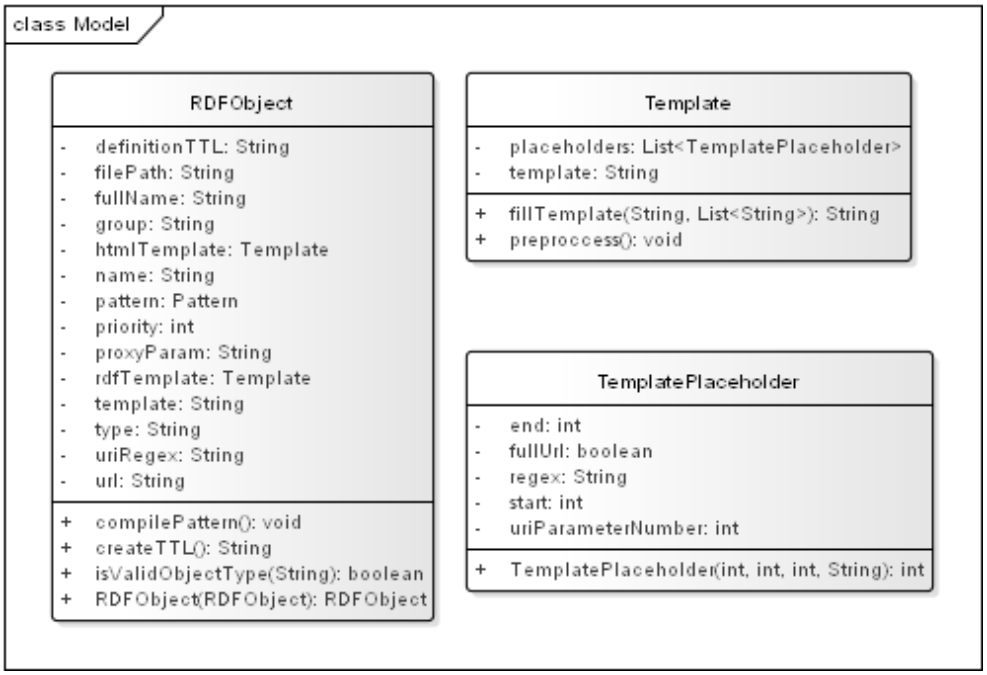
Obrázek 3.3: Třídy reprezentují služby, které mají metody vystavené pro API

tainer. Poslední základní komponentou je Publisher, který zpřístupňuje celý systém klientovi v podobě vygenerovaných objektů.

3.2.1.1 Komponenta API

Tato komponenta slouží pro administraci definic objektů. Jedná se o komponentu, ke které má přístup pouze administrátor systému. Obsahuje dvě třídy, které mají své metody vystavené pro napojení přes API.

Tyto třídy jsou zobrazeny na obrázku 3.3. Třída RDFObjectService slouží k samotnému vytváření, úpravě a mazání definic. ServerService pak obsahuje metodu, která se dá taktéž volat přes API a slouží ke znovunačtení všech definic z nastaveného uložiště.



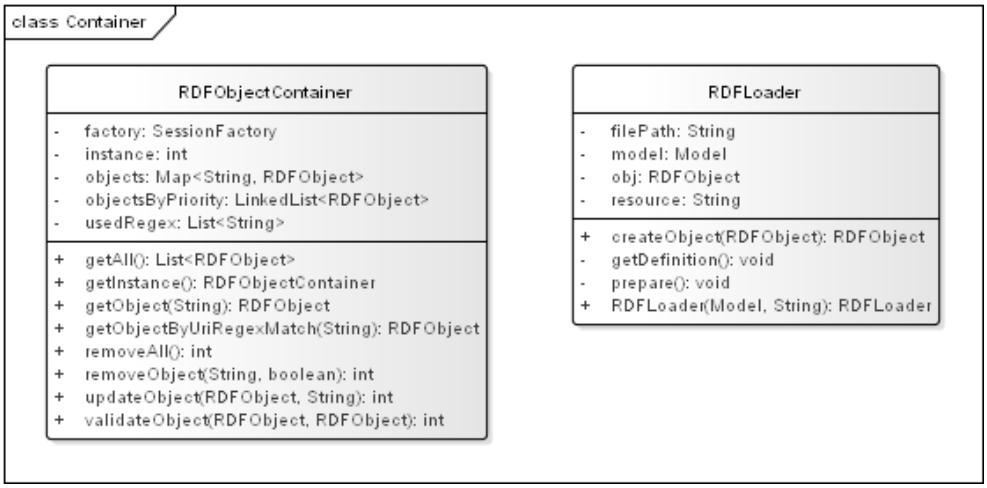
Obrázek 3.4: Třídy v komponentě Model. Pro přehled jsou uvedeny pouze nejdůležitější metody. Třídy jinak obsahují i další metody, převážně settery a gettery.

3.2.1.2 Komponenta Model

Tato komponenta obsahuje tři základní třídy, které jsou využívány zbytkem aplikace. Jedná se o třídy RDFObject, Template a TemplatePlaceholder, které jsou zobrazeny na obrázku 3.4.

RDFObject reprezentuje samotnou definici objektu. Obsahuje dva atributy typu Template. Prvním z nich je šablona definice (RDF serializace, nebo SPARQL dotaz) a druhou je HTML šablona sloužící pro zobrazení objektu v HTML formátu. Dále obsahuje metody jako jsou například validace a kompilace patternu regulárního výrazu pro pozdější identifikace definice dle požadavku klienta. Obsahuje také gettery a settery pro atributy, které ale nejsou na zmíněném diagramu vidět pro jejich primitivnost.

Třídy Template a TemplatePlaceholder zastřešují celý šablonovací systém. Třída Template obsahuje dvě důležité metody pro přdzpracování šablony a následné vyplnění šablony při požadavku. Předzpracování šablony probíhá tak, že se v šabloně naleznou všechny placeholdery a uloží se do seznamu pro pozdější vyplnění. Předzpracování probíhá pouze jendou při nahrátí definic (při startu serveru nebo znovunačtení definic). Cílem takto předzpracované šablony je urychlení generování objktů tak, aby se pouze dosazovaly hodnoty a případně aplikovaly podporované regulární výrazy.



Obrázek 3.5: Komponenta container obsahuje třídy pro administraci objektů

3.2.1.3 Komponenta Container

O správu nahraných definic se stará komponenta Container. Z obrázku 3.5 je patrné, že obsahuje 2 třídy.

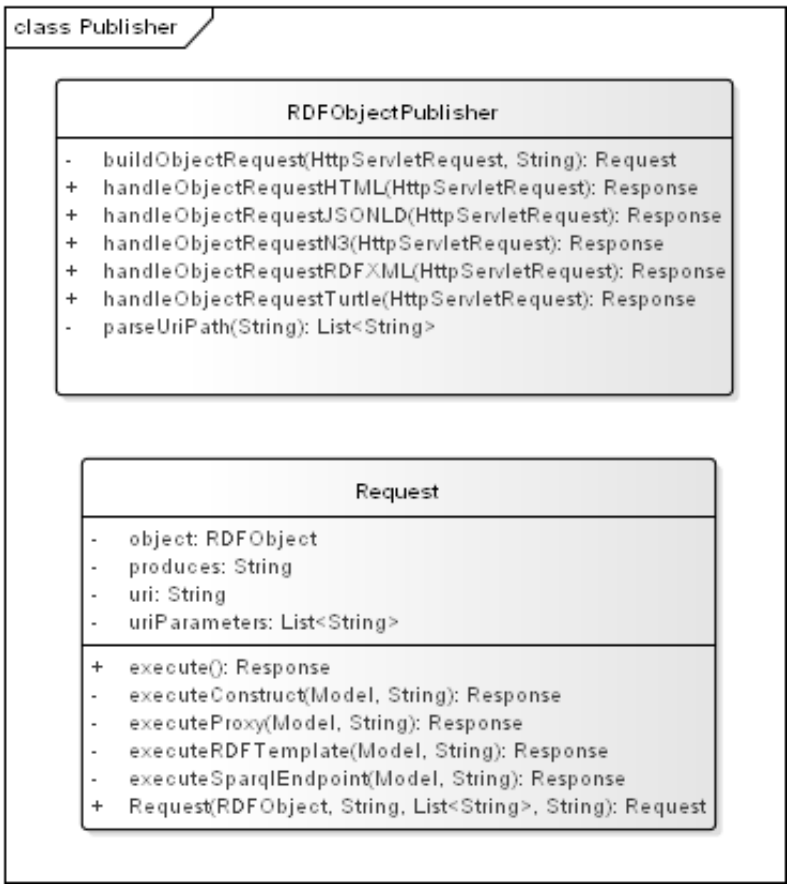
Třída RDFLoader se stará o čtení jednotlivých definic z file systému, validaci a následnou transformaci definice v Turtle formátu do objektu RDFObject, který reprezentuje konkrétní definice.

Třída RDFContainer slouží jako kontejner (malá databáze) všech nahraných objektů. Objekty si drží jak podle celého jména definice pro rychlý přístup k definicím pro API, tak i podle priorit, podle kterých dochází k vyhledávání definic při požadavku klienta. Kontejner se v aplikaci vyskytuje díky použití singleton patternu pouze jeden a je nejvytíženějším objektem v celé aplikaci, protože je využíván všemi komponentami.

3.2.1.4 Komponenta Publisher

Tato komponenta slouží k odbavování požadavků klienta na vygenerování objektu. Na obrázku 3.6 lze vidět 2 třídy, které mají za úkol interakci s klientem.

Třída RDFObjectPublisherService zpracovává prvotní požadavek klienta metodou *handleObjectRequest*(). V rámci systému je tato metoda definována pro všechny podporované serializace výstupu (HTML, RDF/XML, Turtle, JSON-LD a N-Triples). Při zavolání těchto metod je dále vytvořen objekt třídy Request, kterému je předána zodpovědnost za vygenerování výsledného objektu.



Obrázek 3.6: Komponenta container obsahuje třídy pro administraci objektů

3.3 Identifikace objektů, URL

V kontextu RDF se dají objekty identifikovat pouze jedním způsobem, a to URL adresou. V tomto případě ale URL zastává ještě jednu důležitou úlohu. Vzhledem k tomu, že je jediným spojením mezi objektem (a jeho definicí) a vnějším prostředím, tak musí nést i informace, ze kterých bude později vygenerován konkrétní objekt. Návrhu struktury URL byla proto věnována velká pozornost.

3.3.1 Struktura URL v.1

Prvotní návrh byl takový, že se URL rozdělí na následující 3 části:

- Hostname

  Touto částí se rozumí identifikace serveru a protokolu, například *https://dynrdf.com*. Z pohledu objektu slouží jen jako část identifikátoru.

- Identifikace objektu - první část cesty za hostname

  Tato část slouží k identifikaci objektu. Jedná se o unikátní název pro každý objekt. Příkladem může být například objekt časového intervalu s URL začínající *https://dynrdf.com/time-interval*, kde *time-interval* identifikuje tento objekt.

- Parametry objektu

  Zbývající část URL nese informace, které se dosadí do šablon jednotlivých definic objektů. Jednotlivé parametry jsou vždy odděleny lomítkem. Například pro objekt ročního intervalu by mohla URL vypadat následovně *https://dynrdf.com/time-interval/2015/2016*, kde by zvolené roky znamenaly parametry *od* a *do*.

Tento návrh by byl naprosto dostačujícím pro generování objektů všech podporovaných typů. Nicméně s sebou nese dva velké problémy.

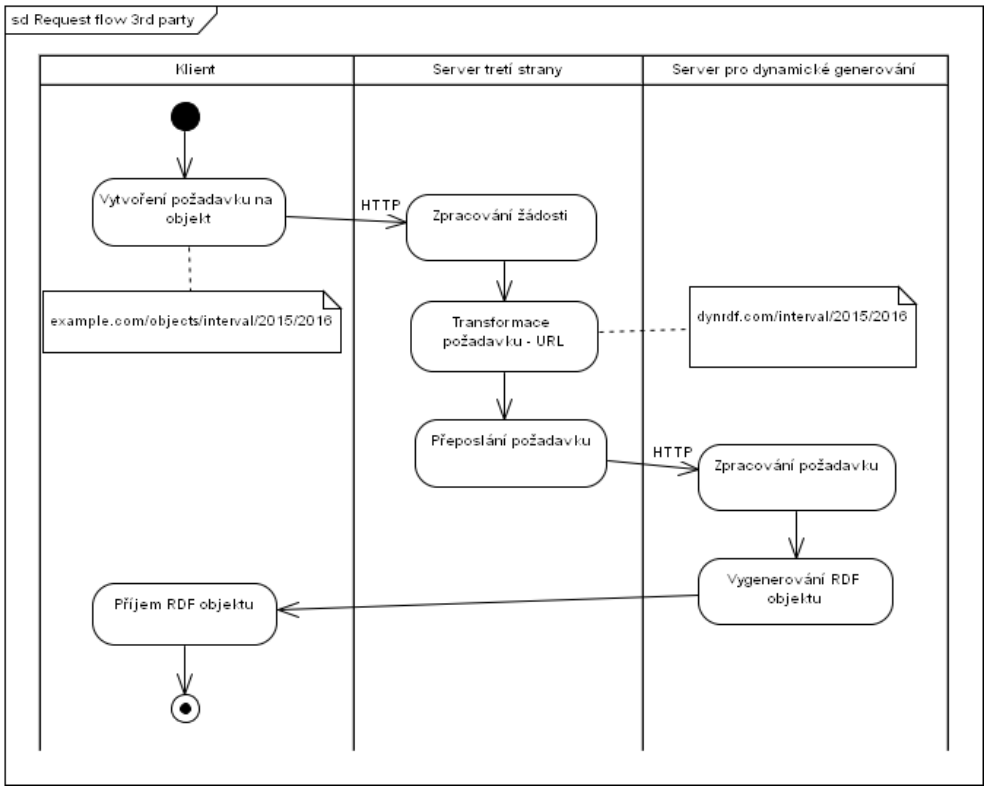
Prvním problémem je ten, že by musel být každý objekt identifikován IP adresou nebo doménou, kde tento server bude spuštěn. To by tedy znamenalo, že by tu nebyla možnost určit pro každý objekt extra URL v identifikátoru. Pokud by například funkce tohoto serveru chtěly využívat dvě různé společnosti, které by svoje objekty dále publikovali, tak by mohlo docházet k nekonzistenci odkazů na objekty. Jinými slovy se dá také zeptat na to, proč by nějaká společnost chtěla publikovat data, kde jejich identifikace není spojena například s názvem společnosti. Vždy by v identifikaci figuroval tento server, který ale s původem dat nemá nic společného. Cílem tohoto systému je pouze generovat objekty a identifikaci a vlastnictví ponechat na autorovi daných definic objektů.

Druhý problém, nebo spíše nepříjemnost nastává v případě, že by se servery společností využívající tohoto dynamického generování chovali jako prostředníci a požadavky na objekty přeposílaly na tento server. Pokud by každá definice měla vždy konkrétní unikátní identifikátor v URL adrese, pak by se pro každý takový objekt muselo přidávat pravidlo ve webových serverech na přeposílání požadavku a případně i dál složitě parsovat URL z požadavku na primárním serveru na URL, která by byla dle zmíněných specifikací.

Pro lepší představu, jak může vypadat požadavek na konkrétní objekt přes prostředníka, třetí stranu, slouží obrázek 3.7. Při požadavku přímo na tento server z klienta se proces na jiném serveru přeskočí. Ani jeden z těchto způsobů právě není bez nějakého výše zmíněného problému.

3.3.2 Struktura URL v.2

Výše zmíněný návrh by znamenal téměř nepoužitelnost tohoto systému. Bylo proto nutné navrhnout jiné řešení. To se od původního návrhu liší na dvou místech, v předání informací o objektu a identifikací objektů.



Obrázek 3.7: Požadavek na objekt přes server třetí strany. Prvotní návrh struktury URL adresy.

3.3.2.1 Informace o objektu v GET parametru

Předání dat o objektu v GET parametru naprosto jednoduše řeší problém s parsováním URL adresy na serverech třetích stran. V případě, kdy na jiný server přijde požadavek na nějaký objekt, tak stačí v konfiguraci webových serverů nastavit pouze jedno pravidlo pro přeposlání požadavku pro všechny objekty.

3.3.2.2 Identifikace objektů, regulární výraz

Předání informací o objektu v GET parametru má ale za následek, že URL adresa už neobsahuje identifikátor objektu jako první část cesty. Objekt se dá identifikovat pouze z informací v GET parametru, pro který účelově není stanovena žádná struktura.

Strukturu URL adresy v GET parametru, která má identifikovat konkrétní objekt, zná pouze autor definice. Ideálním nástrojem, jak docílit mapování URL na konkrétní objekt je zde regulární výraz, který popisuje konkrétní strukturu požadavku.

3.3.3 Shrnutí

Druhým návrhem struktury URL adresy se docílilo toho, že tento server bude schopný generovat objekty, jejichž identifikátor (URL) nebude závislý na adrese, kde tento server bude spuštěn.

Příkladem může být například adresa *http://dynrdf.com/?url=http://dataowner.com/objects/year/2015*. Původní požadavek mohl přijít na server, který je uveden v parametru *url*. Tento požadavek byl následně přesměrován na tento server pouze pomocí jednoho pravidla, kde se za parametr dosadila původní URL adresa požadavku. Následně se vygeneruje objekt roku 2015, jehož identifikátorem je obsah url parametru.

3.4 Definice objektů

Každý objekt, který má být dynamicky generován tímto systémem, musí být nadefinován administrátorem a následně uložen. Základními atributy, které daná definice musí obsahovat jsou:

- název

  Název slouží k identifikaci objektu v rámci seznamu.

- skupina

  Skupina zde označuje například název společnosti, jméno autora, nebo jiný identifikátor autora dané definice objektu. Společně s názvem tvoří plně kvalifikované jméno definice.

- URL regex

  Tento regulární výraz slouží k identifikaci konkrétní definice.

- priorita objektu

  Priorita ovlivňuje pořadí definic, v jakém se pokouší systém najít shodu regulárního výrazu s příchozí URL. Některé regulární výrazy moho popisovat více objektů. Pokud je nějaký výraz více specifičtější než jiný, tak se zvýšením priority dané definice dosáhne k požadované shodě. Priorita se určuje celým číslem. Čím menší číslo, tím větší priorita a tím dříve se bude snažit systém najít shodu s příchozí adresou právě u dané definice.

- typ definice

  Typ definice označuje způsob zadání a vygenerování výsledného objektu. Konkrétním typům dle požadavku 2.3.1.5 se tento text věnuje dále.

- šablona objektu

  Šablonou se zde rozumí text vyplněný placeholdery, do kterých se při generování doplní hodnoty. Šablonou může být SPARQL dotaz nebo kterákoliv podporovaná RDF serializace.

- HTML šablona

  Jedná se o podobnou šablonu jako pro konkrétní objekt, která je určena pro zobrazení ve webových prohlížečích.

U některých typů jsou definovány výjimky, nebo další povinné atributy. Tyto specifikace jsou zmíněny dále u popisu těchto typů.

Definice objektů přes Turtle formát musí splňovat RDF schéma, které je dostpné v přiloženém souboru *object.rdfs*. Ekvivalentní definice v JSON formátu a názvy atributů jsou uvedeny v souboru *object.json*.

3.4.1 RDF serializace

Definice objektu, který je definovám šablonou v RDF formátu, musí obsahovat všechny zmíněné povinné atributy. Do RDF šablony jsou doplněny při generování parametry z URL a vyplněná šablona je poté už výsledným požadovaným objektem.

3.4.2 SPARQL Construct

Šablonou pro tento typ je SPARQL construct dotaz, kde se jednotlivé parametry bindují pomocí funkce *BIND()*. Tento typ definice díky jazyku doplňuje RDF serializace o další funkcionality tohoto dotazovacího jazyka.

Pro vygenerování objektu jsou zde zapotřebí 2 kroky. Dosazení parametrů do šablony jako v případě RDF serializace a následně spuštění SPARQL dotazu lokálně, který vytvoří požadovaný objekt.

3.4.3 SPARQL Endpoint

SPARQL je navržen pro dotazování se nad datasety. Tato data jsou dostupná přes služby běžící na SPARQL protokolu. Tento typ tedy slouží ke konstrukci komplikovanějších objektů, jejichž atributy mohou být výsledkem dalších SPARQL dotazů nad konkrétním datasetem. Výsledkem tohoto typu je RDF dokument construct nebo describe dotazu.

Oproti lokálnímu construct dotazu se tento dotaz vykonává na jiném serveru. Proto je pro tento typ definice dalším povinným atributem URL adresa endpointu.

3.4.4 Proxy

Server má fungovat také jako prostředník mezi klientem a jinou aplikací poskytující RDF objekty. Jinou aplikací se rozumí webová služba, na kterou se bude požadavek přeposílávat. Tyto aplikace nemusí implementovat překlad objektů do jiných RDF serializací. Překlad do požadovaných serializací funguje zde na serveru stejně jako pro ostatní objekty.

Pro tento typ definice není potřeba definovat šablonu objektu, ale dalšími povinnými atributy jsou:

- proxy URL

  Jedná se o URL webové služby pro předaní zodpovědnosti za vygenerován objektu.

- název GET parametru

  Příchozí URL s informacemi o objektu je také přeposlána na danou službu. Názvem GET paramteru se rozumí parametr, do kterého se dosadí tato URL.

3.5 Šablonovací systém

Šablonovací systém je jedním ze stavebních kamenů této práce. Vyplněním šablony vzniká buď už konkrétní objekt, nebo SPARQL dotaz pro vygenerování objektu. Cílem návrhu tohoto systému je jednoduchost, ale zároveň dostatečná funkcionalita pro generování objektů různými způsoby.

3.5.1 Placeholder

Data o objektech jsou do šablon dosazeny přes placeholdery. Placeholderem se rozumí textový objekt, který definuje místo v dokumentu pro dosazení parametrů a má takovou strukturu, aby ho nebylo možné zaměnit s částí textu která nemá být nahrazena. V kontextu této práce se bude placeholderem rozumět textový objekt v tomto formátu:

$$[@ < d > [, < regex >]] \tag{3.1}$$

Placeholder se skládá ze dvou částí. První povinnou částí je parametr *@<d>*, který určuje konkrétní parametr URL adresy objektu, který se místo placeholderu dosadí. Jedná se tedy o *vstup* do placeholderu. Druhým nepovinným parametrem je regulrní výraz, který se může aplikovat před dosazením textu na vstup placeholderu.

3.5.1.1 Vstup placeholderu

Jak bylo zmíněno v kapitole o identifikaci objektu 3.3, každý objekt je definován URL adresou, která je serveru předána GET parametrem. Pomocí informací, které obsahuje tato adresa, je potřeba vygenerovat konkrétní objekt.

URL adresa je pro vstup do placeholderů rozdělena na části mezi lomítky. Na tyto části se dá odkazovat v placeholderu následujícím způsobem. Jako příklad budou uvedeny reference k adrese *http://intervals.com/year-interval/2013/2016*.

- @0 - reference na celou adresu (*http://intervals.com/year-interval/2013/2016*)
- @1, @2, ... (kladné čísla za znakem @) jsou reference na pozice mezi lomítky URL adresy.

  @1 = *http:*

  @2 = prázdný řetězec (mezi //)

  @3 = *intervals.com*

  @4, @5, @6 postupně *year-interval*, *2013* a *2016*

Tento způsob dává autorům šablon celkem snadný způsob, jak přímo z URL adresy dosadit do šablony požadované informace. Nicméně né všechny URL identifikátory objektů nemusí mít takovouto strukturu, aby se dalo snadno referencovat na konkrétní atributy objektu. Pokud by například v uvedeném příkladu nebyly roky rozděleny lomítkem, ale byly by v jednom parametru jako text „*2013-2016*“, nedal by se tento atribut rozdělit v RDF šabloně. Musel by se využít SPARQL construct dotaz, protože SPARQL obsahuje i funkce pro práci z řetězci (regulární výrazy). Proto je v placeholderu jako druhým, nepovinným parametrem regulární výraz.

3.5.1.2 Regulární výraz v placeholderu

Pro možnost extrahování pouze části hodnoty parametru v šabloně slouží nepovinný atribut regulárního výrazu. Tento regulární výraz podporuje Capturing groups [9]. A to způsobem, kdy je ze vstupu placeholderu extrahována hodnota, která se nachází ve skupině číslo 1.

Pokud by se autor šablony chtěl referencovat například na atribut *@5* který by obsahoval„*2013-2016*“, použil by jako placeholdery [@5, “(\d+)-”] pro *2013*, resp. [@5, “-(\d+)”] pro *2016*. Autoři šablon nebudou tedy nuceni používat SPARQL pro případy, kdy by potřebovali získat pouze část atributu, případně část z celé URL.

3.6 API

Závěr

Literatura

[1] W3C: *HTTP/1.1: Content Negotiation [online].* [cit. 2016-04-16]. Dostupné z: `https://www.w3.org/Protocols/rfc2616/rfc2616-sec12.html`

[2] Datahub: *data.gov.uk Time Intervals [online].* [cit. 2016-04-16]. Dostupné z: `https://datahub.io/tr/dataset/data-gov-uk-time-intervals`

[3] Epimorphics: *Using Interval Set URIs in Statistical Data [online].* [cit. 2016-04-16]. Dostupné z: `http://www.epimorphics.com/web/wiki/using-interval-set-uris-statistical-data`

[4] W3C: *SPARQL Query Language for RDF [online].* [cit. 2016-04-16]. Dostupné z: `https://www.w3.org/TR/rdf-sparql-query/`

[5] DuCharme, B.: *Learning SPARQL-Querying and Updating with SPARQL 1.1.* O'Reilly Media, 2011.

[6] W3C: *RDF 1.1 Turtle [online].* [cit. 2016-04-19]. Dostupné z: `http://www.w3.org/TR/2014/REC-turtle-20140225/`

[7] Tim Berners-Lee: *Linked Data - Design Issues [online].* [cit. 2016-04-21]. Dostupné z: `https://www.w3.org/DesignIssues/LinkedData`

[8] Christian Bizer, Andreas Schultz: *The Berlin SPARQL Benchmark [online].* [cit. 2016-04-22]. Dostupné z: `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.161.8030&rep=rep1&type=pdf`

[9] Jan Goyvaerts: *Regular Expression Reference: Capturing Groups and Backreferences [online].* [cit. 2016-04-17]. Dostupné z: `http://www.regular-expressions.info/refcapture.html`

Příloha **A**

Seznam použitých zkratek

RDF Resource description framework

URI Uniform Resource Identifier

URL Uniform Resource Locator

Příloha **B**

Obsah přiloženého CD

object.rdfs

- `readme.txt` ..... stručný popis obsahu CD
- `exe` ..... adresář se spustitelnou formou implementace
- `src`
  - `impl` ..... zdrojové kódy implementace
  - `thesis` ..... zdrojová forma práce ve formátu LaTeX
- `text` ..... text práce
  - `thesis.pdf` ..... text práce ve formátu PDF
  - `thesis.ps` ..... text práce ve formátu PS