



**MATEMATICKO-FYZIKÁLNÍ  
FAKULTA**  
Univerzita Karlova

## **BAKALÁŘSKÁ PRÁCE**

Kristýna Lhoťanová

# **Webová aplikace pro vyhledávání receptů**

Katedra softwarového inženýrství

Vedoucí bakalářské práce: doc. Mgr. Martin Nečaský, Ph.D.

Studijní program: Informatika

Studijní obor: Programování a vývoj software

Praha 2022

Prohlašuji, že jsem tuto bakalářskou práci vypracoval(a) samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů. Tato práce nebyla využita k získání jiného nebo stejného titulu.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V ..... dne .....

Podpis autora

Poděkování.

Název práce: Webová aplikace pro vyhledávání receptů

Autor: Kristýna Lhořanová

Katedra: Katedra softwarového inženýrství

Vedoucí bakalářské práce: doc. Mgr. Martin Nečaský, Ph.D., Katedra softwarového inženýrství

Abstrakt: Cílem této práce je vyvinout webovou aplikaci pro vyhledávání receptů založenou na agregaci datových sad z existujících webových stránek s recepty a jejich obohacení o data ze znalostních grafů. Znalostní grafy byly zastoupeny projekty DBpedia a Wikidata, z nichž byla získána data o ingrediencích a kategoriích jednotlivých receptů. Data byla extrahována s pomocí knihovny Apify a v dokumentovém modelu uložena do databázového systému Apache CouchDB. Aplikace uživateli poskytuje různé možnosti filtrování výsledků včetně fasetového vyhledávání, k čemuž využívá platformu Apache Solr. Zaměřuje se zejména na vyhledávání dle ingrediencí. Jedná se o tzv. single-page aplikaci implementovanou pomocí JavaScriptové knihovny React pro uživatelské rozhraní a frameworku Express.js na straně serveru. Obě části aplikace jsou psány staticky typovaným jazykem TypeScript a komunikují spolu prostřednictvím REST API.

Klíčová slova: webová aplikace, recept, znalostní graf, propojená data

Title: Web application for searching recipes

Author: Kristýna Lhořanová

Department: Department of Software Engineering

Supervisor: doc. Mgr. Martin Nečaský, Ph.D., Department of Software Engineering

Abstract: Abstract.

Keywords: web app, recipe, knowledge graph, linked data

# Obsah

<b>Úvod</b>	<b>2</b>
<b>1 Analýza</b>	<b>5</b>
1.1 Požadavky aplikace . . . . .	5
1.2 Dostupné datové sady . . . . .	5
1.2.1 Recipe1M+ . . . . .	5
1.2.2 Open Recipes . . . . .	5
1.2.3 Food.com Recipes and Interactions . . . . .	6
1.2.4 FoodKG . . . . .	6
<b>2 Architektura řešení</b>	<b>8</b>
2.1 Příprava dat . . . . .	8
2.2 Databázový model . . . . .	8
2.3 Backend . . . . .	8
2.4 Frontend . . . . .	8
2.5 Aplikační logika . . . . .	8
<b>3 Implementace návrhu</b>	<b>9</b>
3.1 Zpracování vstupních dat . . . . .	9
3.2 Databáze Apache CouchDB . . . . .	9
3.3 Vyhledávání pomocí Apache Solr . . . . .	9
3.4 Rozhraní REST API . . . . .	9
3.5 Middleware . . . . .	9
3.6 Single-page aplikace . . . . .	9
<b>4 Testování</b>	<b>10</b>
<b>Závěr</b>	<b>11</b>
<b>Seznam použité literatury</b>	<b>12</b>
<b>Seznam obrázků</b>	<b>13</b>
<b>Seznam tabulek</b>	<b>14</b>
<b>Seznam použitých zkratk</b>	<b>15</b>
<b>A Přílohy</b>	<b>16</b>
A.1 První příloha . . . . .	16

# Úvod

Vyhledávání relevantního obsahu je spolu s elektronickou komunikací jednou z klíčových funkcí internetu. S rostoucím množstvím dostupných informací se filtrování nalezených výsledků stává stále obtížnějším. Tvůrci webových stránek se často zaměřují spíše na uživatelsky přívětivé interaktivní rozhraní, zatímco optimalizace strojového vyhledávání jde stranou. Pro webové vyhledávače, jmenovitě např. Google, Bing nebo Yahoo, je pak náročné analyzovat obsah těchto stránek po sémantické stránce a tedy vyhodnotit, zda obsahují užitečné informace k zodpovězení dotazu uživatele.

V reakci na tuto problematiku vznikl tzv. *Sémantický Web* neboli Web dat jakožto rozšíření původního Webu dokumentů, tak jak jej známe z platformy *World Wide Web*. Sémantický Web lze vnímat jako globální databázi, nad kterou se lze pomocí speciálního jazyka *SPARQL* dotazovat podobně jako nad tradičními databázovými systémy. Data jsou poskytována v různých serializacích formátu RDF a mohou být přímo vnořena do HTML dokumentů nebo zpřístupněna v samostatných souborech. Tato strukturovaná data nazýváme *propojená* (v originále *Linked Data*). Umožňují snadnější hledání souvislostí mezi entitami z různých zdrojů na základě společných slovníků neboli ontologií [1].

V posledních letech termín Sémantický Web ustupuje do pozadí a často je místo něj zmiňován tzv. *znalostní graf* (anglicky *Knowledge Graph*). Začátky fenoménu znalostních grafů bychom mohli datovat do roku 2012, kdy společnost Google představila svůj znalostní graf pro vyhledávání obsahu na webu. K technologii znalostních grafů se brzy poté přihlásily další velké společnosti včetně firem Microsoft, IBM, Facebook, LinkedIn, Amazon, eBay, Airbnb nebo Uber. Grafový model totiž oproti tradičnímu relačnímu modelu nabízí flexibilnější správu dat z oblasti sociálních sítí, dopravních spojení, bibliografických citací a řady dalších odvětví. Výše zmíněné příklady znalostních grafů všechny spadají do kategorie komerčních znalostních grafů, které jsou určeny pro interní využití v rámci dané firmy. Protikladem jim jsou otevřené znalostní grafy poskytující data k volnému využití všem uživatelům internetu. Nejvýznamnějšími představiteli otevřených znalostních grafů jsou aktuálně DBpedia, Wikidata, Freebase a YAGO [2]. První dva zmíněné projekty si představíme v této práci a integrujeme je s aplikací na vyhledávání receptů.

Oblast gastronomie je rozvěž vhodným kandidátem k zapojení do sítě znalostních grafů a propojených dat. Pro tvůrce webových aplikací je poměrně jednoduché publikovat obsah svých stránek ve formátu strukturovaných dat. Vhodným způsobem je např. vložení RDF reprezentace daných entit (receptů, uživatelů, recenzí) ve formátu JSON-LD<sup>1</sup> přímo do hlavičky jednotlivých HTML dokumentů. V takovém případě je žádoucí použít existující ontologie raději než definovat vlastní, byť by mohly být lépe strukturované a uzpůsobené dané doméně. Využití standardizovaných slovníků usnadňuje webovým vyhledávačům interpretaci stránky a je větší šance, že se aplikace dostane na vyšší příčky vyhledávaných výsledků.

Cílem této bakalářské práce je prozkoumat možnosti využití otevřených dat v doméně receptů, propojit je s daty publikovanými na různých webových strán-

---

<sup>1</sup>Koncovka *LD* v názvu JSON-LD odkazuje na pojem Linked Data.

kách shromažďujících recepty a prezentovat tyto výsledky uživateli ve formě vlastní webové aplikace. Zároveň v rámci této aplikace poskytnout užitečné možnosti filtrování agregovaných výsledků včetně fasetového vyhledávání. Proces sběru, konverze a uložení dat by měl být co nejvíce automatizovaný a snadno zreprodukovatelný. Práce se nevěnuje přidávání nových receptů prostřednictvím uživatelského rozhraní. Existujících webové stránky totiž obsahují velké množství dat, které lze díky bohaté historii v podobě hodnocení a recenzí lépe filtrovat. Navíc by bylo potřeba se vypořádat s automatickou kalkulací nutričních hodnot receptu z obsažených surovin, přičemž ne všechny ingredience dokážeme automaticky identifikovat a získat jejich nutriční hodnoty. V budoucnu by funkce nahrávání nových receptů měla být přidána spolu s více lokalizacemi aplikace, registrací uživatelů a celkovou personalizací obsahu pro přihlášené uživatele.

## Volba tématu

Příprava jídla je tématem každodenního života a na webových stránkách, které se této oblasti věnují, má velmi silnou komunitu. Většina z nás se chystání domácích pokrmů z ekonomických důvodů nevyhne, takže se hodí mít po ruce sadu receptů pro inspiraci. Typicky máme na recepty různé požadavky - někdo preferuje rychlejší postup, jiný se dívá po ceně ingrediencí nebo nutričních hodnotách. Občas dostaneme chuť na recept z řecké nebo italské kuchyně a jindy zkrátka chceme experimentovat a najít recept kombinující našich 5 oblíbených surovin. Některé ingredience z receptu nám mohou být neznámé, nebo si jen podle názvu nejsme jistí, zda máme na mysli tu správnou. V takovém případě musíme stránku s receptem opustit a dodatečné informace k ingredienci vyhledat jinde, pokud na ně aplikace přímo neodkazuje. Zde je příležitost zapojit otevřená data a namapovat názvy ingrediencí na jejich odpovídající entity ve znalostních grafech. Data pak můžeme začlenit do aplikace a nabídnout uživateli informace nad rámec samotného receptu, např. popisy a glykemické hodnoty surovin, ilustrační obrázky a podobně. Také můžeme identifikovat ingredience a tranzitivně recepty ze stejných kategorií. Oproti původní datové sadě tak vytvoříme nové vazby a poskytneme uživateli rozmanitější filtrování výsledků.

Doména receptů navíc poskytuje spoustu prostoru pro zajímavá rozšíření se zapojením moderních technologií. Uplatnění by zde našlo například počítačové vidění s rozpoznáváním obrázků. S dostatečně velkou databází bychom díky němu mohli analyzovat fotografii hotového pokrmu a nalézt příslušný recept. Uspadnili bychom tak uživateli práci v situacích jako je návštěva restaurace, při které návštěvníkovi zachutnalo servírované jídlo a chtěl by si jej později připravit v domácích podmínkách. Dalším uplatněním strojového učení by mohlo být vyhledávání na základě příkazů v přirozeném jazyce. Namísto zdlouhavého zadávání nejrůznějších filtrů by stačilo aplikaci položit dotaz: „Jaké recepty z italské kuchyně mohu vyrobit z kuřete, rajčat a parmazánu?“. Této problematice se věnuje například projekt FoodKG konstruující nad recepty a ingrediencemi znalostní graf [3]. Ruku v ruce s touto funkcionalitou jde hlasové zadávání, které by se hodilo zapojit nejen ve fázi vyhledávání receptů, ale také například pro hands-free ovládání aplikace. Uživatel by měl možnost diktovat příkazy k přečtení části receptu, pokud zrovna pracuje na jeho přípravě a nemá volné ruce k listování obsahem. Využití by našlo i populární *full-text* vyhledávání, pomocí kterého lze snadno objevit

recepty na základě klíčových slov v popisku receptu, postupu či recenzích. V komerční sféře by se nabízelo propojení s online supermarkety, konkrétně zrychlení nákupu pomocí vyhledávání surovin k vybranému receptu. S tímto konceptem již na svých stránkách pracuje firma rohlik.cz, nabídka receptů a možnosti filtrování jsou ale omezené. Nepochybně by se hodilo integrovat také doporučovací systém pro ještě snadnější nalezení relevantních výsledků. Aplikace má velký prostor pro škálování objemu dat, přičemž datasety mohou být následně použity jako podklad pro strojové učení.



# 1. Analýza

V této kapitole si zadefinujeme požadavky na funkcionalitu naší aplikace. Také se v kontextu požadavků podíváme na existující webové stránky s recepty a provedeme diskuzi nad jejich funkcemi, možnými vylepšeními a rozšířeními. Následně si rozebereme různé alternativy dostupných datových sad a srovnáme jejich výhody i nevýhody vzhledem k požadavkům aplikace.

## 1.1 Požadavky aplikace

## 1.2 Dostupné datové sady

### 1.2.1 Recipe1M+

V první fázi analýzy se zaměříme na veřejně dostupná zdrojová data s recepty, která by mohla posloužit jako podklad pro naši databázi. Jedním z nejdůležitějších projektů v této oblasti je *Recipe1M+*, strukturovaný korpus obsahující přes 1 milion receptů a 13 milionů souvisejících obrázků jídla. Aktuálně se jedná o největší veřejně dostupnou sadu receptů. Dataset je dostupný pouze přihlášeným uživatelům z ověřené organizace a je povoleno jej využívat výhradně pro účely studia a výzkumu. Z celkového počtu 1 milionu receptů obsahuje 50 000 receptů s nutričními informacemi [4]. V naší aplikaci preferujeme nutriční hodnoty zahrnout, pokud jsou dostupné na zdrojové stránce receptu. Měli bychom tedy k dispozici 50 000 dokumentů s touto informací. Ostatní data jsou určena přednostně pro strojové zpracování prostřednictvím trénování modelů.

Celková velikost datové sady se pohybuje v řádu stovek gigabytů, samotné JSON dokumenty se strukturovanými recepty z adresáře `layers` se ale vejdou do 2 GiB, tudíž by byly vhodné pro potřeby této práce limitované omezenou výpočetní kapacitou. Lze odtud využít 1 029 720 receptů obsahujících název, url, ingredience a postup přípravy. Odkazy na ilustrační fotografie jsou u 402 760 z těchto receptů. Pro příjemnější uživatelský zážitek se omezujeme pouze na recepty s obrázky, takže jsme z datasetu *Recipe1M+* schopni použít přibližně 400 000 receptů, pokud akceptujeme absenci nutričních hodnot. Bylo by spíše obtížnější z tohoto datasetu identifikovat názvy ingrediencí, neboť jsou suroviny uloženy včetně jejich množství a jednotek měření v rozmanitém formátu.

### 1.2.2 Open Recipes

Dalším významným aktérem na poli volně dostupných receptů je iniciativa *Open Recipes*. Autoři Finkler, Shiflett a Birkebæk projekt představují jako otevřenou databázi záložek s recepty. Pojem záložky je použit z důvodu absence instrukcí k přípravě receptu. Dataset má sloužit pouze k vyhledání receptu a pro detailní informace má být uživatel přesměrován na zdroj s kompletním receptem [5]. Tohoto přístupu úspěšně využívají některé z vyhledávačů receptů, např. populární aplikace *SuperCook*. Naše aplikace si ale klade za cíl zpracovat i stránky s detaily receptů, ze kterých lze dále pokračovat na detaily ingrediencí s informacemi ze znalostních grafů. Projekt *Open Recipes* tedy pro náš scénář nebude

vhodnou volbou.

### 1.2.3 Food.com Recipes and Interactions

Rozsáhlý dataset *Food.com Recipes and Interactions* s téměř 200 000 recepty extrahovanými z webové stránky *Food.com* (původního GeniusKitchen) je publikován na portálu *Kaggle*, který shromažďuje podklady pro strojové učení. Datová sada pokrývá 18 let interakce uživatelů včetně hodnocení, počtu recenzí i konkrétních reakcí [6]. Kromě základních informací obsahuje také nutriční hodnoty receptů, datum publikování a rovněž normalizovaná jména ingrediencí. Ta byla získána parsováním originálního textu surovin, kvůli čemuž nejsou vždy zcela spolehlivě přesná (např. ve jménech často zůstala jednotka měření z původního textu). Unikátních ingrediencí je k dispozici kolem 8 000, což by měl být dostatečný základ pro hledání linků s entitami otevřených znalostních grafů. Zároveň ve srovnání s předchozími projekty nabízí nejbohatší informace k jednotlivým receptům.

Nevýhodou datasetu je jeho primární určení pro strojové zpracování. Byl vytvořen jako podklad pro generování personalizovaných receptů na základě dřívějších preferencí uživatele [7]. Syrová data nejsou zamýšlena pro přímou prezentaci, což se negativně odráží na jejich přesnosti a estetice. Slova jsou občas zařazena do špatných kategorií a problematický je zejména plně *lowercase* formát textu, ze kterého nejsme schopni zpětně zrekonstruovat originální text receptu. Dataset bychom tedy nemohli použít samostatně, ale pouze v kombinaci s vlastní extrakcí dat, která by respektovala velikost písma a lépe se vypořádala s parsováním jednotlivých kategorií.

Tento problém je poměrně snadno řešitelný díky struktuře stránky Food.com. Z id receptu lze jednoduše složit url ve formátu `www.food.com/recipe/id` a navíc aplikace podporuje koncept propojených dat, tedy poskytuje recepty ve strukturovaném RDF formátu. Do HTML hlaviček všech dokumentů s recepty vkládá JSON-LD serializaci dle ontologie *Schema.org*. Z připraveného datasetu bychom tedy mohli využít identifikátory receptů a normalizované ingredience, pro každý recept extrahovat jeho JSON-LD a spojit informace dohromady. Zároveň bychom si ušetřili práci s převáděním receptů do JSON-LD formátu a připravené soubory rovnou vložili do hlaviček dokumentů. Nevytvářeli bychom nové entity receptů, pouze bychom změnili prezentační vrstvu RDF dat. Identifikátory entit v podobě IRI by tedy zůstaly nezměněné.

### 1.2.4 FoodKG

Přímo v oblasti znalostních grafů figuruje projekt *FoodKG*, který je postaven nad sadou receptů z již zmíněného datasetu Recipe1M+. Recepty doplňuje o podrobnější data k ingrediencím ze stránky The Cook's Thesaurus a definuje vlastní ontologii. Model ontologie je navržen pro zodpovídání dotazů na recepty dle ingrediencí s přihlédnutím k individuálním potřebám uživatele, jako jsou alergie a intolerance na různé složky potravin.

Vývojáři projektu zpřístupňují skripty k extrakci dat z encyklopedie The Cook's Thesaurus a k vytvoření znalostního grafu. Neposkytují ale žádné nové recepty nad rámec datové sady Recipe1M+, naší horní hranicí by tedy bylo 50 000

receptů s nutričními hodnotami. Ontologie publikovaná na webových stránkách projektu obsahuje 75 entit ingrediencí, které kromě obecného popisu poskytují informace o glykemickém indexu, obsahu lepku a možných náhradách dané ingredience. Výhodou je připravený RDF formát, nad kterým se lze snadno dotazovat pomocí jazyka SPARQL. Autoři Chen a kol. uvádějí ukázky dotazů, vyberme například dotaz vracející recepty, které obsahují banán a zároveň neobsahují vlašské ořechy *FoodKG*:

```
@PREFIX food: <http://purl.org/heals/food/>
@PREFIX ingredient: <http://purl.org/heals/ingredient/>
SELECT DISTINCT ?recipe
WHERE {
    ?recipe food:hasIngredient ingredient:Banana .
    FILTER NOT EXISTS {
        ?recipe food:hasIngredient ingredient:Walnut .
    }
}
```

## 2. Architektura řešení

2.1 Příprava dat

2.2 Databázový model

2.3 Backend

2.4 Frontend

2.5 Aplikační logika

## **3. Implementace návrhu**

**3.1 Zpracování vstupních dat**

**3.2 Databáze Apache CouchDB**

**3.3 Vyhledávání pomocí Apache Solr**

**3.4 Rozhraní REST API**

**3.5 Middleware**

**3.6 Single-page aplikace**

## 4. Testování

# Závěr

# Seznam použité literatury

- [1] The World Wide Web Consortium. Semantic Web, 2015.
- [2] Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d’Amato, Gerard de Melo, Claudio Gutiérrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan F. Sequeda, Steffen Staab, and Antoine Zimmermann. *Knowledge Graphs*. Number 22 in Synthesis Lectures on Data, Semantics, and Knowledge. Morgan & Claypool, 2021.
- [3] Ching-Hua Chen, Deborah L. McGuinness, Nidhi Rastogi, Oshani Seneviratne, Sola Shirai, Ananya Subburathinam, and Mohammed J. Zaki. Foodkg: A semantics-driven knowledge graph for food recommendation.
- [4] Javier Marin, Aritro Biswas, Ferda Ofli, Nicholas Hynes, Amaia Salvador, Yusuf Aytar, Ingmar Weber, and Antonio Torralba. Recipe1M+: A Dataset for Learning Cross-Modal Embeddings for Cooking Recipes and Food Images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2019.
- [5] Ed Finkler, Chris Shiflett, and Andreas Birkebæk. Open Recipes.
- [6] Shuyang Li. Food.com Recipes and Interactions, 2019.
- [7] Bodhisattwa Prasad Majumder, Shuyang Li, Jianmo Ni, and Julian McAuley. Generating Personalized Recipes from Historical User Preferences. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5976–5982, Hong Kong, China, November 2019. Association for Computational Linguistics.



# Seznam obrázků

# Seznam tabulek

# Seznam použitých zkratek

## A. Přílohy

### A.1 První příloha