Západočeská univerzita v Plzni
Fakulta aplikovaných věd
Katedra informatiky a výpočetní techniky

Bakalářská práce

Automatické získání historických údajů z webových zdrojů

Plzeň 2015 Gabriela Hessová

Prohlášení

Prohlašuji, že jsem bakalářskou práci vypracovala samostatně a výhradně s použitím citovaných pramenů.

V Plzni dne 22. června 2015

Gabriela Hessová

Poděkování

Rády bych poděkovala Ing. Richardu Lipkovi, Ph.D. za vstřícnost, trpělivost, cenné rady a věcné připomínky, které mi pomohly tuto bakalářskou práci vypracovat.

Abstract

The topic of this bachelor thesis is automated retrieval of large amount of historical data from web sources and their subsequent transformation into a form usable by other applications which can visualize the information obtained. The theoretical part deals with methods of data retrieval and contains an overview of electronic sources, then deeply describes one of this sources - Wikipedia, The Free Encyclopedia. The practical part describes an implementation of a tool, which transforms data from Wikipedia dump to the final form. The tool focuses on data related to people.

Abstrakt

Předmětem této bakalářské práce je automatické získání většího množství historických údajů z webových zdrojů a jejich následné přetvoření do podoby využitelné aplikacemi, které získané informace vizualizují. Teoretická část se zabývá metodami získávání dat z textu a přehledem elektronických zdrojů, dále pak popisuje vybraný elektronický zdroj - Wikipedii, otevřenou encyklopedii. Praktická část popisuje implementaci nástroje, který data z dumpu Wikipedie automaticky transformuje do konečné podoby. Práce se zaměřuje na data týkající se osob.

Obsah

1 Úvod

V dnešní době je nejvýznamnějším prostředkem pro získávání informací internet. Cílem této práce je prozkoumat dostupné elektronické zdroje údajů spolu s možnostmi jejich získávání a dalšího využití a vytvořit nástroj, který umožní automatizované získání velkého množství dat za účelem dalšího zpracování, konkrétně by se měl stát zdrojem dat pro grafovou databázi a nad ní postavené vizualizační nástroje.

V první části práce budou shrnuty obecné metody získávání dat z textu a popsány některé elektronické zdroje informací.

Ve druhé části pak bude vybrán a prozkoumán jeden z těchto zdrojů a bude navržen a popsán nástroj, který bude jím poskytovaná data s co nejmenším zásahem uživatele převádět do jiné podoby. Uživatel bude moci volit parametry zpracovávaných dat prostřednictvím konfigurovatelného uživatelského prostředí. Nakonec bude z požadovaného množství dat vytvořena grafová databáze.

2 Získávání údajů z textu

2.1 Rozdělení dat z hlediska strukturovanosti

Data na internetu mohou být různého charakteru. Z hlediska získávání údajů pro další zpracování je důležitým faktorem míra jejich uspořádanosti. Podle ní můžeme data rozdělit do tří kategorií:

- nestrukturovaná data - jedná se o běžné texty v přirozeném jazyce, nemají definovaný datový model a nejsou žádným způsobem organizovány. Mezi charakteristické rysy patří nejednoznačnost a nepravidelnost, kdy data stejného významu mohou mít odlišnou reprezentaci i v rámci stejné domény.
- strukturovaná data - jsou přehledně organizována a formátována podle pevně stanoveného schématu tak, aby s nimi bylo možné jednoduše manipulovat a dále je zpracovávat. Nejčastější případy zahrnují databáze, tabulkové procesory a soubory s pevně stanoveným formátem (např. logovací soubory). [1]
- semistrukturovaná data - mají strukturu, která se může nepredikovatelným způsobem měnit. Příkladem mohou být metadata[1], datové sklady[2], bioinformatické databáze nebo soubory ve formátu XML. [1]

Zařazení vyjmenovaných typů dat není deterministické, míra uspořádanosti se případ od případu liší.

2.2 Technologie pro získávání údajů z textu

Přestože toto tvrzení není podloženo žádnými seriózním průzkumy, mnoho zdrojů (např. Gartner[2] nebo Merril Lynch[3]) udává, že více než 80% všech informací na internetu je nestrukturovaných. Z toho důvodu je třeba nalézt

[1] data o datech

[2] neboli DWH (Data Warehouse), zvláštní typ databáze umožňující analytické dotazování nad rozsáhlými soubory dat

metody, které budou v těchto neuspořádaných datech hledat vzorce umožňující získání jejich významu.

2.2.1 Data mining

Existuje několik technologií, které se právě těmito metodami zabývají. Jednou z nich je tzv. data mining (v českém překladu nepříliš používané dolování dat). Data mining je metoda získávání netriviálních skrytých a potenciálně užitečných informací z dat. Definice data mining podle autorů knihy Data Mining, Practical Machine Learning Tools And Techniques[4] zní následovně:

Data mining je proces objevování vzorců v datech. Tento proces musí být automatizovaný nebo poloautomatizovaný (častěji). Smysl nalezených vzorců spočívá v poskytnutí nějakého užitku, obvykle ekonomického. Zpracovávaná vstupní data musí být zastoupena v dostatečném množství.

Data mining je součástí tzv. procesu dobývání znalostí z databází (*Knowledge Discovery in Databases*).

2.2.2 Text mining

Text mining může být obecně definován jako proces založený na vědeckých znalostech, při kterém jsou zpracovávány větší kolekce dokumentů za užití různých analytických nástrojů. Obdobně jako data mining se i text mining snaží extrahovat užitečné informace z datových zdrojů prostřednictvím rozpoznávání vzorců.

V případě text miningu jsou datové zdroje představovány sbírkami dokumentů a vzorce nejsou nacházeny ve formalizovaných záznamech databází, nýbrž v nestrukturovaných textových údajích obsažených právě v těchto sbírkách.

Text mining odvozuje podstatnou část svého zaměření z klíčových výzkumů data miningu. Není tudíž překvapením, že tyto technologie vykazují mnoho podobností. Obě vyžadují předběžné zpracování, algoritmy na vyhledávání vzorců a prvky prezentační vrstvy, jako jsou různé vizualizační nástroje. [5]

2.2.3 NLP - natural language processing

S pojmem *text mining* úzce souvisí další pojem, a to zpracování přirozeného jazyka (*natural language processing*).

NLP je snaha o komplexní extrahování významu slov z textu. To si můžeme zhruba přeložit jako, co, kdo, kdy, kde, jak a proč dělal. NLP typicky využívá lingvistické koncepty jako *part-of-speech* (určování slovních druhů - podstatné jméno, sloveso, přídavné jméno atd.) a gramatickou strukturu, reprezentovanou buď větnými členy, nebo závislostmi (podmět, předmět). Musí se vypořádat s anaforou (opakováním slov) a mnohoznačností (ať už slov nebo gramatických struktur, jako je změna významu za použití jistého slova nebo předložkové fráze).

Za tímto účelem NLP využívá různé reprezentace znalostí jako lexikony slov a jejich významů a gramatických vlastností, sbírky gramatických pravidel a mnohé další zdroje jako ontologie[1] entit a akcí nebo tezaurus synonym a zkratek. [6]

Významným nástrojem pro NLP je například WordNet, rozsáhlá lexikální databáze anglických slov. Podstatná a přídavná jména, slovesa a příslovce jsou seskupována do sad kognitivních synonym (tzv. *synsetů*) , kdy každý vyjadřuje jiný koncept. *Synsety* jsou provázány prostřednictvím konceptuálně-sémantických a lexikálních vztahů. Výsledná síť smysluplně propojených slov a konceptů může být procházena pomocí prohlížeče. WordNet je také volně dostupný ke stažení. [7]

2.2.4 Analýza noisy textů

Analýza tzv. *noisy textů* je odvětví velmi podobné text miningu. Hlavním rozdílem je, že analýza noisy textů pracuje s textem, který vznikl jako produkt procesu extrakce textu z jiných médií než elektronického textu prostřednictvím transkripce nebo OCR [2].

[1] ontologie v informatice je výslovný a formalizovaný popis určité problematiky, obsahuje glosář (definici pojmů) a tezaurus (definici vztahů mezi jednotlivými pojmy)

[2] Optical character recognition - konverze textu vytištěného na papír do elektronické podoby

3 Elektronické zdroje historických údajů

Při zpracovávání většího množství dat je výhodné použít některý z prověřených elektronických zdrojů, které data shromažďují, nějakým způsobem spravují a poskytují nástroje pro práci s nimi.

3.1 Wikipedie

Snad nejznámějším a nejpoužívanějším zdrojem informací je Wikipedie, otevřená encyklopedie. V porovnání s dále uvedenými zdroji zahrnuje zdaleka největší objem dat.

Wikipedie je mnohojazyčná webová encyklopedie s otevřeným obsahem [8]. Na její tvorbě spolupracují dobrovolní přispěvatelé z celého světa. Většina článků může být editována každým, kdo k nim má přístup a řídí se několika základními pravidly. Wikipedie se řadí mezi deset nejoblíbenějších webových stránek světa. Je jedním z projektů Nadace Wikimedia [1], s nimiž je vzájemně provázána.

Data na Wikipedii jsou z větší části semistrukturovaná (texty článků), některé články obsahují strukturovaná data - tzv. infoboxy (tabulky obsahující základní údaje o subjektu). Jednotlivé články jsou provázány odkazy.

Wikipedie nabízí možnost stáhnout si její obsah a pracovat s ním offline. Data na Wikipedii využívají také některé další elektronické zdroje. Více informací o Wikipedii je v kapitole 4.

[1] WMF - Wikimedia Foundation - nezisková nadace, která spravuje projekty Wikipedie, Wikislovník, Wikicitáty, Wikiknihy, Wikizdroje, Wikimedia Commons, Wikizprávy a Wikiverzita

3.2 DBpedia

DBpedia je projekt pro extrahování strukturovaných informací z Wikipedie. Umožňuje klást nad daty z Wikipedie sofistikované dotazy. [9]

Data jsou uložena ve standardizovaném formátu RDF *(Resource Description Framework)*, aneb systému popisu zdrojů. Jeho účelem je popsat data tak, aby byla čitelná jak lidsky, tak strojově. Hlavní myšlenkou RDF je k popisovanému zdroji přiřadit výraz ve tvaru podmět – vlastnost – předmět (též subjekt – predikát – objekt). Pro tento výraz se také používá termín trojice (anglicky *triple*). RDF je zároveň grafový datový model, data jsou orientované označené grafy. Jedna hrana RDF grafu je označená trojicí. Trojice jsou organizovány do pojmenovaných grafů (čtveřic). Uzly, hrany a pojmenované grafy jsou označeny pomocí URI (*Unified Resource Identifier*). [11]

RDF je jednou ze tří základních technologií sémantického webu[1]. Dalšími jsou pak dotazovací jazyk SPARQL *(SPARQL Protocol and RDF Query Language*) a ontologický jazyk OWL (*Web Ontology Language)*. SPARQL je používán pro dotazování nad DBpedií. DBpedia pak poskytuje nástroje pro tvorbů dotazů (tzv. *Query Buildery*):

- OpenLink iSPARQL Visual Query Builder
- DBpedia Query Builder

DBpedia je základem pro mnoho dalších aplikací, většina jich je však ve fázi vývoje, např.:

- spacetime – engine pro vyhledávání a zobrazování, zatím pouze demo
- OpenLink Virtuoso built-in Faceted Browser - vyhledávač entit podle zadaného textu
- gFacet - vizualizační nástroj pro prohlížení RDF dat jakožto grafových struktur

DBpedia stejně jako Wikipedie poskytuje velké množství dat pro stažení. Mezi nevýhody tohoto projektu patří fakt, že stále prochází vývojem a webová služba často není dostupná.

[1] web, ve kterém jsou informace strukturovány a uloženy podle standardizovaných pravidel, což usnadňuje jejich vyhledání a zpracování

3.3 YAGO

YAGO (*Yet Another Great Ontology*) je rozsáhlá sémantická znalostní databáze (ontologie), která je odvozena z informačních zdrojů, jako jsou Wikipedie, WordNet nebo GeoNames. Vznikla v Institutu Maxe Plancka pro informatiku v Saarbrückenu. Obsahuje údaje o více než 10 milionech entit (osobách, organizacích, místech atd.) a více než 120 milionů faktů souvisejících s těmito entitami. Přesnost ontologie byla manuálně vyhodnocena a dosahuje 95%. Yago je ontologie, která je zakotvena v čase a prostoru, připojuje k údajům časovou a prostorovou dimenzi. [10]

Yago má stejný cíl jako DBpedie, a tím je přetvoření dat z Wikipedie do strukturované podoby. Projekty se však liší ve svém zaměření. Yago klade důraz na přesnost a taxonomickou strukturu.

Pracovat s ontologií YAGO lze buď online přes webové rozhraní nebo SPARQL interface, nebo offline stažením souborů v RDF formátu.

3.4 Freebase

Dalším elektronickým zdrojem dat je otevřená grafová databáze Freebase spravovaná společností Metaweb Technologies. Freebase pracuje s daty z Wikipedie, MusicBrainz a dalších zdrojů. Zahrnuje téměř 40 milionů témat (*topics*) představujících reálné subjekty jako lidi, místa, předměty atd. Tato témata pak tvoří jednotlivé uzly grafu. Ne každý uzel však musí být téma.

Mnohostrannou povahu některých témat pomáhá zachytit koncept typů (*types*). Jednomu tématu může být přiřazen libovolný počet typů. Např. tématu Bob Dylan přísluší typ zpěvák, textař, hudební skladatel, autor knihy atd. Každý typ pak s sebou nese příslušnou sadu vlastností (*properties*).

Tak jako jsou vlastnosti seskupeny do typů, typy samotné jsou seskupeny do domén. Každá doména má přiřazený identifikátor. Příkladem mohou být `/business`, `/film`, `/medicine`, `/music` ... [12]

Freebase poskytuje možnost rychlého procházení dat za použití vyhledávače, výsledkem hledání jsou však nestrukturované články. K získání strukturované informace slouží dotazovací jazyk MQL (*Metaweb Query Language*), který

pracuje s JSON[2] objekty a je v mnoha ohledech intuitivnější než SPARQL.

3.5 History World

HistoryWorld[13] je encyklopedie více než 10 tisíc světových událostí, umožňuje jednoduché vyhledávání a zobrazování událostí na časové ose. HistoryWorld vychází z informací z Encyclopedia Britannica. Obsahuje nestrukturovaná data. Citace více než 250 slov textu pro komerční účely je možná jen po domluvě s autory.

3.6 Ancient History Encyclopedia

Ancient History Encyclopedia[14] se zaměřuje na údaje o antické historii. Jedná se o neziskovou vzdělávací webovou stránku. Poskytuje vyhledávač, vizualizaci pomocí časových os, obrázky, videa a další. Jedná se o nestrukturovaná data. Veškerý původní obsah je dostupný pod licencí Creative Commons[3], která umožňuje jakékoli další použití a distribuci pro nekomerční účely. Ancient History Encyclopedia se nachází na seznamu OER[4] (Open educational resources) a také sdílí svůj obsah prostřednictvím univerzitní sítě Pelagios[5].

3.7 HyperHistory Online

HyperHistory Online[15] je součástí projektu The World History. Výsledky jsou zobrazovány na časové ose, detaily subjektů mohou být zobrazeny ve formě nestrukturovaných článků.

[2]JavaScript Object Notation - objektový formát zápisu dat nezávislý na platformě
[3]http://creativecommons.org/
[4]https://www.oercommons.org/
[5]http://pelagios-project.blogspot.cz/p/about-pelagios.html

3.8 Infoplease

Informační portál Infoplease[16] umožňuje vyhledávání podle klíčových slov. Představuje encyklopedii, slovník, atlas a další. Je součástí rodiny Family Education Network[6].

3.9 Encyclopedia.com

Encyclopedia.com[17] sdružuje data z encyklopedií jako The Columbia Encyclopedia, Oxford's World Encyclopedia nebo the Encyclopedia of World Biography. Umožňuje online vyhledávání. Veškerý obsah a nástroje spadají pod licenci High Beam Research[7] a jsou volně k použití pro nekomerční účely.

3.10 Encyclopedia Britannica

Encyclopedia Britannica[18] je aktualizovaná elektronická verze největší tištěné encyklopedie na světě. Informace v ní bývají měřítkem přesnosti článků různých elektronických zdrojů. Kopírovat, tisknout nebo stahovat obsah je možno pouze pro osobní, nekomerční použití.

3.11 Who's Who

Who's Who[19] je databáze obsahující krátké životopisy vlivných lidí v Británii. Je aktualizovanou elektronickou verzí publikace pocházející z roku 1849. Každý rok přibyde okolo 1000 nových záznamů. Obsah Who's Who spadá pod Oxford University Press. Pro přístup k datům je nutná registrace.

[6] http://fen.com/resources/agreeDisclaim.html
[7] http://www.highbeam.com/about-us

3.12 MusicBrainz

Jako poslední zdroj uvádím MusicBrainz[20], elektronickou databázi hudebníků, nebo přesněji řečeno všech lidí, kteří kdy měli co do činění s hudbou. Najdeme zde informace jak o Johannu Sebastianu Bachovi, tak např. o Jindřichu VIII. Výsledkem hledání na MusicBrainz jsou částečně strukturovaná data. U každého subjektu jsou vyplněny záznamy jako jméno, typ, pohlaví, oblast působení a začátek a konec působení. Součástí záznamu je také nestrukturovaný článek z Wikipedie. Data v MusicBrainz jsou pod licencemi Creative Commons - CC0 a Creative Commons Attribution-NonCommercial-ShareAlike 3.0.

V tabulce 3.1 je přehled elektronických zdrojů informací.

Název	**URL**	**Licence/společnost**
Wikipedie	www.wikipedia.org	CC BY-SA 3.0, GFDL
DBpedia	www.dbpedia.org	CC BY-SA 3.0, GFDL
YAGO	www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/	CC BY-SA 3.0,
Freebase	www.freebase.com	CC BY
History World	http://www.historyworld.net/	nenalezeno
Ancient History Encyclopedia	http://www.ancient.eu/	CC BY
HyperHistory	www.hyperhistory.com	nenalezeno
Infoplease	www.infoplease.com	FEN
Encyclopedia.com	www.encyclopedia.com	HighBeam Research
Encyclopedia Britannica	www.britannica.com	Encyclopedia Britannica
Who's Who	www.ukwhoswho.com	CC BY-SA 3.0, GFDL
MusicBrainz	https://musicbrainz.org	CC0, CC BY-SA 3.0

Tabulka 3.1: Přehled elektronických zdrojů informací

4 Wikipedie

Jako zdroj informací jsem si zvolila Wikipedii kvůli velkému objemu dat a vysoké přesnosti i úplnosti v porovnání s jinými elektronickými zdroji.

Obsah Wikipedie lze stáhnout ve formě tzv. data dumps[1] v 10 různých jazycích: angličtině, němčině, francouzštině, italštině, čínštině, japonštině, polštině, portugalštině, ruštině a španělštině.

Soubory, se kterými chceme dále pracovat, jsou ty s názvem ve tvaru *xx*`wiki-`*yyyyMMdd*`-pages-articles.xml.bz2`, kde *xx* je zkratka jazyka daného dumpu a *yyyyMMdd* datum (např. `enwiki-20141106-pages-articles.xml.bz2`). Tyto soubory obsahují pouze aktuální revize, žádné diskusní nebo uživatelské stránky. Nové verze vycházejí asi jednou až dvakrát za měsíc.

Revize anglické Wikipedie z 6. 11. 2014 zabírá v komprimované formě 10,5 GiB paměti, po rozbalení dostaneme jeden XML soubor o velikosti 46,7 GiB a 800 milionech řádků. V některých případech však lze pracovat i se samotným komprimovaným souborem. Soubor je možné stáhnout běžným způsobem nebo přes BitTorrent. Dále lze stáhnout následující soubory:

- pages-meta-current.xml.bz2 – všechny stránky (včetně diskusních), jen aktuální revize
- abstract.xml.gz – abstrakty stránek
- all-titles-in-ns0.gz – jen titulky stránek (s přesměrováním)
- SQL soubory pro stránky, odkazy
- Latest Dumps - všechny revize všech stránek – tyto soubory mohou mít až několik terabytů textu

4.1 Prohlížení obsahu offline

Pro prohlížení obsahu Wikipedie bez přístupu k internetu lze využít různých prohlížečů.

[1] http://meta.wikimedia.org/wiki/Data_dump_torrents#enwiki

4.1.1 BzReader

Hlavní a zároveň jediný účel aplikace BzReader[21] je prohlížení Wikipedie bez přístupu k internetu. Pracuje přímo s komprimovaným souborem typu `pages-articles.xml.bz2`, takže jednou z jeho výhod je úspora místa na disku. Převádí text Wikipedie do HTML. BzReader je volně k dispozici, je určen primárně pro operační systém Windows.

Po jeho instalaci je třeba nejdříve vytvořit indexy pro rychlý přístup k jednotlivým stránkám. Tato operace zabere několik hodin. Výsledkem je složka s názvem např. `enwiki-20141008-pages-articles.xml.idx`, která obsahuje mimo jiné soubor typu cfs (*Compact File Set*) o velikosti zhruba 1,33 GiB.

Dump Wikipedie pak lze prohlížet úplně stejně jako její webovou verzi (viz obr. 4.1). Stránky jsou opět provázány odkazy, neobsahují však obrázky a tabulky zvané infoboxy také nejsou správně zobrazeny. BzReader je nástroj určený čistě pro čtení, neobsahuje žádné další funkce pro práci s nalezenými výsledky.

Dostupný z: `https://code.google.com/p/bzreader/`

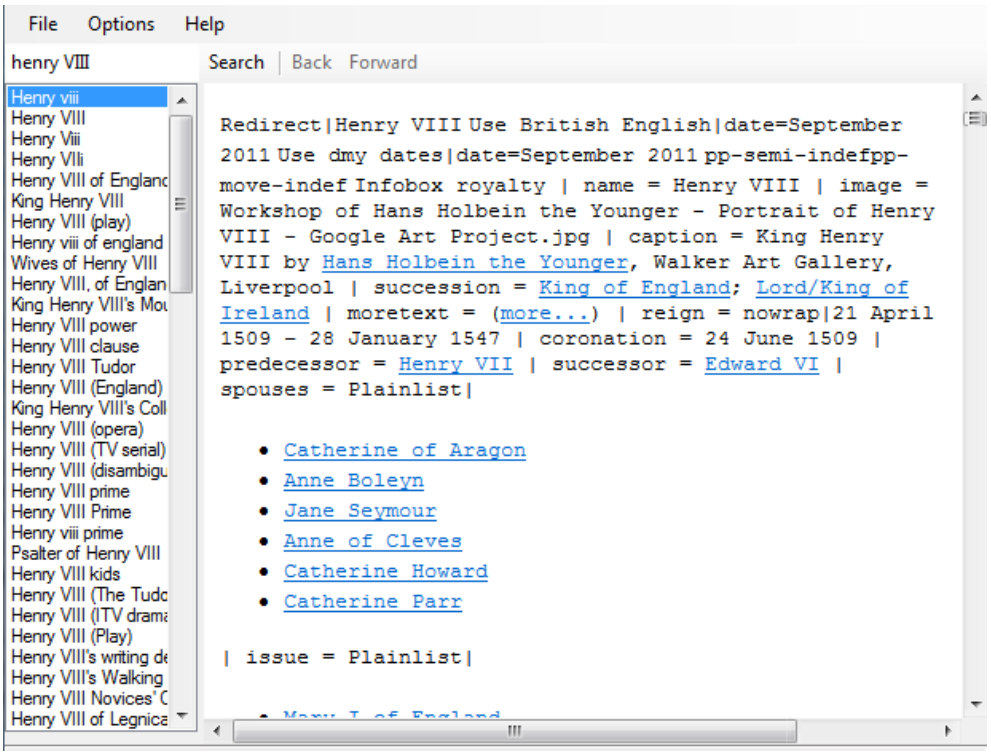
4.1.2 MzReader

MzReader je nadstavba BzReaderu, provádí propracovanější renderování textu do HTML, takže jsou jím vytvořené stránky lépe čitelné. Vyžaduje Microsoft Visual Basic 6.0 Runtime.

Dostupný z: `http://homepage.ntlworld.com/bharat.vadera/MzReader/`

4.1.3 Kiwix

Kiwix [22] je offline prohlížeč obsahu webových stránek. Jeho původní účel je zpřístupnit Wikipedii pro práci v režimu offline, ale je možné ho využít pro prohlížení jakýchkoli HTML stránek.



Obrázek 4.1: Prohlížeč BzReader

Jedním z rozdílů oproti BzReaderu je formát souboru, se kterým pracuje. Kiwix používá soubory ve formátu ZIM[23] (*Zeno IMproved*), což je vysoce komprimovaný otevřený formát s doplňujícími informacemi (metadaty).

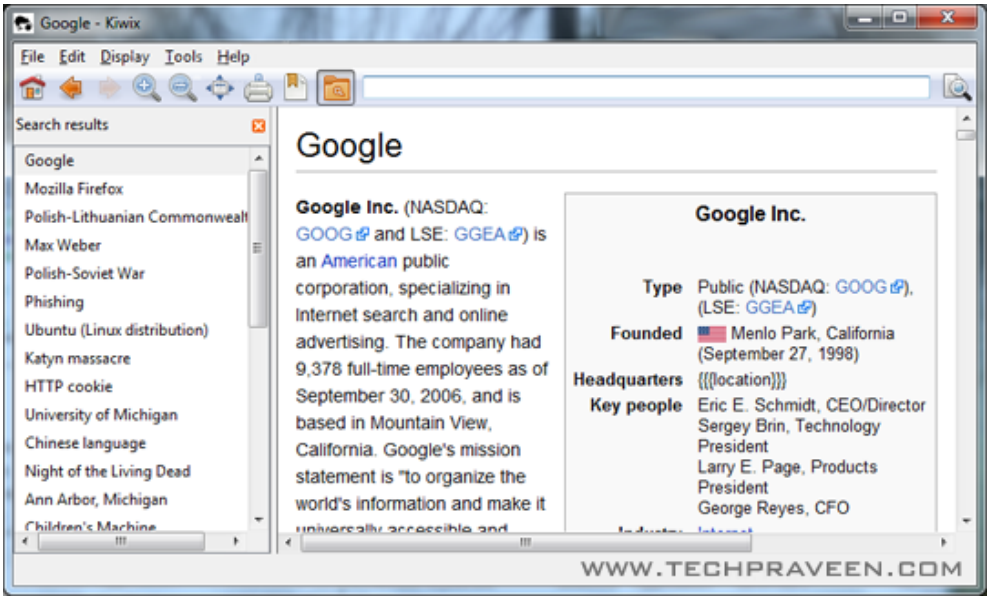
Dalším významným rozdílem je fakt, že Kiwix na rozdíl od BzReaderu poskytuje některé další funkce pro pohodlné používání:

- fulltextový vyhledávač
- záložky a poznámky
- HTTP server
- export do PDF/HTML
- uživatelské rozhraní ve více než 100 jazycích
- navigace
- integrovaný správce obsahu a nástroj pro stahování

Potřebné soubory lze stáhnout přímo z oficiálních stránek Kiwix[1].

Stránky zobrazované Kiwixem jsou nerozeznatelné od webové Wikipedie (viz obr. 4.2). Kiwix je dostupný pro Windows, Mac OS X, Linux i Android.

Dostupný z: `www.kiwix.org`



Obrázek 4.2: Prohlížeč Kiwix [24]

4.1.4 WikiTaxi

WikiTaxi[25] je prohlížeč pro všechna data ve formátu MediaWiki[2]. Umožňuje prohlížení stránek, jako jsou Wikipedie, Wikiquote nebo WikiNews. Nepodporuje prohlížení obrázků. Dovede pracovat s mnoha různými jazyky jako angličtinou, němčinou či turečtinou, problém nastává při práci s jazyky psanými zprava doleva. Je určený pro OS Windows.

Dostupný z: `www.wikitaxi.org`

[1] http://www.kiwix.org/wiki/Main_Page#Wikipedia_files nebo http://download.kiwix.org/zim/wikipedia/

[2] MediaWiki – engine všech projektů Wikipedia Foundation

4.2 Zpracování textového obsahu

Nástroje pro offline prohlížení Wikipedie neposkytují žádné funkce využitelné pro další práci s textem. Proto je třeba použít nějaký parser neboli syntaktický analyzátor, pomocí kterého získáme z daného xml souboru informace, které potřebujeme.

Na stránkách mediawiki.org je článek o několika alternativních parserech pro převedení textu v syntaxi používané MediaWiki do jiné podoby[3]. Jedná se však většinou o již opuštěné nebo příliš úzce zaměřené projekty, proto je nasnadě napsat si parser, který bude sloužit jen pro naše vlastní účely.

4.2.1 Struktura pages-articles.xml souboru

Soubor `enwiki-20141008-pages-articles.xml` zabírá 46,7 GiB na disku a na to je třeba při práci s ním myslet. Běžné programy jako Notepad nebo Internet Explorer nejsou schopné zobrazit jeho obsah, protože se snaží načíst celý soubor do operační paměti. Jednou z možností, jak si prohlédnout jeho vnitřní strukturu, je použít vestavěný prohlížeč programu Total Commander - Lister.

Soubor je ve formátu XML (*EXtensible Markup Language*), což je značkovací jazyk, jehož základem jsou elementy a atributy. XML soubory jsou textové soubory, používají kódování Unicode, obvykle UTF-8. Specifikace XML formátu je uvedena na stránkách w3schools[26].

Úvodní řádky souboru jsou zobrazeny na ukázce zdrojového kódu 4.1.

Soubor obsahuje kořenový element `mediawiki`. Následuje element `siteinfo`, který obsahuje informace o obsahu souboru. Odtud až do konce souboru už následují samotné stránky Wikipedie představující jednotlivé články, kterým odpovídají elementy `page`. Ty jsou na stejné úrovni jako `siteinfo`. Může se jednat jak o plnohodnotné stránky, tak o pouhé rozcestníky. Každý element `page` obsahuje element `revision` představující poslední revizi stránky.

[3] http://www.mediawiki.org/wiki/Alternative_parsers

```
<mediawiki xmlns="http://www.mediawiki.org/xml/export-0.9/" ...>
  <siteinfo>
    <sitename>Wikipedia</sitename>
    <dbname>enwiki</dbname>
    <base>http://en.wikipedia.org/wiki/Main_Page</base>
    <generator>MediaWiki 1.25wmf1</generator>
    <case>first-letter</case>
    <namespaces>
      <namespace key="-2" case="first-letter">Media</namespace>
                  ...
      <namespace key="2600" case="first-letter">Topic</namespace
          >
    </namespaces>
  </siteinfo>
  <page>
    <title>AccessibleComputing</title>
    <ns>0</ns>
    <id>10</id>
    <redirect title="Computer accessibility" />
    <revision>
```

Zdrojový kód 4.1: Úvodní řádky souboru `enwiki-20141008-pages-articles.xml`

```
<page>
    <title>AccessibleComputing</title>
    <ns>0</ns>
    <id>10</id>
    <redirect title="Computer accessibility" />
    <revision>
      <id>381202555</id>
      <parentid>381200179</parentid>
      <timestamp>2010-08-26T22:38:36Z</timestamp>
      <contributor>
        <username>OlEnglish</username>
        <id>7181920</id>
      </contributor>
      <minor />
      <comment>[[Help:Reverting|Reverted]] ...</comment>
      <text xml:space="preserve">#REDIRECT ...</text>
      <sha1>lo15ponaybcg2sf49sstw9gdjmdetnk</sha1>
      <model>wikitext</model>
      <format>text/x-wiki</format>
    </revision>
  </page>
```

Zdrojový kód 4.2: Struktura elementu `page`

`title`	název stránky
`ns`	namespace (jmenný prostor)
`id`	identifikační číslo stránky
`redirect`	tento element je přítomen pouze v případě, že se jedná o stránku typu přesměrování, obsahuje název stránky, na kterou přesměrovává
`revision`	poslední revize

Tabulka 4.1: Podelementy elementu `page`

`id`	identifikační číslo revize
`parentid`	rodičovské id
`timestamp`	časová značka
`contributor`	přispěvatel
`minor`	minor
`comment`	komentář
`text`	textový obsah stránky – to, co se zobrazuje
`sha1`	otisk z hashovací funkce
`model`	model, většinou `wikitext`
`format`	formát, většinou `text/x-wiki`

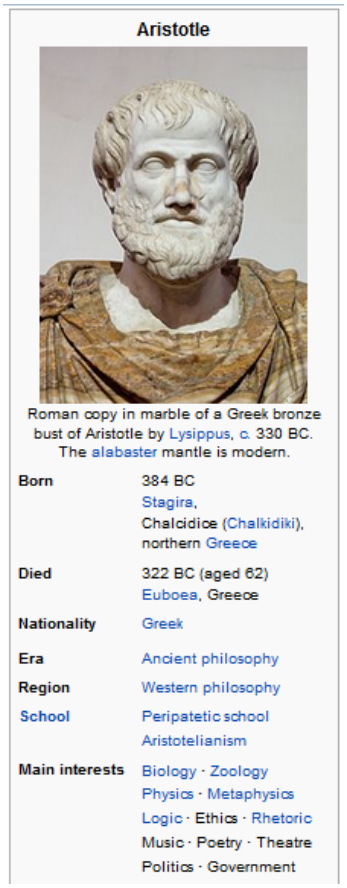
Tabulka 4.2: Podelementy elementu `revision`

Texty stránek, obsažené v elementech `page`, lze považovat za semistrukturovaná data - články v přirozeném jazyce jsou doprovázeny značkovacím jazykem - tzv. *Wiki markupem* [27], jehož specifikaci najdeme na oficiálních stránkách Wikipedie.

Kódování souboru je UTF-8.

4.2.2 Infoboxy

Některé články obsahují strukturované informace ve formě tabulek zobrazujících se většinou na pravé straně stránky – tzv. infoboxů. Souhrnné informace o podobě a funkci infoboxů najdeme na stránkách Wikipedie (viz odkaz [28]).



Obrázek 4.3: Infobox tak, jak je zobrazen na stránkách Wikipedie

```
{{Infobox philosopher
| name = Aristotle
| image = Aristotle Altemps Inv8575.jpg
| caption = {{longitem|line-height:1.25em|Roman copy in ...
| birth_date = {{BCE|384}} {{longitem|padding-top:0; ...
| death_date = {{nowrap|{{BCE|322}} (aged 62)&lt;br/&gt; ...
| nationality = [[Greeks|Greek]]
| era = [[Ancient philosophy]]
| region = [[Western philosophy]]
| school_tradition = {{Unbulleted list |[[Peripatetic ...
| main_interests = {{hlist|[[Biology]]|[[Zoology]]}} ...
| notable_ideas = {{Unbulleted list |[[Golden mean ...
| influences = {{hlist |[[Parmenides]] |[[Socrates]] |...
| influenced = {{longitem|Virtually all subsequent [[Western ...
}}
```

Zdrojový kód 4.3: Struktura infoboxu v textovém souboru

Infoboxy jsou součástí elementu `text`. Často následují bezprostředně po startovcí značce `<text>`.

Infobox obsahuje důležité informace, které jsou společné pro subjekty stejného typu. Například každá osoba má nějaké jméno a datum narození, u zvířat je zase uvedena vědecká klasifikace (rod, čeleď atd.). Údaje v infoboxu by měly být stručné, přesné, relevantní k subjektu a měly by již být obsaženy na jiném místě v článku.

Šablona infoboxu je ohraničena dvojicí otevíracích ({{) a uzavíracích (}}) složených závorek. Hlavička infoboxu je uvozena klíčovým slovem *Infobox*. Na stejné řádce, oddělen mezerou, následuje typ infoboxu. Existuje mnoho typů infoboxů; například co se týče osob, můžeme v souboru najít *person, royalty, officeholder, monarch, philosopher, scientist, writer, artist, musical artist, military person, prime minister* a mnoho dalších. Všechny typy infoboxů jsou specifikovány v poměrně rozsáhlém seznamu infoboxů[29]. V tomto seznamu jsou uvedeny i šablony jednotlivých infoboxů spolu s atributy, které jim přísluší.

Každému typu infoboxu přísluší pevně stanovená množina atributů, některé z nich jsou povinné a jiné volitelné, přičemž volitelné často nejsou vyplněny nebo úplně chybí. Atributy jsou představovány dvojicí *název atributu = hodnota*. Jednotlivé atributy jsou odděleny znakem '|'. Někdy bývá hodnota reprezentována seznamem (např. seznamem potomků). Atribut, jehož název je nesprávně napsán nebo nepatří do množiny definovaných atributů, se vůbec nezobrazí. Cokoliv nepatří do šablony daného infoboxu, je ignorováno. V názvech atributů jsou rozlišována velká a malá písmena. Návrhy na nové atributy lze podávat na diskusní stránce k příslušné šabloně.

5 Návrh nástroje pro extrahování údajů z Wikipedie

Mým hlavním úkolem je najít způsob, jak zpracovat velké množství dat tak, aby s nimi bylo možné dále pracovat a efektivně v nich vyhledávat. Výsledkem první části mé práce by měla být aplikace, která podle uživatelem definovaných parametrů extrahuje potřebné informace z dumpu Wikipedie a uloží je do jednoduché relační databáze. Druhou částí pak bude získání patřičných informací o množině historických osobností a vztahů mezi nimi a převedení této množiny do grafové databáze.

5.1 Možnosti vyhledávání

K nástroji pro vyhledávání bude uživatel přistupovat prostřednictvím grafického uživatelského rozhraní. Hlavním účelem nástroje bude extrahování infoboxů z dumpu Wikipedie do samostatného souboru, jehož obsah bude moci být následně přetvořen do databáze. Nebude to však jediný účel, bude poskytovat možnost jednoduchého, ale nepříliš efektivního vyhledávání nebo provádění statistik týkajících se např. toho, kolik infoboxů se v souboru vyskytuje a jaké je zastoupení jednotlivých typů.

GUI bude konfigurovatelné, nabídne uživateli možnost volby typu nebo specifikaci názvu hledaných infoboxů. Typ infoboxu bude vybrán z omezené množiny předem definované v konfiguračním souboru. Uživatel nebude muset prohledávat soubor vždy od začátku, ale bude si moci zvolit počáteční pozici.

Prohledávání necelých 50 GiB dat zabere několik hodin, uživatel bude tudíž informován o jeho průběhu.

5.2 Čtení dumpu Wikipedie

Dump Wikipedie je sice XML soubor, ale nejedná se o příliš členité XML, proto pro čtení souboru nebudu používat žádnou z metod parsování XML souborů (SAX, DOM). Soubor bude sekvenčně čten za účelem vyhledávání

posloupností znaků odpovídajících šablonám infoboxů. Během čtení budou zaznamenávány titulky stránek, na kterých se právě nacházíme.

5.3 Vytvoření databáze

Kvůli objemu zpracovávaných dat bude vytvoření databáze sestávat ze dvou oddělených kroků:

1. extrahování infoboxů a jim příslušných názvů stránek z dumpu Wikipedie
2. zpracování souboru, který vznikl jako výsledek předchozího kroku, a přesunutí dat v něm obsažených do databáze

Databázi bude tvořit jedna velká tabulka, kde řádky budou jednotlivé infoboxy a sloupce budou id, title, type, body a category.

- `id` - identifikátor
- `title` - titulek stránky, na které se infobox nachází (považován za název infoboxu)
- `type` - typ infoboxu (person, officeholder apod.)
- `body` - tělo infoboxu v nezpracované textové podobě
- `category` - kategorie infoboxu (osoba, místo, událost apod.)

Ačkoliv je vyhledávání záznamů v relační databázi podstatně rychlejší než v pouhém textovém souboru, pro automatické vyhledávání většího množství údajů však stále může být relativně pomalé, provádění jednoho dotazu SELECT se může pohybovat v řádu vteřin až minut. Proto je třeba nad sloupcem, podle kterého budeme záznamy vyhledávat (v tomto případě názvem infoboxu), vytvořit tzv. index.

Index je pomocná datová struktura umožňující rychlé vyhledávání ve větších objemech dat, což má za následek zvýšení nároků databázového serveru

na operační paměť a diskový prostor. Indexy se používají k rychlému vyhledání dat bez nutnosti procházení celé tabulky při každém dotazu. Index může být použit ve spojení s jedním nebo více sloupci. Většinou má podobu B-stromu[1].

5.4 Další práce s daty

Získaná data je třeba dále zpracovávat a získávat tak smysluplné hodnoty např. z řádky s datem narození nebo seznamem potomků. Data z dumpu Wikipedie jsou sice částečně strukturovaná, obsahují však mnoho nepravidelností a chyb, nelze tedy pro získávání informací z nich použít žádných existujících technik či nástrojů. Před vkládáním uzlů do grafové databáze budou vždy upraveny hodnoty atributů představujících významná data a osoby v životě daného člověka.

Nakonec přijde samotné vkládání do grafové databáze. Graf je datová struktura sestávající z vrcholů, které jsou propojeny hranami. Znázorňuje se obvykle jako množina bodů spojených čárami. Formálně je graf uspořádanou dvojicí množiny vrcholů V a množiny hran E [30]. V našem případě budou vrcholy představovat jednotlivé infoboxy a hrany vztahy mezi nimi.

[1] vyvážený strom, ve kterém operace přidání, vybírání a vyhledávání prvků probíhají v logaritmickém čase

6 Implementace nástroje

Aplikace je naprogramována v jazyce Java. Skládá se ze tří částí: tříd pro zpracování dumpu Wikipedie, pro vytvoření relační databáze a pro převedení dat do grafové databáze.

Hlavními třídami aplikace jsou dvě třídy implementující rozhraní `IReader` - `DumpReader` a `InfoboxFileReader`.

6.1 Třída `DumpReader`

Jak už název napovídá, v prvním případě se jedná o třídu, jejímž účelem je zpracování XML souboru s dumpem Wikipedie v té podobě, v jaké ho stáhneme z webu. Soubor je prohledáván čtením po řádcích pomocí třídy `java.io.RandomAccessFile`, která mimo jiné umožňuje přímý přístup k souboru, takže soubor nemusíme prohledávat od začátku, ale můžeme zvolit počáteční pozici prohledávání (v bytech). V případě nalezení shody a úspěšného vymezení hranic infoboxu je z `java.io.RandomAccessFile` vytvořen `java.io.BufferedReader`, pomocí něhož je znovu přečten celý infobox kvůli správnému kódování. `RandomAccessFile` považuje všechny znaky za 1-bytové, vstupní soubor je však v kódování UTF-8, což je formát kódování, který používá proměnnou délku znaku (1 až 4 byty) a třída `BufferedReader` umožňuje kódování explicitně nastavit.

Prohledávání dumpu může probíhat v 6 různých režimech, které jsou vyjmenovány v enumu `EMode`:

- `KEYWORDS` - vyhledávání specifikované názvem hledaného infoboxu/infoboxů
- `TYPES` - vyhledávání specifikované typem hledaného infoboxu/infoboxů
- `KEYWORDS_AND_TYPES` - vyhledávání specifikované názvem i typem hledaného infoboxu/infoboxů
- `INFOBOXES_HEADS` - vyhledání hlaviček všech infoboxů

- `INFOBOXES_ATTRIBUTES` - vyhledávání infoboxů daných typů (ukládání do samostatných souborů pro další zpracování - např. zjišťování frekvence vyplňování daných atributů apod.)
- `INFOBOXES_ALL` - extrakce všech infoboxů z dumpu Wikipedie bez ohledu na jejich typ nebo název

Posledním režimem je `DATABASES` (viz později), ten však není určen pro zpracovávání dumpu Wikipedie.

Při prohledávání v režimech `KEYWORDS`, `TYPES`, `KEYWORDS_AND_TYPES` a `INFOBOXES_ALL` jsou získané výsledky zapisovány do textového souboru. Jednotlivé záznamy jsou řazeny za sebou ve formátu:

```
<title> titulek stránky, ze které infobox pochází</title>
hlavička infoboxu
tělo infoboxu
obsahující jeho atributy
```

```
<title>Autism</title>
{{Infobox disease
| Name = Autism
| Image = Autism-stacking-cans 2nd edit.jpg
| Alt = Young red-haired boy facing away from camera, stacking a
    seventh can atop a column of six food cans on the kitchen
    floor. An open pantry contains many more cans.
| Caption = Repetitively stacking or lining up objects is a
    behavior sometimes associated with individuals with autism.
| DiseasesDB = 1142
| ICD10 = {{ICD10|F|84|0|f|80}}
| ICD9 = 299.00
| ICDO =
| OMIM = 209850
| MedlinePlus = 001526
| eMedicineSubj = med
| eMedicineTopic = 3202
| eMedicine_mult = {{eMedicine2|ped|180}}
| MeshID = D001321
| GeneReviewsNBK = NBK1442
| GeneReviewsName = Autism overview
    <title>Alabama</title>
{{Infobox U.S. state
|Name = Alabama
|Fullname = State of Alabama
|Flag = Flag of Alabama.svg
...
```

Zdrojový kód 6.1: Ukázka souboru s extrahovanými infoboxy

Módy `INFOBOXES_ATTRIBUTES` a `INFOBOXES_HEADS` vytvoří soubory, které mohou být dále zpracovány pro statistické účely (např. četnost výskytů infoboxů daného typu nebo vyplnění konkrétních atributů). Režim `INFOBOXES_ATTRIBUTES` vytvoří pro každý typ infoboxu zvláštní soubor a ukládá do něj jen těla infoboxů bez názvu a hlavičky.

6.2 Balík `database`

Balík `database` obsahuje třídy potřebné pro vytvoření relační databáze. Používaným typem databáze je MySQL, což je zajištěno konfigurací ve třídě `shared.Controller`. Vstupním souborem je soubor s extrahovanými infoboxy popsaný v sekci 6.1. Ten je zpracováván třídou `InfoboxFileReader`. Soubor je čten pomocí `java.io.BufferedReader`. Přístup do databáze je zprostředkován třídou implementující rozhraní `IDatabaseManager`, konkrétně `MySQLManager`.

Výstupem je jednoduchá relační databáze s názvem, který zvolil uživatel. Obsahuje jednu tabulku, jejíž název je určen hodnotou konstanty `TABLE` ve třídě `IDatabaseManager` - defaultně `infoboxes`.

název sloupce	typ a parametry	význam
`id`	`INT NOT NULL AUTO_INCREMENT PRIMARY KEY`	identifikátor
`title`	`VARCHAR(256)`	titulek stránky, na které se infobox nachází
`type`	`VARCHAR(256)`	typ infoboxu
`body`	`VARCHAR(20000)`	tělo infoboxu
`category`	`INT`	kategorie infoboxu

Tabulka 6.1: Struktura tabulky `infoboxes`

Maximální délky řetězců ukládaných do sloupců tabulky `infoboxes` jsou definovány jako konstanty ve třídě `IDatabaseManager` - `TITLE_MAX_LENGTH`, `TYPE_MAX_LENGTH` a `BODY_MAX_LENGTH`.

Další konstantou, která stojí za zmínku, je počet infoboxů ukládaných do databáze v rámci jednoho dotazu - konstanta `FLUSH` třídy `InfoboxFileReader`, defaultně nastavena na hodnotu 500.

Infoboxy jsou v závistlosti na typu řazeny do 5 kategorií definovaných konstantami třídy `Infobox` (viz tabulka 6.2).

název	hodnota	význam
`CATEGORY_PERSON`	1	osoba
`CATEGORY_PLACE`	2	místo
`CATEGORY_EVENT`	3	událost
`CATEGORY_ITEM`	4	předmět
`CATEGORY_OTHER`	5	ostatní

Tabulka 6.2: Kategorie infoboxů

Pro zjednodušení jsem v rámci své práce rozlišovala jen kategorie `osoba` a `ostatní`. Do kategorie `osoba` spadají typy v seznamu, který je součástí tohoto dokumentu jako příloha B. Seznam byl vytvořen na základě statistik, kdy byl prohledán celý dump Wikipedie (revize z 8. 10. 2014) za účelem zjištění četností jednotlivých typů infoboxů. Výtah z těchto statistik je v příloze C. Do užšího seznamu pak byly vybrány typy s četností výskytů vyšší než 100, uvedené v souboru `CATEGORY_PERSON.txt`.

6.3 Balík `graph_database`

6.3.1 Vytváření grafové databáze

Balík `graph_database` obsahuje třídy potřebné pro vložení dat do grafové databáze `TimelineDatabase`, vytvářené v rámci jedné diplomové práce na ZČU. K této databázi je přistupováno prostřednictvím metod API.

O vytváření databáze se stará třída `GraphDatabaseCreating`. Databáze je vytvořena ve dvou fázích. V první fázi jsou do grafové databáze vloženy všechny uzly z připravené relační databáze, které spadají do kategorie `osoba`. Každá instance třídy `database.Infobox` je transformována do instance třídy `cz.zcu.fav.kiv.timeline.entity.Node`. V konstruktoru třídy `Node` jsou uzlu nastaveny následující atributy:

- `id` - id uzlu, nastaveno pomocí globálního čítače
- `name` - jméno - název infoboxu

- `description` - popis - typ infoboxu
- `stereotype` - `NodeStereotype.PERSON`
- `begin` - datum narození
- `end` - datum úmrtí
- `tags` - tagy; název a typ infoboxu
- `properties` - vlastnosti, dvojice klíč a hodnota získané z těla infoboxu

Ve druhé fázi jsou doplněny hrany mezi vrcholy. Vrcholy jsou zpracovávány popořadě. Z `properties` jsou vybrány atributy představující příbuzné a další významné osoby, které jsou/byly součástí života zpracovávané osoby. Názvy těchto atributů jsou specifikovány ve třídě `graph_database.Attributes`. Jedná se o atributy z následujícího seznamu, který je možno dále rozšířit:

spouse	successor
partner	preceded
parents	succeeded
children	leader
issue	monarch
offspring	vicepresident
mother	prime_minister
father	deputy
relatives	lieutenant
predecessor	alongside

Když jsou nalezena jména všech příbuzných, je pak pro každé z těchto jmen v grafové databázi vyhledán odpovídající uzel, a pokud existuje, je vytvořena hrana (instance třídy `cz.zcu.fav.kiv.timeline.entity.Bond`) a vložena do databáze. Vytváření databáze tak probíhá dvouprůchodově.

6.3.2 Získávání dat a jmen z textu

Formát dat v dumpu Wikipedie není jednotný, a proto bylo třeba implementovat algoritmus pro získávání požadovaných informací tak, aby bylo dosaženo optimálního poměru mezi přesností a úplností.

Na oficiálních stránkách Wikipedie je uvedeno, že pro každý typ infoboxu je definovaná konečná množina atributů. Všechny však nemusí být uvedeny vždy nebo nemusí být vyplněny. Ve třídě `graph_database.Attributes` jsou uvedeny vybrané názvy několika významných a často se opakujících atributů.

V následující tabulce jsou ukázky reprezentace některých dat z dumpu.

`birth_date = {{birth date|mf=yes|1905|2|2}`
`birth_date = {{birth date and age|1947|04|01|df=y}}`
`birth_date = {{Birth date|df=yes|1885|4|3}}`
`birth_date = March 3, 1847`
`birth_date = 2 July 1884`
`birth_date = c. 446 BC`
`birth_date = 18 June c.|980 [[Common Era|CE]]`
`birth_date = 304 BCE, Close to 7th Aug`
`death_date = 232 BCE (aged 72)`
`death_date = {{death date and age|161|3|7|86|9|19|df=y}}`
`death_date = 19 August AD 14 (aged 75)`
`death_date = {{death date and age|df=yes|1836|6|10|1775|1|20}}`
`death_date=April 4, 397`
`birth_date = {{BCE|384}} {{longitem|padding-top:0;line-height:1.4em |[[Stagira (ancient city)|Stagira]], Chalcidice ([[Chalkidiki]]), northern [[Greece]]}}}`

Tabulka 6.3: Ukázky reprezentace dat v dumpu Wikipedie

V řetězcové reprezentaci je hledáno správné datum (rok, měsíc a den), což zajišťuje třída `graph_database.DataMining`. Z těchto tří hodnot je následně vytvořena instance třídy `org.joda.time.DateTime`. Pokud není specifikován den nebo měsíc, je jejich číslo nahrazeno hodnotou 1.

V řetězcích se mohou také vyskytovat zkratky specifikující letopočet.

BC - z angl. Before Christ (před Kristem)
BCE - z angl. Before Common Era (před naším letopočtem)
AD - z lat. Anno Domini (léta Páně)
CE - z angl. Common Era (našeho letopočtu)

Aplikace kontroluje výskyt prvních dvou zkratek a případně nastaví instanci třídy `DateTime` příslušnou éru.

Data i jména jsou z řetězců získávána za využití regulárních výrazů a hledání výskytů různých znaků. U jmen algoritmus předpokládá, že alespoň jedno jméno začíná velkým písmenem. Při získávání data jsou vyhledávány číslice nebo názvy měsíců v angličtině.

V tabulce 6.4 je ukázka reprezentace seznamů jmen v dumpu.

`spouse = [[James Innes-Ker, 7th Duke of Roxburghe]]`
`children = Nathalie Felber, Jacqueline Felber`
`children = 4 sons`
`children = Lady Margaret Ewing [[Henry Innes-Ker, 8th Duke of Roxburghe]] Lady Victoria Villiers  Lady Isabel Wilson Lord Alastair Innes-Ker Lady Evelyn Collins Lord Robert Innes-Ker`
`relatives = [[Myat Paya Lat|Myat Phayalat]]`
`children = {{collapsible list|title=7|[[William Montagu Douglas Scott, 6th Duke of Buccleuch]]|[[Henry Douglas-Scott-Montagu, 1st Baron Montagu of Beaulieu]]|Lord Walter Montagu Douglas Scott|[[Lord Charles Montagu Douglas Scott]]|Victoria Kerr, Marchioness of Lothian|Lady Margaret Cameron|Lady Mary Trefusis}}`
`parents = ubl | [[Carlo Bugatti]] | Teresa Lorioli`
`predecessor =[[Sir Herbert Williams, 1st Baronet|Herbert Williams]]`
`successor = Constituency abolished`
`predecessor2 = [[Bill Woodroffe]]}}`

Tabulka 6.4: Ukázky reprezentace seznamů jmen v dumpu Wikipedie

Ukázka získání jmen ze seznamu

Vstupní řetězec:
`Spencer-Churchill, 7th Duke of Marlborough]]<br>[[Frances Anne Spencer-Churchill, Duchess of Marlborough\|Lady Frances Vane]]`

- řetězec je rozdělen s využitím separátoru daného regulárním výrazem `][^\[]*\[\[|&lt`
- výsledkem jsou dvě části:
  `Spencer-Churchill, 7th Duke of Marlborough`

`Frances Anne Spencer-Churchill, Duchess of Marlborough|Lady Frances Vane]]`

- každá část je následně rozdělena podle znaku '`|`', protože ten od sebe odděluje různé varianty jednoho jména, a dále je zpracovávána jen první část, ve které se hledá slovo začínající velkým písmenem a zpracovávají se pak znaky za ním následující
- výsledkem jsou dvě nalezená jména:
  `John Spencer-Churchill, 7th Duke of Marlborough`
  `Frances Anne Spencer-Churchill, Duchess of Marlborough`

Poznámka: Pokud se v řetězci vyskytují jména uzavřená v dvojitých hranatých závorkách (např. `[[Maximilian Agassiz]]`), jedná se o odkaz na existující stránku s tímto názvem.

7 Testy a možnosti rozšíření

Tato kapitola pojednává o testování aplikace, jejích nedostatcích a možnostech dalšího rozšíření.

7.1 Několik statistik

Zpracování dumpu

Extrahování všech infoboxů z dumpu Wikipedie je operace, která zabere několik hodin. V rámci testování na různých strojích se doba běhu pohybovala v rozmezí od 4 do 56 hodin (viz tab. 7.1).

Následující tabulka obsahuje srovnání parametrů strojů, na kterých byla testována doba extrakce.

Stroj	CPU	RAM	HDD	OS	Doba extrakce
A	IC i3-2120	8 GB	500GB HDD	Win 7 Pro 64b	56h 10m
B	IC i5 4210H	8 GB	500GB SSHD	Win 8.1 Pro 64b	21h 37m
C	IC i5 4210H	8 GB	240GB SSD	Debian 8.0 64b	4h 23m

Tabulka 7.1: Srovnání testovacích strojů

Počet infoboxů v dumpu se pohybuje v řádu jednotek milionů, z revize z 6. 11. 2014 jich bylo úspěšně extrahováno 2508151, dalších 386 bylo označeno jako vadných, protože nebyly nalezeny hranice infoboxu. Z těchto 2508151 infoboxů jich přibližně 28% (698620) spadá do kategorie osoba.

Velikost dumpu je 47 GiB, soubor s extrahovanými infoboxy zabírá 2,5 GiB.

Relační databáze

Vytvoření relační databáze ze souboru se všemi extrahovanými infoboxy zabere dobu v řádu jednotek až desítek minut. Při vytváření přesahuje 76 infoboxů stanovenou maximální délku (20000 znaků) a jsou na tuto hodnotu ořízнuty. Soubor s databází má po vytvoření indexu nad sloupcem `title` celkovou velikost 3,3 GiB.

Dotazy nad sloupcem `title` bez vytvořeného indexu se pohybují v řádu desítek vteřin až minut, s vytvořeným indexem v řádu milisekund.

Grafová databáze

Vytvoření grafové databáze ze všech infoboxů v dumpu Wikipedie zabere několik desítek hodin. Samotné vložení všech uzlů do databáze trvalo na testovacím stroji (stroj A v tabulce 7.1) 25 hodin, vložení hran pak bude trvat několikanásobně delší dobu. Velikost databáze obsahující všech 2,5 milionů uzlů (bez hran) je přibližně 3 GiB.

7.2 Testy

Aplikace byla testována na přesnost a úplnost, co se týká způsobu, jakým získává informace z řetězců představující hodnoty atributů z infoboxů extrahovaných z dumpu Wikipedie. Jejím úkolem v této oblasti je získávat korektní hodnoty dat (dat narození, úmrtí atd.) a jmen (popř. titulů či přízvisek) osob ze seznamů.

V obou případech byly provedeny 3 testy o 50 otázkách. Výsledky byly vyhodnocovány ručně. Vstupní hodnoty byly náhodně vybrány z množiny všech 2508151 infoboxů dané revize.

7.2.1 Testy přesnosti a úplnosti dat

V případě dat bylo za správnou hodnotu považováno odpovídající datum ve formátu *dd.MM.yyyy* a letopočet (v případě éry před naším letopočtem je v instancích třídy `org.joda.time.DateTime` uváděn rok jako záporná hodnota). Pokud není v textu specifikován den nebo měsíc, je za správnou hodnotu považováno datum s příslušnými hodnotami nahrazenými číslem 1. Pokud není datum v textu vůbec uvedeno, je za správnou hodnotu považováno `null`.

Úspěšnost získávání korektních dat z řetězců je relativně vysoká. Poslední hodnotu v tabulce 7.2 sice algoritmus vyhodnotil nesprávně, jedná se však o neobvyklé datum, ve kterém chybí rok, což algoritmus nepředpokládá.

Řetězec	<<Birth date and age|1972|10|27>>
Výstup	**27.10.1972 (správně)**
Řetězec	<!– <<Death date and age|YYYY|MM|DD|1972|10|27>> –>
Výstup	**27.10.1972 (správně)**
Řetězec	October 4, 1918
Výstup	**4.10.1918 (správně)**
Řetězec	January 9, 1998 (aged 79)
Výstup	**9.1.1998 (správně)**
Řetězec	1144
Výstup	**1.1.1144 (správně)**
Řetězec	<<death year and age|1200|1144>>
Výstup	**1.1.1200 (správně)**
Řetězec	1978<!– <<Birth date and age|YYYY|MM|DD>> –>
Výstup	**1.1.1978 (správně)**
Řetězec	<!– <<Death date and age|YYYY|MM|DD|YYYY|MM|DD>> (death date then birth date) –>
Výstup	**null (správně)**
Řetězec	<<Birth-date and age|1933>>
Výstup	**1.1.1933 (správně)**
Řetězec	September 1, 2008
Výstup	**1.9.2008 (správně)**
Řetězec	c. 1892
Výstup	**1.1.1892 (správně)**
Řetězec	14 July
Výstup	**1.1.14 (špatně)**

Tabulka 7.2: Ukázka vyhodnocení správnosti dat

Číslo testu	Počet korektních hodnot	Úspěšnost
1	50	100%
2	50	100%
3	49	98%

Tabulka 7.3: Výsledky testu na správné získávání dat

7.2.2 Testy přesnosti a úplnosti jmen

Za správnou odpověď je považována ta, ve které se nacházejí všechna jména ze seznamu včetně přízvisek a titulů, případně prázdný řetězec, pokud se v původním řetězci žádné jméno nevyskytuje.

Získaná jména jsou oddělena dvojitým středníkem.

Řetězec	Miriam Szenberg
Výstup	**Miriam Szenberg (správně)**
Řetězec	Elizabeth Ellen Webster m. 1895
Výstup	**Elizabeth Ellen Webster;; (správně)**
Řetězec	<<marriage|Daniel Chao|2011>>
Výstup	**Daniel Chao (správně)**
Řetězec	Pamela Maturana Rivera María Gabriela Maturana Rivera
Výstup	**Pamela Maturana Rivera;; María Gabriela Maturana Rivera (správně)**
Řetězec	"'Son:"' Paul Leone Peters "'Daughter:"' Gail Peters Beitz>>
Výstup	**Son;; Daughter (špatně)**
Řetězec	Veza Taubner-Calderon (1934-?) Hera Buschor (m. 1971)>>
Výstup	**Veza Taubner-Calderon ;; Hera Buschor (správně)**
Řetězec	1978<!– [[Princess Altinaï of Montenegro|Princess Altinaï]] [[Boris, Hereditary Prince of Montenegro|Prince Boris]]
Výstup	**Princess Altinaï of Montenegro;; Boris, Hereditary Prince of Montenegro (správně)**
Řetězec	[Sherill Lynn Rettino]], [[Mitchell Wayne Katzman]], [[Frank Katzman]]
Výstup	**Sherill Lynn Rettino;; Mitchell Wayne Katzman;; Frank Katzman (správně)**

Tabulka 7.4: Ukázka vyhodnocení správnosti jmen

Číslo testu	Počet korektních hodnot	Úspěšnost
1	45	90%
2	46	92%
3	45	90%

Tabulka 7.5: Výsledky testu na správné získávání jmen ze seznamů

7.2.3 Testy vkládání hran

Při vytváření grafové databáze bylo testováno, kolik hran bylo skutečně vytvořeno na základě seznamu příbuzných osob každého uzlu.

Byly provedeny 3 testy s náhodným vzorkem dat, testy byly vyhodnoceny automaticky. Výsledky jsou zobrazeny v tabulce 7.6.

Vytvoření hrany mezi dvěma uzly ovlivňují dva faktory:

- přesnost a úplnost algoritmu na získávání jmen ze seznamů
- přítomnost infoboxu představovaného daným jménem ve vzorku dat, v rámci celé Wikipedie je třeba, aby skutečně existovala stránka s daným jménem a navíc obsahovala infobox, jehož typ spadá do kategorie osoba

Počet infoboxů	Počet vytvořených uzlů (osob)	Počet nalezených jmen příbuzných	Počet vytvořených uzlů	Úspěšnost
1000	312	895	49	5.5%
10000	3267	7898	1260	16.0%
50000	6140	14955	2913	19.5%

Tabulka 7.6: Výsledky testu vkládání hran

Tyto testy nevypovídají o skutečné procentuální úspěšnosti nástroje, protože nebyl testován s úplnými vstupními daty z Wikipedie, je však vidět, že s větším testovacím vzorkem dat roste úspěšnost vytváření hran mezi uzly.

7.2.4 Nedostatky a návrhy na vylepšení

Nástroj na automatické získávání historických údajů s sebou nese mnoho nedostatků. Jedním z nich je reprezentace dat, kdy se i přes jejich strukturovanost v textu vyskytuje mnoho nepravidelností a neúplností. Zajímavým faktem je např. to, že se v dumpu Wikipedie nevyskytuje infobox popisující jednu z nejvýznamnějších historických událostí, a to 1. světovou válku. Zdrojový kód stránky obsahuje zvláštní typ infoboxu bez výčtu atributů:

```
{{World War I infobox}}
```

V aplikaci je velký prostor pro vylepšení. Bylo by dobré důkladněji prozkoumat typy infoboxů tak, aby bylo možné každý z nich zařadit do jedné z připravených kategorií:

- osoba
- místo
- událost
- předmět
- ostatní

Zároveň by pak bylo vhodné zpracovávat větší množinu atributů určitých typů infoboxů a nezaměřovat se jen na data a jména.

Co se týče dat a jmen, je zde možnost vylepšit stávající algoritmus na jejich získávání nebo navrhnout jiný, který by dosahoval větší úspěšnosti.

Aplikace je navržena jen pro práci s dumpem anglické verze Wikipedie, typy infoboxů a atributy jsou pro každý jazyk specifické.

8 Závěr

Hlavním úkolem této bakalářské práce bylo prozkoumat různé webové zdroje, vybrat z nich jeden, z něhož bude možné čerpat při vytváření databáze, a implementovat nástroj, který umožní automatizované získání dat z tohoto zdroje a jejich transformování do požadované podoby.

Jako takový zdroj dat byla vybrána Wikipedie kvůli velmi vysoké přesnosti dat vzhledem k jejich množství a také díky strukturovaným tabulkám zvaným infoboxy. Důležitým faktorem byla také možnost stažení obsahu stránek, což umožňuje další práci bez přístupu k internetu.

Výsledkem práce je nástroj, který v několika fázích dovede zpracovat vstupní soubor s obsahem Wikipedie a naplní požadovanými daty nejdříve relační a poté grafovou databázi vytvořenou v rámci projektu zaměřeného na vytváření a vizualizaci časové osy, který byl předmětem několika diplomových prací na ZČU.

Nástroj je poměrně snadno konfigurovatelný a kromě nastavení parametrů před každou fází zpracování nevyžaduje zásah uživatele.

Seznam použitých zkratek

- CC-BY-SA - Creative Commons Attribution-ShareAlike
- DOM - Document Object Model
- FEN - Family Education Network
- GFDL - GNU Free Documentation License
- HTML - HyperText Markup Language
- JSON - JavaScript Object Notation
- MQL - Metaweb Query Language
- NLP - Natural Language Processing
- OCR - Optical character recognition
- OER - Open educational resources
- OS - operační systém
- OWL - Web Ontology Language
- RDF - Resource Description Framework
- SAX - Simple API for XML
- SPARQL - SPARQL Protocol and RDF Query Language
- URI - Unified Resource Identifier
- UTF-8 - UCS Transformation Format - způsob kódování řetězců znaků Unicode/UCS do sekvencí bajtů

- WMF - Wikimedia Foundation
- XML - EXtensible Markup Language
- YAGO - Yet Another Great Ontology
- ZIM - Zeno IMproved

Seznam obrázků

Seznam tabulek

Literatura

[1] SHETTAR, Rajashree, SHOBHA G, Dr., *Survey on Mining in Semi-Structured Data* [online], 2007, cit[2015-04-12], http://paper.ijcsns.org/07_book/200708/20070832.pdf

[2] *Gartner, Inc.* [online], cit[2015-04-12], http://www.gartner.com/technology/

[3] *Merrill Lynch, Bank of America Corporation* [online], cit[2015-04-12], https://www.ml.com/

[4] WITTEN, Ian H., FRANK, Eibe, HALL, Mark A. *Data mining, Practical Machine Learning Tools and Techniques*, 3. vyd., USA 2011, cit[2015-04-12]

[5] FELDMAN, Ronen, SANGER, James, *The text mining handbook, Advanced Approaches in Analyzing Unstructured Data*, Cambridge University Press, 2007, cit[2015-04-12]

[6] KAO, Anne, POTEET, Stephen R., *Natural Language Processing and Text Mining*, Springer-Verlag London Limited 2007, cit[2015-04-12]

[7] *Wordnet, A lexical database for English* [online], cit[2015-04-12], https://wordnet.princeton.edu/

[8] *Wikipedia, the free encyclopedia* [online], cit[2015-04-13], http://en.wikipedia.org/wiki/Wikipedia

[9] *DBpedia* [online], cit[2015-04-13], http://http://dbpedia.org/About

[10] *Max Planck Institut Informatik* [online], cit[2015-04-14], http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/faq/

[11] *Cambridge Semantics* [online], cit[2015-04-13], http://www.cambridgesemantics.com/semantic-university/rdf-101

[12] *Google Developers* [online], cit[2015-04-13], https://developers.google.com/freebase/guide/basic_concepts

[13] *History world* [online], cit[2015-04-13], http://www.historyworld.net/

[14] *Ancient History Encyclopedia* [online], cit[2015-04-13], http://www.ancient.eu/

[15] *HyperHistory Online* [online], cit[2015-04-13], http://www.hyperhistory.com/online_n2/History_n2/a.html

[16] *Infoplease* [online], cit[2015-04-13], http://www.infoplease.com/

[17] *Encyclopedia.com* [online], cit[2015-04-13], http://www.encyclopedia.com/

[18] *Encyclopedia Britannica* [online], cit[2015-04-13], http://www.britannica.com/

[19] *Who's Who* [online], cit[2015-04-13], http://www.ukwhoswho.com/

[20] *MusicBrainz* [online], cit[2015-04-13], https://musicbrainz.org/

[21] *Google Code: BzReader* [online], cit[2015-04-18], https://code.google.com/p/bzreader/

[22] *Kiwix* [online], cit[2015-04-18], http://www.kiwix.org/wiki/Main_Page

[23] *openZIM* [online], cit[2015-04-18], http://www.openzim.org/wiki/ZIM_file_format

[24] http://www.techpraveen.com/2011/01/kiwix-read-wikipedia-contents-offline.html

[25] *WikiTaxi* [online], cit[2015-04-18], http://www.wikitaxi.org/delphi/doku.php/products/wikitaxi/index

[26] *w3schools: XML* [online], cit[2015-04-19],
`http://www.w3schools.com/xml/`

[27] *Wikipedia: Wiki markup* [online], cit[2015-04-19],
`http://en.wikipedia.org/wiki/Help:Wiki_markup#Redirects`

[28] *Wikipedia: Help:Infobox* [online], cit[2015-04-19],
`http://en.wikipedia.org/wiki/Help:Infobox`

[29] *Wikipedia: List of infoboxes* [online], cit[2015-04-19],
`http://en.wikipedia.org/wiki/Wikipedia:List_of_infoboxes`

[30] *Wikipedie: Teorie grafů* [online], cit[2015-06-12],
`http://cs.wikipedia.org/wiki/Teorie_grafů`

A Uživatelská dokumentace

A.1 Příprava dat

Dump Wikipedie lze stáhnout z následujícího odkazu:

`http://meta.wikimedia.org/wiki/Data_dump_torrents`

Aplikace pracuje s atributy v anglickém jazyce, proto stáhneme dump anglické Wikipedie, např. soubor `enwiki-20150602-pages-articles-xml.bz2`.

Samotné stažení může trvat několik hodin (až okolo 10). Soubor bude zabírat okolo 10 GiB, po rozbalení okolo 50 GiB. Lze ho dekomprimovat např. pomocí programů WinRAR nebo bzip2.

Aplikace pak pracuje s dekomprimovaným souborem `enwiki-20150602-pages-articles.xml`.

A.2 Zpracování dat

Nástroj pro zpracovávání dat je aplikace s názvem `DataRetrieval`.

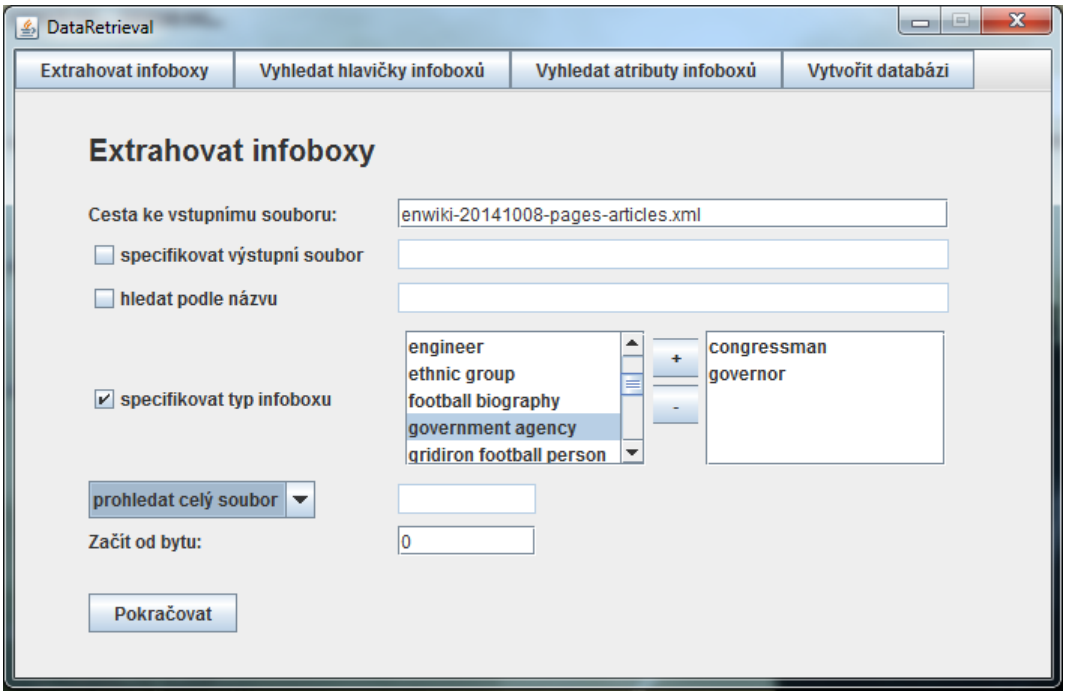
Aplikace se spouští s příkazové řádky příkazem

```
java -jar DataRetrieval
```

Zobrazí se okno (viz obr. A.1), které nabízí uživateli 4 možnosti.

- Extrahovat infoboxy
- Vyhledat hlavičky infoboxů
- Vyhledat atributy infoboxů
- Vytvořit databázi

Vstupním souborem u prvních tří možností je stažený XML soubor.



Obrázek A.1: Panel `Extrahování infoboxů`

A.2.1 Extrahování infoboxů z dumpu

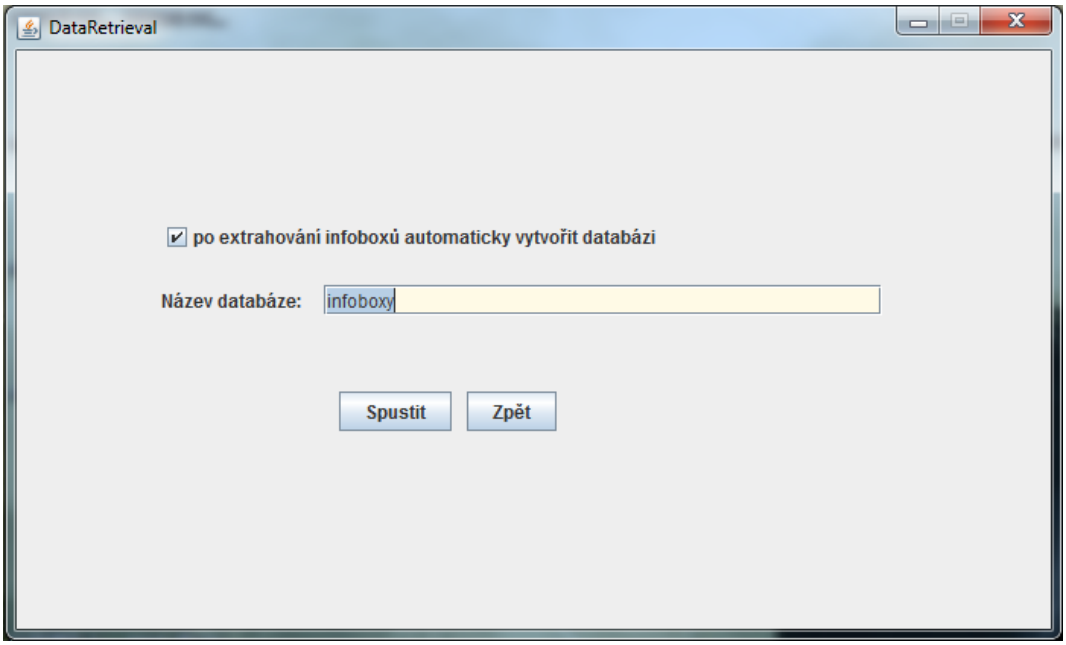
Uživatel může specifikovat název výstupního souboru, pokud to neudělá, bude vytvořen soubor s názvem ve tvaru `wiki_output_dd.MM.yy_HH.mm.txt`.

Uživatel může hledat infoboxy podle názvu či názvů oddělených symbolem ',' nebo podle typu, který může vybrat ze seznamu.

Lze také specifikovat množství hledaných výskytů nebo pozice v souboru (byte), od které má prohledávání začít.

Po nastavení parametrů vyhledávání se uživatel stiskem tlačítka `Pokračovat` přesune na další obrazovku (viz obr. A.2), která mu nabídne možnost automatického vytvoření MySQL databáze po extrahování všech infoboxů. K tomu je ovšem potřeba patřičně nastavený a běžící MySQL server. O tom více v sekci A.2.3.

Pokud extrahujeme všechny infoboxy nezávisle na názvu či typu, bude výsledný soubor zabírat necelé 3 GiB. Tato operace trvá několik hodin.



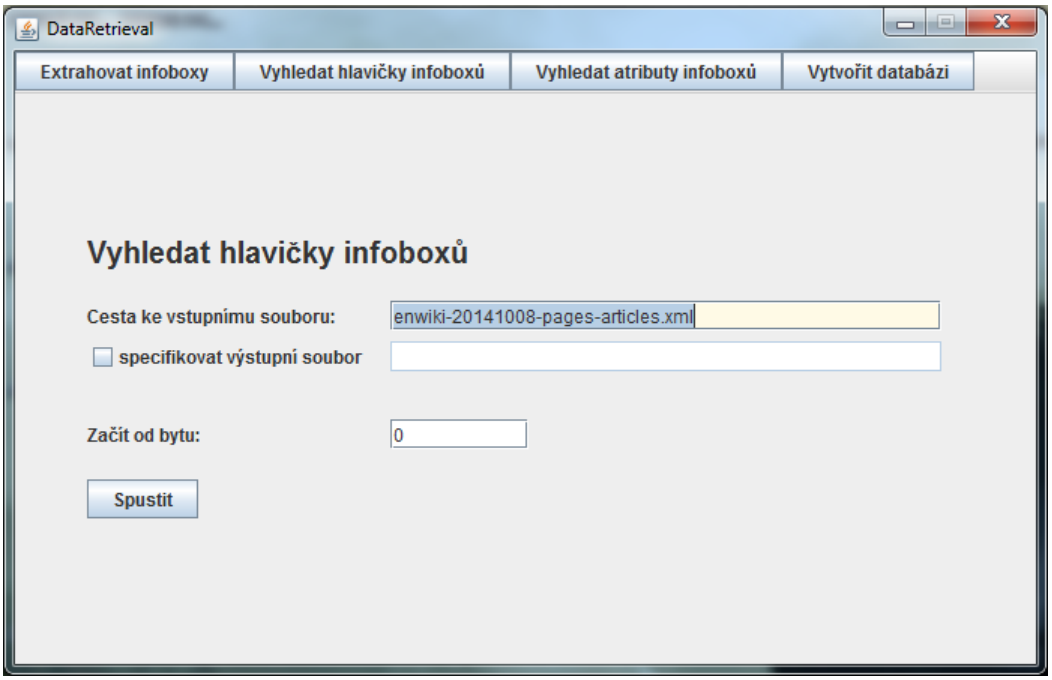
Obrázek A.2: Panel `Pokračovat vytvořením databáze`

A.2.2 Vyhledávání hlaviček a atributů

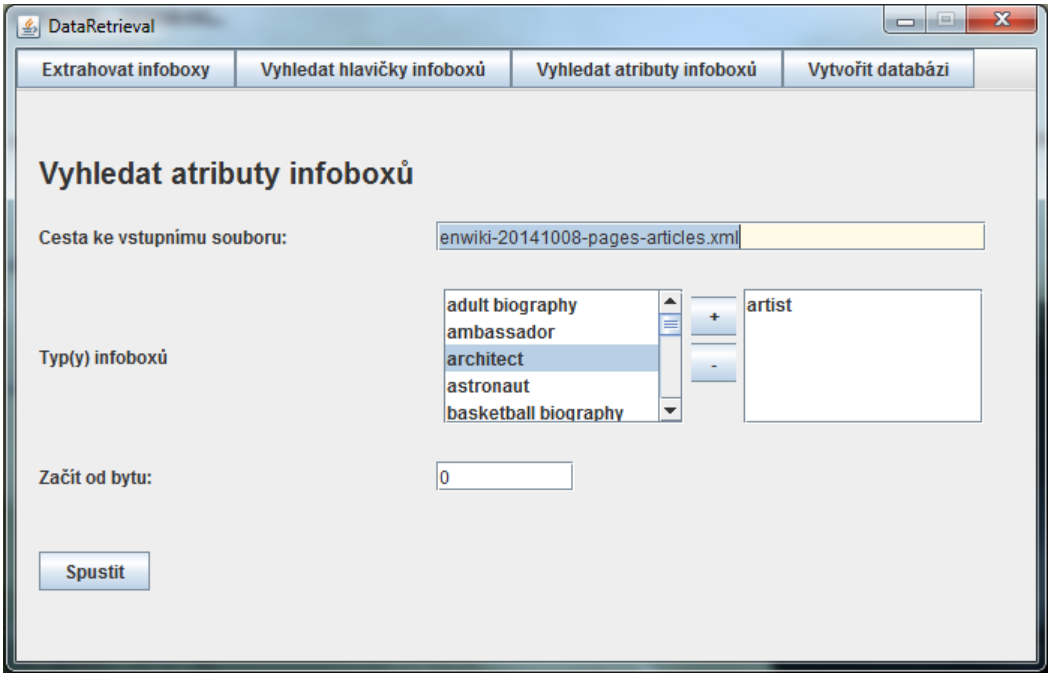
Následující dvě funkce jsou spíše doplňkové a lze je využít pro statistické účely.

Funkce `Vyhledat hlavičky infoboxů` extrahuje do souboru řádky s hlavičkami všech infoboxů. Tato funkce může sloužit například pro zjištění četností výskytů jednotlivých typů infoboxů.

V okně `Vyhledat atributy infoboxů` uživatel specifikuje zkoumané typy infoboxů a do samostatných souborů pak budou extrahovány atributy infoboxů těchto typů. Název souborů bude ve tvaru `wiki_TYPE_dd.MM.yy_HH.mm.txt`.



Obrázek A.3: Panel `Vyhledávání hlaviček infoboxů`



Obrázek A.4: Panel `Vyhledávání atributů infoboxů`

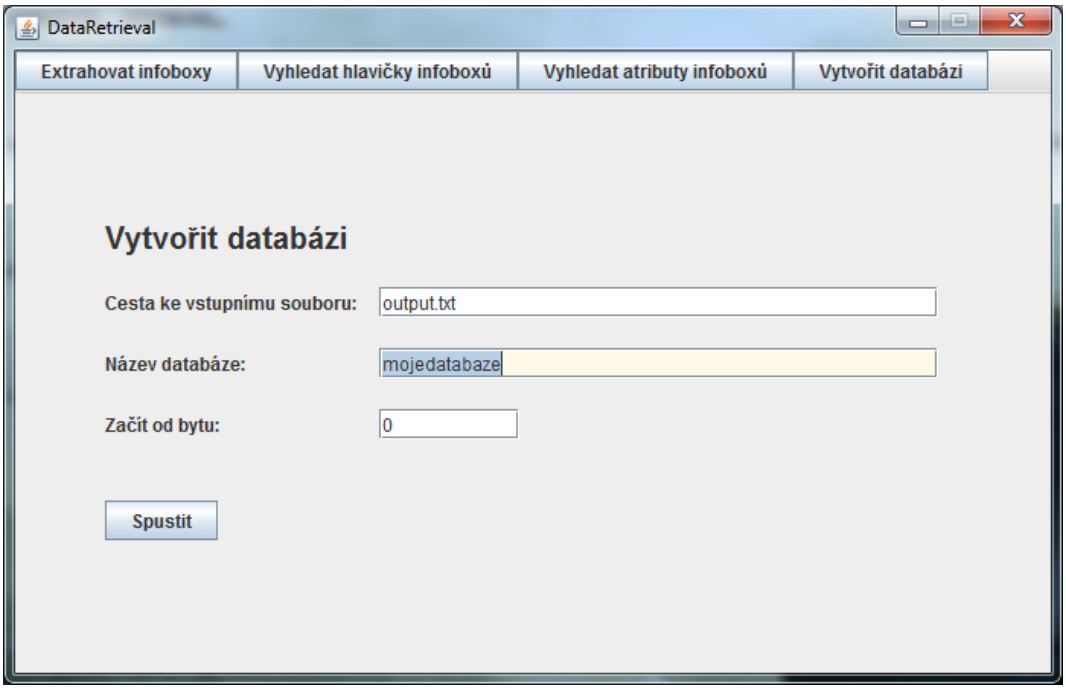
A.2.3 Příprava a vytváření relační databáze

Ještě před vytvářením relačním databáze je třeba nainstalovat a zapnout MySQL server. Aplikace se bude do databáze připojovat s následujícími parametry:

- jméno serveru: `localhost`
- uživatel: `root`
- heslo: bez hesla

Jedná se o defaultní konfiguraci, takže většinou není třeba nic nastavovat.

Vstupním souborem pro vytváření databáze je soubor s extrahovanými infoboxy vytvořený funkcí popsanou v A.2.1. Pokud už máte takový soubor k dispozici, databázi lze vytvořit v záložce `Vytvořit databázi` (viz obr A.5).



Obrázek A.5: Panel `Vytvoření databáze`

Název databáze musí být jedinečný, databáze s daným jménem nesmí existovat, jinak bude zobrazena chybová hláška. Název databáze nesmí obsahovat znaky z následujícího seznamu: `<>/\:"*`. V databázi bude vytvořena jediná tabulka s názvem `infoboxes` obsahující všechny záznamy.

Čas vytvoření databáze se obvykle pohybuje v řádu minut.

V jednom dotazu je do databáze vloženo 500 infoboxů, jedná se tak někdy o pakety větší než 1 MB. V případě problémů s velikostí přenášených paketů lze nastavit maximální hodnotu v konfiguračním souboru MySQL (`my.ini` nebo `my.cnf`):

```
max_allowed_packet = 10M
```

A.2.4 Vytváření grafové databáze

K vytvoření grafové databáze slouží doplňková aplikace `GraphDatabaseCreating.jar`.

Spustí se příkazem

`java -jar GraphDatabaseCreating.jar 1 REL_DB_NAME GRAPH_DB_PATH`
pro vytvoření kompletní grafové databáze

`java -jar GraphDatabaseCreating.jar 2 REL_DB_NAME GRAPH_DB_PATH`
pro vložení pouze uzlů do grafové databáze

`java -jar GraphDatabaseCreating.jar 3 GRAPH_DB_PATH`
pro vytvoření hran v existující grafové databázi.

`REL_DB_NAME` je název relační databáze, ze které aplikace čerpá data.

`GRAPH_DB_PATH` je cesta k vytvářené grafové databázi.

A.2.5 Logování

Zpracování velkého množství dat je poměrně složitá záležitost, a proto je třeba různé události (varování a výjimky) logovat do souborů.

- `ERRORLOG_WIKI_DUMP.log` - log pro zpracovávání dat z dumpu Wikipedie
- `ERRORLOG_RELATIONAL_DB.log` - log pro vytváření relační databáze
- `ERRORLOG_GRAPH_DB.log` - log pro vytváření grafové databáze

Na konci každého čtení je do logů zapsána sumarizace zvláštních událostí.

B Infoboxy s nejčastějším výskytem v revizi z 8. 10. 2014 (kategorie osoba)

Typ	Počet výskytů	Typ	Počet výskytů
person	139425	saint	2992
football biography	116540	athlete	2955
musical artist	74365	figure skater	2908
officeholder	33257	martial artist	2892
military person	26931	professional wrestler	2815
mlb player	19054	golfer	2770
writer	18452	state senator	2747
scientist	18310	monarch	2609
sportsperson	17793	comics creator	2588
ice hockey player	14279	nba biography	2278
nfl player	13740	pageant titleholder	2278
cricketer	13495	president	2174
artist	10514	nobility	1929
politician	9879	indian politician	1921
royalty	8549	judge	1844
cyclist	6666	racing driver	1820
basketball biography	6147	criminal	1716
gridiron football person	6011	baseball biography	1713
rugby league biography	5390	prime minister	1672
rugby biography	5171	model	1559
state representative	5171	philosopher	1557
afl biography	5130	architect	1552
christian leader	5048	adult biography	1524
tennis biography	4164	gymnast	1522
MP	4019	mayor	1512
congressman	3952	canadianMP	1487
swimmer	3855	handball biography	1440
gaa player	3321	rugby union biography	1433
boxer	3138	cfl player	1318
governor	3016	volleyball player	1291

Tabulka B.1: 60 nejčastějších typů infoboxů v kategorii osoba

C Infoboxy s nejčastějším výskytem v revizi z 8. 10. 2014

Typ	Počet výskytů	Typ	Počet výskytů
settlement	350321	station	13991
person	139425	airport	13969
album	126196	nfl player	13740
football biography	116540	cricketer	13495
film	89714	political party/seats	13054
musical artist	74365	german location	13016
single	49616	military conflict	12647
company	46705	planet	11400
nrhp	45196	aircraft begin	11050
french commune	36790	artist	10514
officeholder	33257	aircraft type	9942
book	31838	software	9898
ship career	31203	politician	9879
television	29447	river	9728
ship characteristics	28448	building	9629
ship image	27351	lake	9545
military person	26931	organization	9419
school	22033	australian place	8765
uk place	20237	royalty	8549
mlb player	19054	rockunit	8368
writer	18452	stadium	8297
radio station	18323	automobile	8137
scientist	18310	italian comune	8110
road	17826	language	8033
sportsperson	17793	ncaa team season	7870
university	16925	television episode	7758
football club	16859	korean name	7705
video game	16429	football club season	7644
military unit	15496	election	7591
mountain	15418	hurricane small	6904
ice hockey player	14279	cyclist	6666

Tabulka C.1: Prvních 62 typů infoboxů s nejčastějším výskytem