ZÁPADOČESKÁ UNIVERZITA V PLZNI
FAKULTA APLIKOVANÝCH VĚD

Katedra informatiky a výpočetní techniky

BAKALÁŘSKÁ PRÁCE

Konfigurovatelné získávání grafových dat z otevřených zdrojů

Plzeň 2021 Karel Šilhavý

ZÁPADOČESKÁ UNIVERZITA V PLZNI
Fakulta aplikovaných věd
Akademický rok: 2020/2021

ZADÁNÍ BAKALÁŘSKÉ PRÁCE
(projektu, uměleckého díla, uměleckého výkonu)

Jméno a příjmení:	**Karel ŠILHAVÝ**
Osobní číslo:	**A18B0325P**
Studijní program:	**B3902 Inženýrská informatika**
Studijní obor:	**Informatika**
Téma práce:	**Konfigurovatelné získávání grafových dat z otevřených zdrojů**
Zadávající katedra:	**Katedra informatiky a výpočetní techniky**

Zásady pro vypracování

1. Seznamte se s problematikou analýzy částečně strukturovaných dat.
2. Analyzujte možnosti získávání grafových dat z dostupných zdrojů.
3. Navrhněte konfigurovatelný nástroj pro získávání stromových dat.
4. Implementujte desktopovou aplikaci, která umožní konfigurovat získávání dat, provést jejich stažení a následně je procházet a upravovat podle potřeby.
5. Proveďte důkladné testování implementovaného nástroje, zejména s ohledem na přesnost a úplnost.

Rozsah bakalářské práce:	**doporuč. 30 s. původního textu**
Rozsah grafických prací:	**dle potřeby**
Forma zpracování bakalářské práce:	**tištěná**

Seznam doporučené literatury:

Dodá vedoucí bakalářské práce.

Vedoucí bakalářské práce:	**Ing. Richard Lipka, Ph.D.** Katedra informatiky a výpočetní techniky
Datum zadání bakalářské práce:	**5. října 2020**
Termín odevzdání bakalářské práce:	**6. května 2021**

L.S.

Doc. Dr. Ing. Vlasta Radová
děkanka

Doc. Ing. Přemysl Brada, MSc., Ph.D.
vedoucí katedry

V Plzni dne 26. října 2020

ABSTRAKT

V rámci této bakalářské práce byl vytvořen nástroj, který umožňuje konfigurovatelné získávání grafových dat z otevřených zdrojů informací. Konkrétně z částečně strukturovaných tabulek na Wikipedii zvaných infoboxy. Na základě interakce s uživatelem data propojuje požadovaným způsobem a výsledek uloží v dále použitelném formátu pro zobrazení v časové ose Timeline. Uživateli je tak umožněno téměř automatizovaným postupem sestavovat například rodokmeny historických osobností, vládnoucí linie a různé další mapy vztahů mezi osobami.

KLÍČOVÁ SLOVA

konfigurovatelná těžba dat, otevřené zdroje informací, Wikipedie, infobox, infoboxy

ABSTRACT

Within this bachelor thesis was created a tool that allows configurable obtaining graph data from open sources of the information. The tool is design to work specifically with the semi-structured tables on Wikipedia called infoboxes. It links the data in the desired way based on the interaction with user and saves the result in another usable format for the Timeline. Therefore user can construct for example family trees of historical figures, ruling dynasties and other various relationship maps by almost automated manner.

KEY WORDS

configurable data mining, open sources of information, Wikipedia, infobox, infoboxes

Prohlášení

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

V Plzni dne 24. června 2021

Karel Šilhavý

Poděkování

Rád bych poděkoval vedoucímu bakalářské práce panu Ing. Richardu Lipkovi, Ph.D. za trpělivost, vstřícný přístup, cenné rady a připomínky a za čas věnovaný konzultacím při vypracovávání této bakalářské práce.

Obsah

1 Úvod

S rostoucím využívání Internetu se zvyšuje množství dat, která jsou z něj dostupná. Každým rokem se zvyšuje míra digitalizace téměř ve všech oborech lidské činnosti. Oproti předešlým způsobům uchovávání, zpracování a sdílení informací je práce s jejich digitální podobou rychlejší, levnější a v drtivé většině případů jednodušší. Avšak velká část těchto dat je nestrukturovaná, což znesnadňuje automatické strojové zpracování a je stále nutné využívat lidský faktor pro zpracování dat.

Dříve člověk, který chtěl zjistit informace o svých předcích, musel navštívit matriku či dokonce více matrik v různých městech České republiky anebo i v cizině. Dalším příkladem jsou staré i nové knihy. Staré knihy jsou často převáděny do digitální podoby formou naskenovaného dokumentu, bývají psané nestandardizovaným formátem, případně ručně psaným písmem. Na takto naskenované dokumenty je použita technika OCR (převod psaného textu do digitální podoby). I přes dnešní pokročilou techniku není tento proces vždy úspěšný, a proto z toho vnikají neúplně strojově zpracovatelná data. Kdyby lidé před digitalizováním těchto informací určili náležité struktury pro tato data a tyto struktury důsledně dodržovali při převodu, byly by tato data už snadněji strojově zpracovatelná. Otázkou zůstává, zda je vůbec možné určit formální strukturu pro obecnou knihu a jak náročné by bylo toto provedení.

Opačným příkladem může například být zpracování faktur. V dnešní době se posílají buď jako jejich oskenovaná papírová podoba nebo dokonce jako fotografie z mobilního telefonu. Další běžná forma faktury je elektronická verze. Na první pohled by se dalo říci, že je tato forma snadno strojově zpracovatelná, ale není úplně tomu tak. Existuje velké množství programů na generování faktur bez jednotného formátu, z čehož plyne, že každá faktura je jinak strukturovaná, obsahuje jiné údaje či jsou jinak rozmístěné. Z toho důvodu nejsou snadno strojem zpracovatelná. Přívětivější možností je možnost generování QR kódu, který drží všechny potřebné informace v předem definované struktuře, kterou lze z faktury programově velice snadno zpracovat.

Dnes již téměř všechna odvětví lidské činnosti mají možnost pracovat s digitálními daty a mohou využívat nestrukturované nebo částečně strukturované zdroje digitálních informací. Příkladem takového zdroje je jedna z nejnavštěvovanějších webových stránek na světě,

otevřená internetová encyklopedie Wikipedie. Přibližně třetina článků na Wikipedii obsahuje tzv. infoboxy (1). To jsou specifické šablony, které vytvářejí přehlednou tabulku. Její záznamy mají zpravidla podobu klíče, ke kterému je přiřazena hodnota.

Cílem této bakalářské práce je vytvořit nástroj, který uživateli umožní konfigurovatelným způsobem získávat data z těchto struktur, na základě interakce s uživatelem je vhodným způsobem propojovat, vizualizovat je pro lepší přehled uživatele a výsledek uložit v dále použitelném formátu pro časovou osu Timeline (2).

2 Strukturovanost dat

Data můžeme obecně rozdělit podle míry jejich uspořádání na strukturovaná, částečně strukturovaná (semi-strukturovaná) a nestrukturovaná. V této kapitole je dále přiblížena každá z těchto skupin.

2.1 Strukturovaná data

Strukturovaná data se pevně drží jejich formální definice struktury, jsou tak uchovávána jak data samotná, tak i jejich vztahy, pokud jsou danou strukturou definovány. Příkladem jsou databáze. Podle datového modelu je dále můžeme dělit na (3):

- Hierarchický model – Data jsou rozdělena do jednotlivých úrovní (vrstev). Vazby mezi vrcholy jsou pouze mezi těmi o jednu úroveň níže nebo výše. Specifickým případem jsou stromy. Ty mají, oproti obecnému hierarchickému modelu, pouze jeden výchozí (na nejvyšší úrovni) vrchol, často nazývaný jako kořen. Podle typů stromů je můžeme dále dělit na binární stromy, n-stromy.

- Síťový model – Obecná grafová struktura, kde nejsou definované úrovně a vrcholy mohou mít mezi sebou vztahy libovolně.

- Relační model – Často využívaný model pro databáze. Data jsou ukládány do řádků tabulek, které mají definované vlastnosti těchto dat jako vlastnosti sloupců. Dále jsou definované vztahy mezi jednotlivými tabulkami.

2.2 Nestrukturovaná data

Nestrukturovaná data jsou lidmi psané texty v přirozeném jazyce, jde tedy o obecný text, kde struktura není definovaná či formálně popsaná, což komplikuje jejich strojové zpracování. Neznamená to ovšem, že při psaní těchto textů nejsou dodržována žádná pravidla. V takovém případě by nikdo, kromě autora, takovému textu neporozuměl. Nejedná se však o pravidla, které jsou naprosto jasně definována a musí být vždy dodržena, tak jako je to v případě strukturovaných dat, které jsou díky tomu strojově zpracovatelná.

Obecně u každého formálního jazyka můžeme definovat gramatiku. Gramatika určuje pravidla pro daný jazyk a skládá se z (4):

- Množiny terminálních symbolů – Jedná se o abecedu, ve které je jazyk psaný. Jde o neprázdnou množinu znaků, kterou nazýváme písmena, z těch se skládají slova daného jazyka.

- Množiny neterminálních symbolů – Představují určité mezistupně, se stejnými nebo podobnými vlastnostmi, které můžeme dále přepsat na další terminální nebo neterminální symboly. Můžeme za ně například považovat slovní druhy, větné členy.

- Počátečního symbolu – Jeden prvek z množiny neterminálních symbolů, který se přepisuje dále.

- Množiny přepisovacích pravidel – Jde o pravidla, která určují, jak lze počáteční symbol přepisovat dále na další neterminální či terminální symboly. Například každá věta v českém jazyce má základní větné členy – podmět a přísudek. Za podmět lze dosadit podstatné jméno, za přísudek sloveso.

Toto pojetí vychází z práce Syntaktický struktur Noama Chomského. Původní snaha byla použít tento přístup i pro zpracování přirozených jazyků, nicméně pro úplnou strojovou zpracovatelnost přirozených jazyků se jeví jako nedostatečný zejména pro „*nedostatečné respektování některých jazykových rovin a vztahů atd*“. (5)

2.3 Částečně strukturovaná data

Částečně strukturovaná (též semi-strukturovaná či polostrukturovaná) data jsou na pomezí mezi strukturovanými a nestrukturovanými daty. Na rozdíl od těch strukturovaných zcela nedrží definovanou strukturu, ale jsou doplněna o značky, anotace nebo klíčová slova, která data popisují. Případně alespoň dodržují nějaká konkrétní pravidla, jako například pořadí určeného typu dat, daná pro jistý formát. Doplněním těchto popisných informací je jim dána alespoň částečně struktura, která umožňuje jejich strojové zpracování. Jedná se například o značkovací jazyky XML a HTML nebo formáty JSON a CSV.

2.3.1 XML

XML (*extensible markup language*) je obecným značkovacím jazykem vytvořeným uznávaným konsorciem W3C, který vychází ze staršího jazyka SGML. Byl vyvinut jako platformě nezávislý formát pro výměnu informací. Mimo jeho snadnou strojovou zpracovatelnost je zamýšlený také jako formát snadno čitelný pro člověka. Párovými značkami (tagy) se ohraničují jednotlivé elementy. Tagy mohou mít atributy, které blíže specifikují daná data. Každý element může obsahovat další elementy, čímž vzniká stromová struktura dat. Jeho univerzálnost spočívá v tom, že neobsahuje předurčené tagy. Je tedy nutné si definovat tagy vlastní. (4) Mezi tagy mohou být další tagy či obecný nestrukturovaný text.

XMP Parsery

S parserem se lze nejčastěji setkat jako s ucelenou knihovnou, která poskytuje nástroje pro snadnou práci s dokumenty. Z hlediska zpracování XML souborů rozlišujeme dva základní přístupy:

- **DOM** (*document object model*) – Při tomto přístupu ke zpracování je soubor nejdříve celý načten. Vytvoří se hierarchická stromová struktura dokumentu, ve kterém každý uzel stromu odpovídá danému elementu v XML souboru. Strom lze libovolně procházet a vytvářet, upravovat či mazat jednotlivé uzly. U objektově orientovaných programovacích jazyků je možné, při znalosti schématu XML dokumentu, reprezentovat daný dokument třídou s příslušnými atributy. Ty pak odpovídají možným elementům XML souboru. (5)

- **SAX** *(simple API for XML*) – Jedná se o takzvaný „*event-driven*“ přístup. XML soubor je zpracováván sekvenčně a vždy po načtení elementu (případně skupiny elementů) je zavolána příslušná metoda či funkce, která mu odpovídá a reaguje na něj. Oproti DOM se tak vyznačuje vyšší rychlostí a menší náročností na paměť. SAX je tedy vhodný pro použití při zpracovávání velkých XML souborů.

2.3.2 JSON

JSON (*JavaScript Object Notation*) podobně jako XML byl vytvořen pro přenos dat nezávislý na platformě s lehce pro člověka čitelným i jednoduše strojově zpracovatelným formátem.

Využívá dvě základní struktury – objekt, kterým je kolekce párů klíč:hodnota a seřazené pole hodnot. Hodnoty mohou být z hlediska datového typu: objekt, řetězce znaků, čísla (celá i reálná), boolean (true/false) a null. (5) Je hojně využívaný zejména webovými službami a implementace knihoven pro práci s ním je dostupná ve značném množství programovacích jazyků.

2.3.3 HTML

Značkovací jazyk HTML (*Hypertext Markup Language*) vznikl už v roce 1990 pro tvorbu webových stránek. První definicí HTML specifikoval Tim Berners-Lee ve snaze usnadnit fyzikům vzájemnou komunikaci a sdílení výsledků jejich výzkumů. Podobně jako značkovací jazyk XML je dnes dále vyvíjený a udržovaný konsorciem W3C. Oproti XML má jasně definované tagy, které lze použít, a jejich atributy. Tagy mohou být párové i nepárové. Webové prohlížeče ho umí nativně zpracovávat. W3C také vyvinulo XHTML, tedy HTML definované pomocí XML, ve snaze zavést formát pro webové stránky, který je možné validovat. (7)

3 Otevřené zdroje informací

Následuje přehled vybraných otevřených zdrojů informací, které jsou dostupné online a mohly by být potenciálně použity jako zdroj pro získávání grafových dat.

3.1 Wikipedie

Otevřená internetová encyklopedie Wikipedie[1] je podle analytického webu společnosti SimilarWeb, Ltd.[2] sedmou nejnavštěvovanější webovou stránkou na světě. Je postavena na redakčním systému wiki, který umožňuje uživatelům přidávat nové články a upravovat stávající. Na tvorbě jejího obsahu se podílejí lidé z celého světa. Aktuálně je aktivních 310 jazykových mutací. Byla spuštěna 15. 1. 2001 a je v současné době vlastněna a spravována společností Wikimedia Foundation, Inc. Nejobsáhlejší jazykovou mutací je anglická, která má přes 6 milionů článků. Česká jich obsahuje přibližně dvanáctkrát méně. (6)

Někdy bývá považována za nedůvěryhodný zdroj informací, jelikož může kdokoliv provádět změny obsahů jednotlivých článků. Pokud se jedná o úmyslném vkládání nepravdivých informací či mazání užitečného obsahu, pak hovoříme o vandalismu. Wikipedia bojuje proti tomuto chování tím, že ukládá veškeré změny každého článku a kdokoliv, kdo zaznamená nežádoucí změnu, může obnovit předchozí verzi. Při sporných úpravách rozhodují správci o tom, která z úprav bude využita pro aktuální podobu článku. Při rozhodování mohou přihlédnout k diskuzi daného článku, kde může kdokoliv vyjádřit svůj názor na změny. Při každé editaci je navíc evidována IP adresa, z které k ní došlo, případně uživatelský účet, pokud byl použit. Při opakovaném vandalismu mají správci možnost zablokovat těmto uživatelům další případné úpravy článků. Správci mohou také omezit možnost editace článků, které se stávají častým terčem vandalství, že je mohou upravovat pouze sami správci nebo jen registrovaní uživatelé, kteří již provedli určitý počet úprav jiných článků. (7)

Články na Wikipedii jsou psané ve značkovacím jazyce Wikitext (někdy též uváděno jako Wiki markup či Wikicode). Značky přinášejí další popisné informace o jinak naprosto obecném textu. Jejich gramatika je formálně popsána (8), tudíž i snadno strojově zpracovatelná. Jedná se tedy o další příklad částečně strukturovaných dat. Uživatelé při

[1] https://www.wikipedia.org
[2] https://www.similarweb.com/

vytváření a úpravě článků mohou využít WYSIWYG[3] editor či editor, kde je použití Wikitextu přímo vidět. Oba editory mají zjednodušené rozhraní podobné používaným textovým procesorům MS Word či LibreOffice.

3.1.1 Infoboxy

Jedná se specifické šablony systému wiki, které vytvářejí přehlednou tabulku zpravidla v pravém horním rohu článku, mající za cíl přehledně zobrazovat souhrny informací, které mohou mít podobné typy článků společné a tím vylepšit možnosti navigace mezi nimi. Dodržování struktury infoboxů není striktně kontrolováno, je tedy na každém editorovi, zda doporučenou podobu infoboxu pro daný typ článku použije. (1) Následuje příklad struktury konkrétního infoboxu ve výpisu 3.1.

Výpis 3.1 Příklad struktury infoboxu - Charles IV, Holy Roman Emperor

```
{{Infobox royalty
| name        = Charles IV
| image       = Charles_IV-John_Ocko_votive_picture-fragment.jpg
| caption     =
| succession  = [[King of Bohemia]]
| reign       = 26 August 1346 – 29 November 1378
| predecessor = [[John of Bohemia|John]]
| successor   = [[Wenceslaus IV of Bohemia|Wenceslaus IV]]
| coronation  = 2 September 1347, [[Prague]]
| succession1 = [[King of the Romans]]<br>(Roman-German King)
| reign1      = 11 July 1346 – 29 November 1378
| predecessor1= [[Louis IV, Holy Roman Emperor|Louis IV]]
| successor1  = [[Wenceslaus IV of Bohemia|Wenceslaus IV]]
| coronation1 = 26 November 1346, [[Bonn]]
| succession2 = [[Holy Roman Emperor]], [[King of Italy]]
| reign2      = 1355 – 29 November 1378
| predecessor2= [[Louis IV, Holy Roman Emperor|Louis IV]]
| successor2  = [[Sigismund, Holy Roman Emperor|Sigismund]]
| coronation2 = {{plainlist|
* 6 January 1355, [[Milan]] (Italian)
* 5 April 1355, [[Rome]] (imperial)}}
| spouse      = {{plainlist|
* {{marriage|[[Blanche of Valois]]|1329|1348|end=d}}
* {{marriage|[[Anne of Bavaria]]|1349|1353|end=d}}
* {{marriage|[[Anna von Schweidnitz]]|1353|1362|end=d}}
* {{marriage|[[Elizabeth of Pomerania]]<br>|1363}}
}}
| issue       = {{plainlist|
* [[Wenceslaus, King of the Romans]]
* [[Sigismund, Holy Roman Emperor]]
* [[John of Görlitz]]
* [[Margaret of Bohemia, Queen of Hungary|Margaret, Queen of Hungary]]
* [[Catherine of Bohemia]]
* [[Elisabeth of Bohemia (1358–1373)|Elisabeth, Duchess of Austria]]
* [[Anne of Bohemia|Anne, Queen of England]]
* [[Margaret of Bohemia, Burgravine of Nuremberg|Margaret, Burgravine of
```

jednoduché pole hodnot

jednoduché pole hodnot

[3] WYSIWYG - What you see is what you get

```
Nuremberg]]}}
| house       = [[House of Luxembourg|Luxembourg]]
| father      = [[John of Bohemia]]
| mother      = [[Elisabeth of Bohemia (1292–1330)|Elisabeth of Bohemia]]
| birth_date  = 14 May 1316
| birth_place = [[Prague]]
| death_date  = 29 November 1378 (aged 62)
| death_place = [[Prague]]
| place of burial= [[St. Vitus Cathedral]], [[Prague]]
| religion    = [[Catholic Church|Roman Catholicism]]
}}
```

Jak můžeme vidět na příkladě výše, každý infobox je uvozený dvojicí levých – otevíracích složených závorek. Dále následuje klíčové slovo *Infobox* a jeho typ. Potom následují dvojice klíč (také atribut) a hodnota. Klíč je vždy uvozen svislítkem, mezi klíčem a hodnotou je znak rovná se. Ukončen je dvojicí pravých – zavíracích složených závorek. Klíče mohou být i několikaslovné a určitá podmnožina klíčů bývá stejná pro stejné typy infoboxů. Například u osob to bude jméno, místo a datum narození a úmrtí, rodiče, potomci, choť. Pokud je osoba ještě blíže specifikovaným typem, jako je infobox Karla IV. zde v ukázce typu royalty (člen královské rodiny), můžeme dále očekávat klíče jako rod, vláda, předchůdce, následovník a korunovace. Tyto klíče jsou očekávány, protože náleží dané šabloně. Jako hodnoty klíčů mohou být také uvedené složitější objekty. Na výše uvedeném příkladu výpisu 3.1 je zvýrazněn jeden z těchto typů objektů, a to jednoduché pole hodnot, které je zde dvakrát za sebou. Je uvozeno dvojicí otevíracích složených závorek, následuje klíčové slovo *plainlist*. Každá z hodnot tohoto pole je oddělena hvězdičkou. Pole je poté ukončeno dvojicí zavíracích složených závorek. V příkladu lze také vidět způsob zápisu odkazů. Ty se zapisují do dvojice hranatých závorek. Pokud se mezi nimi nachází i svislítko, pak odděluje zobrazovaný text od hodnoty odkazu. V části před svislítkem (až po dvojici otevíracích hranatých závorek) je uveden text odkazu, po něm (až ke dvojici zavíracích hranatých závorek) pak následuje text, který je zobrazován.

Na Wikipedii je také uveden rozcestník pro všechny šablony infoboxů. Obsahuje odkazy na jejich konkrétní předpisy a také počet použití šablon daného typu. (9) Šablona pro infoboxy obecné osoby je jednou z vůbec nejpoužívanějších, její předpis se základními parametry je uveden v následujícím výpisu 3.2 (10).

Výpis 3.2 Šablona pro infoboxy obecné osoby

```
{{Infobox person
| name        = <!-- use common name/article title -->
| image       = <!-- filename only, no "File:" or "Image:" prefix, and no enclosing [[brackets]] -->
| alt         = <!-- descriptive text for use by speech synthesis (text-to-speech) software -->
| caption     =
```

```
| birth_name    = <!-- only use if different from name -->
| birth_date    = <!-- {{Birth date and age|YYYY|MM|DD}} for living people supply only the year with
{{Birth year and age|YYYY}} unless the exact date is already widely published, as per [[WP:DOB]].
For people who have died, use {{Birth date|YYYY|MM|DD}}. -->
| birth_place   =
| death_date    = <!-- {{Death date and age|YYYY|MM|DD|YYYY|MM|DD}} (DEATH date then
                  BIRTH date) -->
| death_place   =
| nationality   =
| other_names   =
| occupation    =
| years_active  =
| known_for     =
| notable_works =
}}
```

Také existuje šablona pro obecné osoby se všemi parametry. Je doporučováno v šabloně nechat všechny základní parametry, i když není uvedená hodnota. V takovém případě se ve výsledném náhledu nezobrazí, ale zůstanou pro případné budoucí doplnění jejich hodnot. Ostatní klíče z rozšířené šablony mají být použity pouze tehdy, pokud jsou relevantní k dané konkrétní osobě a jejich hodnota bude vyplněna.

Hodnoty mohou být oproti klíčům různorodé. V příkladu infoboxu Karla IV. vidíme kromě jednoduchých textů, časové údaje s různým formátem zápisu, seznamy s výčtem dalších hodnot či odkazy na jiné stránky Wikipedie. Odkazy jsou ohraničeny dvojicí hranatých závorek. Pokud je mezi nimi svislítko, text před ním se použije jako odkaz a text za ním bude zobrazen. V některých případech může nastat situace, kdy hodnotou je celý další infobox. Příkladem je infobox kosmického letu Apollo 11, který jich takto vnořených obsahuje hned několik, jak lze vidět na výpisu 3.3. Také je běžné, že stránka obsahuje více infoboxů za sebou.

Výpis 3.3 Apollo 11 – příklad vnořených infoboxů (infobox jako hodnota)

```
{{Infobox spaceflight
| name              = Apollo 11
| image             = Aldrin Apollo 11 original.jpg
| image_caption       = [[Buzz Aldrin]] on the Moon as photographed by [[Neil Armstrong]]
(Armstrong seen in the visor reflection)
... zkráceno ...
| orbit_period        = 2 hours<ref name="orbit" />
| apsis             = cynthion
|interplanetary       =
 {{Infobox spaceflight/IP
  |type              = orbiter
  |object            = [[Moon|Lunar]]
  |component         = [[Apollo command and service module|Command and service module]]
  |orbits            = 30
  |arrival_date      = July 19, 1969, 17:21:50 UTC{{sfn|Orloff|2000|p=106}}
  |departure_date    = July 22, 1969, 04:55:42 UTC{{sfn|Orloff|2000|p=109}}
 }}
 {{Infobox spaceflight/IP
```

```
  |type              = lander
... zkráceno ...
  |surface_EVAs       = 1
  |surface_EVA_time    = 2 hours, 31 minutes, 40 seconds
 }}
| docking             =
 {{Infobox spaceflight/Dock
  | docking_target     = LM
  | docking_type       = dock
  | docking_date       = July 16, 1969, 16:56:03 UTC{{sfn|Orloff|2000|p=106}}
  | undocking_date     = July 20, 1969, 17:44:00 UTC{{sfn|Orloff|2000|p=107}}
  | time_docked        =
 }}
 {{Infobox spaceflight/Dock
  | docking_target     = LM ascent stage
  | docking_type       = dock
  | docking_date       = July 21, 1969, 21:35:00 UTC{{sfn|Orloff|2000|p=109}}
  | undocking_date     = July 21, 1969, 23:41:31 UTC{{sfn|Orloff|2000|p=109}}
  | time_docked        =
 }}
| crew_size           = 3
... zkráceno ...
| previous_mission     = [[Apollo 10]]
| next_mission         = [[Apollo 12]]
| programme            = [[Apollo program]]
}}
```

3.1.2 Přístupy k datům Wikipedie

Pro většinu uživatelů je nejběžnějším přístupem k Wikipedii její webová stránka. Na výchozí stránce se nachází rozcestník s výběrem deseti jazykových mutací s přímým odkazem na hlavní stránku dané verze. Následuje možnost vyhledávání klíčových slov s výběrem libovolné jazykové mutace. Dále zde jsou odkazy na ostatní projekty společnosti provozující Wikipedii a na stažení aplikace Wikipedie na Google Play a App Store.

Další možností je stažení obsahu celé Wikipedie formou tzv. data dumpu. Jejich základní dělení je podle jazyků. Dále se jednotlivé balíky rozdělují podle obsahu na dumpy s aktuální verzí článků bez diskuzí a uživatelských stránek, aktuální verzí článků s uživatelskými diskuzemi, pouze s úvody článků, pouze názvy článků a jejich přesměrování, SQL zálohy a dumpy s veškerou historií všech článků. Jde o velké komprimované soubory ve formátu XML. (11) Například soubor *enwiki-20210201-pages-articles-multistream.xml*, který obsahuje pouze aktuální verze článků psané v angličtině k 01. 02. 2021, je velký 76,1 GB, komprimovaný má 17,9 GB. K čtení dumpů mohou být použity offline programy, takovými jsou například *WikiTaxi* či *XOWA*. Následuje ukázka z uvedeného XML dumpu ve výpisu 3.4.

Obsahuje začátek odpovídající Wikipedie stránky Karla IV. v angličtině až po konec jeho infoboxu.

Výpis 3.4 Stránka Charles IV, Holy Roman Emperor v XML dumpu

```
<page>
  <title>Charles IV, Holy Roman Emperor</title>
  <ns>0</ns>
  <id>38895</id>
  <revision>
    <id>1000054104</id>
    <parentid>997081383</parentid>
    <timestamp>2021-01-13T09:24:04Z</timestamp>
    <contributor>
      <username>Grblomerth</username>
      <id>948536</id>
    </contributor>
    <minor />
    <comment>King of the Romans as Wenceslaus, not Wenceslaus IV</comment>
    <model>wikitext</model>
    <format>text/x-wiki</format>
    <text bytes="34639" xml:space="preserve">{{short description|14th century Holy Roman
Emperor of the House of Luxembourg}}
{{More footnotes|date=March 2010}}
{{Infobox royalty
| name        = Charles IV
| image       = Charles_IV-John_Ocko_votive_picture-fragment.jpg
| caption     =
| succession  = [[King of Bohemia]]
| reign       = 26 August 1346 – 29 November 1378
| predecessor = [[John of Bohemia|John]]
| successor   = [[Wenceslaus IV of Bohemia|Wenceslaus IV]]
| coronation  = 2 September 1347, [[Prague]]
| succession1 = [[King of the Romans]]&amp;lt;br /&amp;gt;(Roman-German King)
| reign1      = 11 July 1346 – 29 November 1378
| predecessor1= [[Louis IV, Holy Roman Emperor|Louis IV]]
| successor1  = [[Wenceslaus IV of Bohemia|Wenceslaus]]
| coronation1 = 26 November 1346, [[Bonn]]
| succession2 = [[Holy Roman Emperor]], [[King of Italy]]
| reign2      = 1355 – 29 November 1378
| predecessor2= [[Louis IV, Holy Roman Emperor|Louis IV]]
| successor2  = [[Sigismund, Holy Roman Emperor|Sigismund]]
| coronation2 = {{plainlist|
* 6 January 1355, [[Milan]] (Italian)
* 5 April 1355, [[Rome]] (imperial)}}
| spouse      = {{plainlist|
* {{marriage|[[Blanche of Valois]]|1329|1348|end=d}}
* {{marriage|[[Anne of Bavaria]]|1349|1353|end=d}}
* {{marriage|[[Anna von Schweidnitz]]|1353|1362|end=d}}
* {{marriage|[[Elizabeth of Pomerania]]&amp;lt;br /&amp;gt;|1363}}
}}
| issue       = {{plainlist|
* [[Wenceslaus, King of the Romans]]
* [[Sigismund, Holy Roman Emperor]]
* [[John of Görlitz]]
* [[Margaret of Bohemia, Queen of Hungary|Margaret, Queen of Hungary]]
* [[Catherine of Bohemia]]
```

```
* [[Elisabeth of Bohemia (1358–1373)|Elisabeth, Duchess of Austria]]
* [[Anne of Bohemia|Anne, Queen of England]]
* [[Margaret of Bohemia, Burgravine of Nuremberg|Margaret, Burgravine of Nuremberg]]}}
| house       = [[House of Luxembourg|Luxembourg]]
| father      = [[John of Bohemia]]
| mother      = [[Elisabeth of Bohemia (1292–1330)|Elisabeth of Bohemia]]
| birth_date  = 14 May 1316
| birth_place = [[Prague]]
| death_date  = 29 November 1378 (aged 62)
| death_place = [[Prague]]
| place of burial= [[St. Vitus Cathedral]], [[Prague]]
| religion    = [[Catholic Church|Roman Catholicism]]
}}
```

3.1.3 Zdroje informací související s Wikipedií

DBPedie

Cílem DBPedie je vytěžit všechna strukturovaná data z projektů nadace Wikimedia, mezi něž patří i Wikipedie. Z anglické verze se podařilo získat přibližně 4 a půl milionu záznamů, přičemž napříč všemi jazykovými verzemi (zpracováno celkem 125 jazyků) se jedná o 38,3 milionu záznamů.

Získaná data jsou ukládána ve schématu RDF, jehož základní prvek informace je tzv. trojice, která se skládá z podmětu, vlastnosti a předmětu. Přístup k nim je umožněn přes dotazovací jazyk SPARQL. Data jsou také dostupná přes webové stránky začínající adresou *http://dbpedia.org/page/* a končící stejným názvem jako ve Wikipedii. (12)

3.2 Encyclopædia Britannica

Jedná se o uznávanou původně Britskou encyklopedii vydávanou od 2. poloviny 18. století. Poslední tištěná verze byla vydána ve 32 svazcích s celkovým počtem 32 640 stan. Od roku 2012 již není publikována v tištěné verzi a je dostupná online[4]. Jelikož každý z jejích článků má jednoznačně dohledatelného autora či skupiny autorů, je považována za přesný a spolehlivý zdroj informací. (13)

[4] https://www.britannica.com

3.3 Internetová encyklopedie encyclopedia.com

Internetová encyklopedie *encyclopedia.com* obsahuje věrohodné články vydané vydavatelstvím *Oxford University Press* a americkou vzdělávací společností *Cengage*, dále zahrnuje články z encyklopedie *Columbia Encyclopedia*. Celkem je dostupných přes 300 tisíc článků. Podobně jako *Britannica* také neumožňuje široké veřejnosti úpravu svého obsahu, tím si udržuje svoji důvěryhodnost. (14)

4 Návrh nástroje

4.1 Specifikace zadání

Hlavním úkolem této práce je vytvořit nástroj, který umožní uživateli konfigurovatelným způsobem získávat grafová data z otevřeného zdroje. Na základě vstupu od uživatele data zobrazí a nabídne možnosti získání dalších dat, které s nimi souvisejí. Po výběru uživatele se pokusí vyhledat tato související data, veškeré entity a jejich vazby zobrazí. Vzhledem k podstatě dat a jejich propojení se nabízí jejich vizualizace jako graf či tabulka. Uživatel může svými zásahy doplňovat informace pro další získávání požadovaných dat. Dále tato data vyexportovat do podoby použitelné časovou osou *Timeline*. (2)

Jako nejvhodnější zdroj grafových dat se jeví anglická jazyková mutace Wikipedie. Zejména z důvodu její velké obsáhlosti a nejrychlejšího rozvoje obsahu. Dalším, neméně významným, důvodem je existence infoboxů, ze kterých lze automatizovaným postupem získávat potřebné informace o vazbách mezi jednotlivými entitami.

Aplikace umožní uživateli vyhledávat další entity z té výchozí pomocí klíčů, které si uživatel vybere z navržené množiny. Tato množina bude získána prohledáním infoboxu náležející stránce, která odpovídá vybrané entitě. Již nalezené entity uživateli zobrazí a umožní mu z každé z nich dále vyhledávat další entity podle navržených klíčů. Pokud by data, podle kterých se entity propojují, byla chybná či úplně chyběla, umožní program uživateli tato data doplnit.

Na Wikipedii je mnoho stran, obsahujících stejný typ či podtyp šablony infoboxu, tedy budou obsahovat stejné klíče. Z toho důvodu bude možné tyto entity propojovat. Jelikož entity typu osoba patří na Wikipedii k jedněm z nejzastoupenějších, bude se tento program zaměřovat primárně na práci s nimi. V případě osob mají hodnoty klíčů většinou také jako hodnotu stránku jiné osoby. Pokud tato stránka také obsahuje infobox, bude také s vysokou pravděpodobností obsahovat stejnou podmnožinu klíčů (editoři stránky nemusejí uvést všechny klíče podle dané šablony, protože k nim například není hodnota známá). Díky této skutečnosti bude možné sestavovat i relativně rozvětvené grafy.

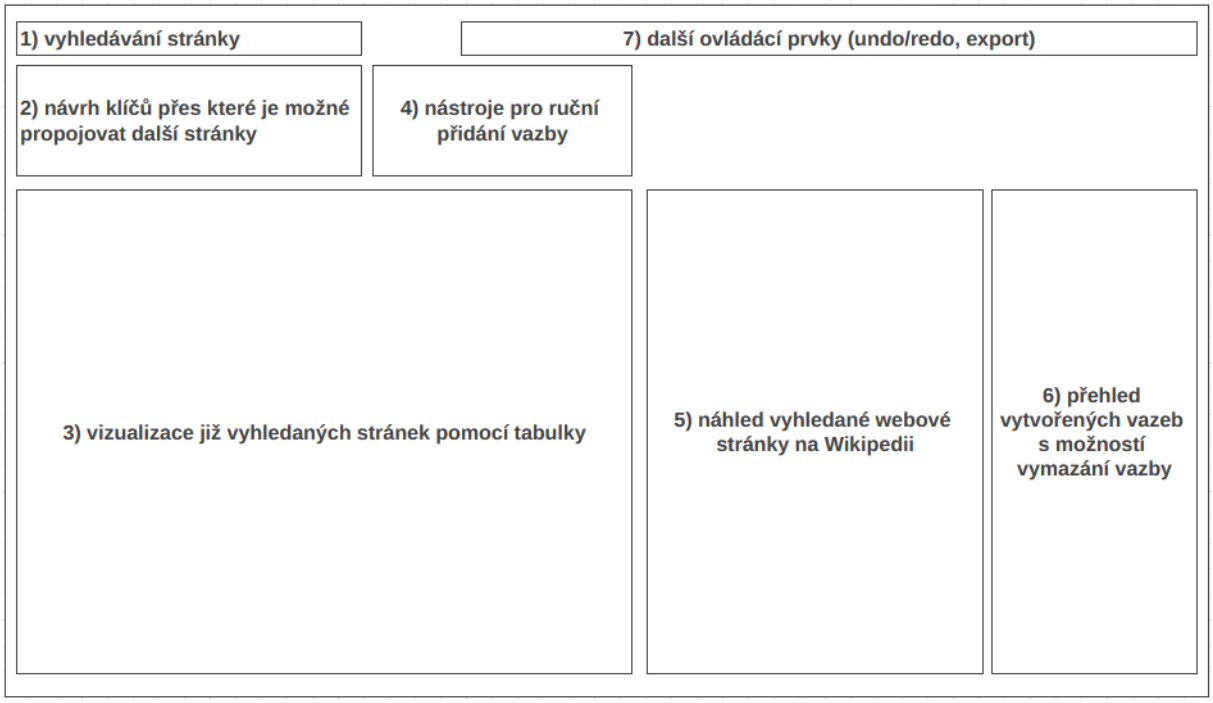
Typickým příkladem použití tohoto nástroje je sestavování rodokmenů historických rodů. Počátečním bodem je libovolný člen rodové linie. Od něj se začne tvořit graf přidáváním hran

typu *Parents*, *Parent(s)* či ve dvou krocích typy *Mother* a *Father*. Druhým směrem je graf sestavován přes atributy *Issue*, *Issue(s)* nebo *Children* či *Children(s)*. Tímto způsobem je možné aplikaci použít například pro relativně snadné vyhledání žijících přímých potomků historických osobností.

Aplikace umožní uživateli zobrazit náhled na webovou stránku Wikipedie dané entity.

4.2 Interakce s uživatelem

Pro snadnější sestavování grafu se bude aplikace ovládat pomocí GUI. V něm bude mít uživatel možnost určit, podle jakých klíčů se bude graf sestavovat. I když je záměrem proces sestavování co nejvíce zautomatizovat, přepokládají se zásahy uživatele i při jeho vytváření, zejména z důvodů nedodržení struktur ve vstupních datech (například použití jiných klíčů než předepisuje šablona, neaktualizovaná šablona používající starou podobu klíče atd.). Návrh GUI je na následujícím obrázku 4.1.



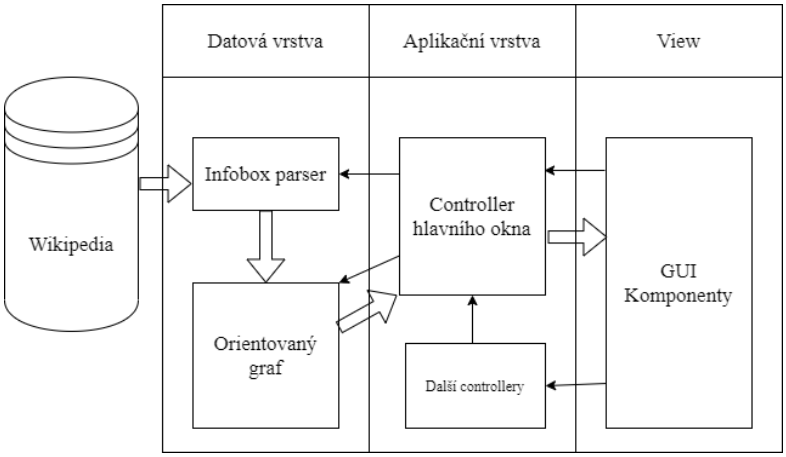
Obrázek 4.1 Návrh GUI

4.3 Architektura aplikace

Jak vyplývá z předchozích kapitol, zdrojem dat bude anglická jazyková mutace Wikipedie. Data je možné získávat za běhu aplikace při sestavování grafu uživatelem online přímo ze stránek Wikipedie v podobě zdrojového HTML kódu stránky. Další možností je stažení XML

dumpu a jeho předzpracování do databáze. Tento přístup je výhodnější zejména při sestavování větších grafů z důvodu rychlosti, také zde odpadá riziko dočasného zákazu další dotazů, které je sice malé, ale při online přístupu existuje.

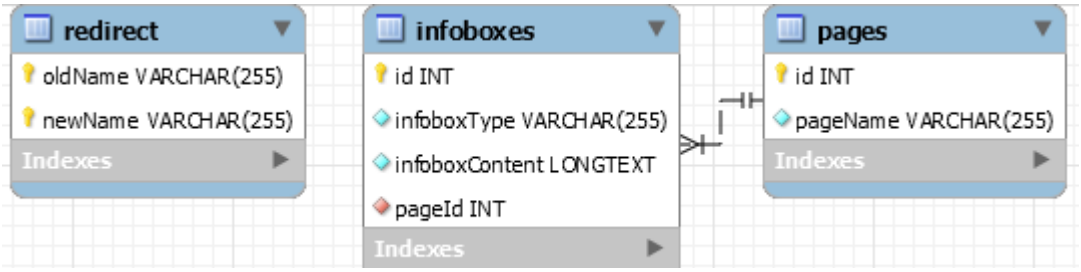
Při sestavování grafu budou data, z nichž bude tvořen, ukládána do orientovaného grafu. Informace o jednotlivých stránkách budou vrcholy tohoto grafu a hrany budou reprezentovat vztahy mezi nimi. Uživatel bude s programem interagovat prostřednictvím grafického uživatelského rozhraní (GUI). Požadavky převezme controller hlavního okna, ten je dále předá parseru infoboxu, pokud se bude jednat o požadavek o nová data, nebo vyžádá již zpracovaná data od vrstvy spravující orientovaný graf. Diagram architektury je na následujícím obrázku 4.2.



Obrázek 4.2 Diagram architektury aplikace

4.4 Databázové schéma

Data budou uložena do databáze z Wikipedia dumpu pomocí importovací části programu. Odhadované množství importovaných řádků je v řádech miliónů pro každou tabulku. Z toho důvodu byla vybrána relační databáze MySQL, protože je zdarma a je vhodná pro práci s tímto rozsahem dat. Do databáze budou ukládány následující údaje: název stránky, typ infoboxu, obsah infoboxu a přesměrování stránek. Relační schéma databáze je uvedeno na následujícím obrázku 4.3.



Obrázek 4.3 Databázové schéma

Tabulka *pages* obsahuje název stránky Wikipedie a primárním klíčem je celočíselný identifikátor. Každému záznamu z této tabulky může náležet vícero záznamů z tabulky *infoboxes*. Ta uchovává záznamy jednotlivých nalezených infoboxů. Propojení je zajištěno pomocí cizího klíče *pageId*, který odpovídá id z tabulky *pages*. Dále tabulka *infoboxes* uchovává celý text infoboxu pod atributem *infoboxContent* a typ infoboxu pod *infoboxType*. Primárním klíčem každého záznamu je také celočíselný identifikátor.

Dále tato databáze bude obsahovat tabulku *redirect*. Ta uchovává nalezené záznamy o přejmenování stránek Wikipedie. Atribut *oldName* obsahuje starý název stránky, *newName* potom nový název, na který byla stránka přejmenována. Tato tabulka je důležitá z toho důvodu, že v neaktualizovaných textech infoboxů bývají odkazy na staré názvy stránek.

Pro rychlé vyhledávání, po naplnění tabulek daty, budou vytvořeny indexy nad sloupečky, podle kterých se bude vyhledávat. Typické vyhledávání v této databázi pro potřeby programu bude následující:

1. Získání *id* stránky na základě jejího názvu *pageName*.

2. Nalezení textu a typu infoboxu na základě získaného *id* stránky (cizí klíč *pageId* v tabulce *infoboxes*)

3. Pokud stránka podle jejího názvu nebude nalezena v tabulce *pages*, bude následovat dotaz na její možný název v tabulce *redirect*. Dokud budou nové výsledky *newName*, zpět na 1.

Sloupce *oldName a newName*, mají index vytvořený jako součást primárního klíče. Dalším sloupcem, přes který se vyhledává, je *pageName* tabulky *pages*, ten však není součástí

primárního klíče, proto nad ním musí být index posléze vytvořen, stejně tak jako nad sloupci *pageName* a *pageId*.

5 Implementace nástroje

Celý nástroj je rozdělený na dva samostatné programy. Pro implementaci obou byla použita technologie Java. První program nazvaný *DatabaseInserter* slouží k zpracování dumpu Wikipedie a naplnění databáze požadovanými daty. Druhým programem, který umožňuje uživateli data z databáze spojovat na základě jeho požadavků a dále exportovat, je *LinkingApp*.

5.1 DatabaseInserter

Program je členěn do 3 balíčků: *database, dump, main*.

V balíčku *main* se nachází třída *Main*, která obstarává inicializaci ostatních základních tříd potřebných pro běh programu a také kontroluje požadované vstupy od uživatele. Dále je zde knihovní třída *Constants*. Ta obsahuje pouze veřejné statické atributy, představující konstanty použité v programu. Konkrétně je zde definovaný název konfiguračního souboru, který je možný využít místo předání parametrů programu z příkazové řádky (či terminálu), otevírací a ukončovací tagy pro úsek textu reprezentující jednotlivou stránku Wikipedie (<*page*> a </*page*>) a název dané stránky (<*title*> *a* </*title*>), úvodní řetězec každého infoboxu ({{*Infobox*), znak oddělující jednotlivé hodnoty v infoboxu („|“), úvodní sekvence pro přesměrování stránky (<*redirect title*=) a ukončení tohoto tagu („/>“). Toto přesměrování je na Wikipedii využito při přejmenování konkrétní stránky na jiný nový název, přičemž odkazy na ostatních stránkách Wikipedie odkazují ještě na jméno staré.

Druhý balíček má název *dump* a obsahuje pouze jednu třídu, kterou je *DumpReader*. Ta je inicializována v třídě *Main*. V konstruktoru přijímá jeden parametr a tím je cesta k souboru XML dumpu. Jejím účelem je číst tento soubor řádek po řádku pomocí třídy *BufferedReader* z balíku *java.io* a po zavolání veřejné metody *getNextPage()* vrací list s řádky odpovídající další nalezené stránce Wikipedie.

Posledním balíčkem je *database*. V něm jsou třídy *DatabaseHandler*, *DatabaseInserter* a pomocná třída *InfoboxEntry*. *DatabaseHandler* je incializován v třídě *Main* a vyžaduje jako parametry tyto údaje: url databáze, uživatelské jméno k databázi a heslo pro daný uživatelský účet, případně cestu ke konfiguračnímu souboru, kde jsou tyto údaje uvedeny. Pomocí třídy *DriveManager* z balíku *java.sql* vrací vytvořené připojení k databázi splňující požadavky

rozhraní *Connection* z téhož balíku. Třída *DatabaseInserter* v konstruktoru vyžaduje instance tříd *DatabaseHandler a DumpReader*. Probíhá zde parsování jednotlivých stran Wikipedie a vkládání dat do databáze pomocí vytvořeného připojení z třídy *DatabaseHandler* a dalších tříd z balíku *java.sql*. Pokud je na stránce nalezen alespoň jeden infobox, je vložena do databáze. Dále se pomocí parity dvojice složených závorek určí hranice infoboxu, vyextrahuje se jeho typ a obsah v podobě dvojic klíč:hodnota a vloží se také do databáze. Pro usnadnění předávání těchto dvojic a obsahu slouží třída *InfoboxEntry*. Jestliže na stránce není nalezen infobox žádný, do databáze se nic neukládá a je načtena další stránka.

5.2 LinkingApp

Pro usnadnění ovládání je tento program ovládán pomocí GUI. Na tvorbu grafického uživatelského rozhraní byla využita možnost jeho deklarativní tvorby pomocí FXML souborů. K jejich sestavení byl použit nástroj *Scene Builder* od organizace Gluon. (15) Funkce grafických komponent obsluhují příslušné controllery.

Zdrojový kód tohoto programu je také rozdělen do 3 balíčků nazvaných *datahandlers*, *fxmlcontrollers*, *main*. Navíc je ještě potřeba přilinkovat resource složku s FXML soubory.

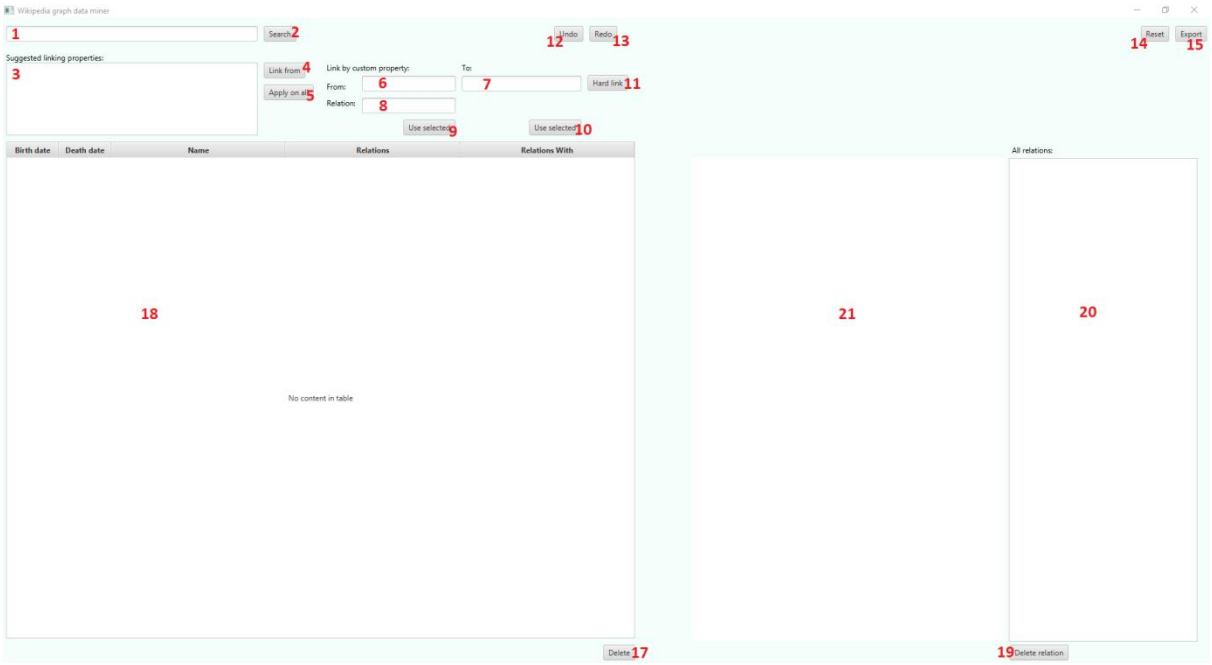
Balíček *main* obsahuje třídu *Main*, která je zodpovědná za vytvoření základní scény a načtení příslušného controlleru. Dále je zde třída *Constants* s konstantami použitými v programu.

V balíčku *fxmlcontrollers* jsou všechny controllery odpovídající FXML souborům. Nejpodstatnější z nich je *MainWindowController*, který implementuje všechny funkce hlavního okna.

Poslední balíček se nazývá *datahandlers*. Třídy *DatabaseConnectionCreator* a *DatabaseHandler* v něm obsažené zabezpečují vytvoření spojení s databází a komunikaci s ní. Třída *JgraphtDataHandler* využívá knihovny JGraphT (18) pro práci s orientovaným grafem, který uživatel postupně na základě svých požadavků vytváří. Dále zde jsou pomocné třídy *InfoboxEntry*, *InfoboxLink*, *KeyWithProperty*, *TableRowInfoboxData* pro předávání potřebných dat a třída *TimelineExporter*, jejíž funkce je popsána v podkapitole 5.2.2.

5.2.1 Interakce s uživatelem

Uživatel program ovládá přes GUI. Popis jednotlivých prvků a funkcionalit je na následujícím obrázku 5.1.



5.1 Popis prvků GUI

1. vyhledá stránku Wikipedie s infoboxem se zadaným jménem
2. pole pro zadání vyhledávané stránky s infoboxem
3. přehled klíčový atributů infoboxu
4. začne se tvořit graf ze zvoleného infoboxu přes zvolený atribut
5. začne se tvořit graf ze všech již nalezených infoboxů přes zvolený atribut
6. ruční přidání vazby – název stránky, ze kterého má být link veden
7. ruční přidání vazby – název stránky, ke kterému má být link veden
8. ruční přidání vazby – název ručně vytvořeného linku
9. ruční přidání vazby – předvyplní 6. a 8. na základě výběru textu z 23.
10. ruční přidání vazby – předvyplní 7. na základě výběru textu z 23.
11. vytvoří vazbu mezi stránkami zadané v 6. a 7. přes zadaný atribut v 8.
12. undo – vrátí stav o krok zpět
13. redo – vrátí stav o krok vpřed
14. zahodí všechna doposud nalezená a zadaná data
15. vytvoří souboru ve formátu pro časovou osu
16. zobrazí okno pro editaci datumů
17. vymaže řádek vybraný v 18.
18. tabulka zobrazující nalezené stránky a jejich vazby, případně datumy
19. vymaže řádek vybraný v 20.
20. přehled všech vazeb
21. náhled stránky vybrané v 20.

Všechny získané klíče pro daný infobox, které jsou uložené v databázi, se zobrazí uživateli v seznamu vlastností infoboxu navržených pro další linkování (bod 3.). Po vybrání klíče může uživatel rozhodnout, zda má být vyhledávání dalších stránek provedené jen z právě vybraného (4.) nebo ze všech již nalezených (5.). Veškeré nalezené stránky, na které vedly nalezené odkazy, jsou v tabulce (20.), která mimo jejich názvů zobrazuje jejich vazby s ostatními. Také je zde uveden datum jejich narození a smrti, jelikož je nástroj zamýšlen zejména pro osoby. Část programu parsující datumy je převzata z bakalářské práce Gabriely Hessové, která proklamuje úspěšnost 98%-100% (v závislosti na provedeném testu) (18). Uživateli je dále umožněno datumy editovat – přímo v tabulce, po dvojkliku na příslušnou buňku. Dále uživatel pomocí GUI může smazat již nalezené stránky (19.), ručně stránky přidávat i vazby mezi nimi (6. až 12.) a mazat vazby již vytvořené (21.). Pro usnadnění manuálního přidávání dalších stránek je zobrazován náhled právě zvolené stránky (23.). K tomu je využita instance třídy Webview z balíku *javafx.scene.web*.

5.2.2 Exporter pro časovou osu Timeline

Další možností, kterou uživatel má, je exportování aktuálního stavu do souboru ve formátu pro časovou osu Timeline (2). Jedná se o JSON formát, ve kterém jsou dvě pole – první obsahuje jednotlivé uzly grafu, druhé hrany mezi nimi. Každý uzel je reprezentován jako objekt hodnot: id – unikátní celočíselný identifikátor; stereotype – typ, v tomto případě osoba, tedy *person*; description – libovolný popis uzlu; begin – v tomto případě datum narození osoby; end – datum úmrtí osoby; properties – ostatní vlastnosti uzlu, například *startPrecision* a *endPrecision* – míra přenosti pro hodnoty begin a end. Datumy musí být uvedeny ve formátu YYYY-MM-DD.

6 Testy úplnosti a přesnosti

Při provádění testů úplnosti a přesnosti byla porovnávána data získané aplikací vůči odpovídajícím datům v infoboxech ve Wikipedii. Míra úplnosti vyjadřuje poměr mezi nalezenými stránkami z výchozí stránky, po vytvoření linků na další stránku podle vybraných klíčů, vůči očekávaným. Přesnost určuje, zda nalezená stránka odpovídá požadované. Rovněž bude posouzena přesnost získaných datumů jejich narození a smrti.

6.1 Albert Einstein – genealogie školitelů

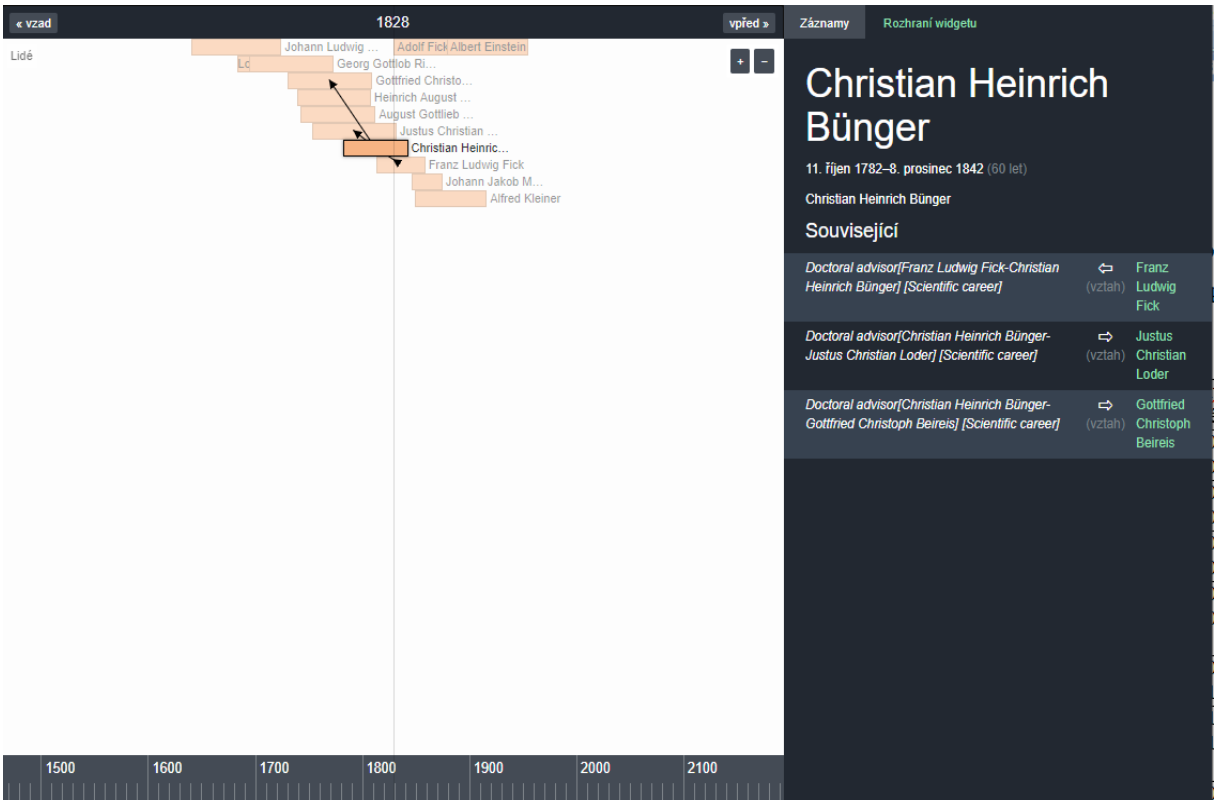
V první testovací sadě je výchozí osobností Albert Einstein. Od něj je sestavován graf pomocí klíče *Doctoral advisor*, je tedy od něj sestavována genealogie školitelů disertačních prací. Přehled je na následující tabulce 6:1.

Tabulka 6:1

Výchozí hodnota	Očekáváná hodnota	Obdržená hodnota	Data očekávané	Data získané
Albert Einstein	Alfred Kleiner	Alfred Kleiner	1879-03-14 1955-04-18	1879-03-14 1955-04-18
Alfred Kleiner	Johann Jakob Müller	Johann Jakob Müller	1849-04-24 1916-07-03	1849-04-24 1916-07-03
Johann Jakob Müller	Adolf Eugen Fick	Adolf Eugen Fick	1846-03-04 1875-01-14	1846-03-04 1875-01-14
Adolf Eugen Fick	Franz Ludwig Fick	Franz Ludwig Fick, link na prázdnou hodnotu	1829-09-03 1901-08-21	1829-09-03 1901-08-21
Franz Ludwig Fick	Christian Heinrich Bünger	Christian Heinrich Bünger	1813-05-18 1858-12-31	1813-05-18 1858-12-31
Christian Heinrich Bünger	Justus Christian Loder; Gottfried Christoph Beireis	Justus Christian Loder; Gottfried Christoph Beireis	1782-10-11 1842-12-08	1782-10-11 1842-12-08
Justus Christian Loder	Heinrich August Wrisberg, August Gottlieb Richter	Heinrich August Wrisberg, August Gottlieb Richter	1753-03-12 1832-04-16	1753-03-12 1832-04-16
Gottfried Christoph Beireis	Lorenz Heister	Lorenz Heister	1730-03-02 1809-09-18	1730-03-02 1809-09-18
Heinrich August Wrisberg			1739-06-20 1808-03-29	1739-06-20 1808-03-29
August Gottlieb Richter	Georg Gottlob Richter	Georg Gottlob Richter	1742-04-13 1812-07-23	1742-04-13 1812-07-23
Lorenz Heister			- -	- -
Georg Gottlob Richter	Johann Ludwig Hannemann	Johann Ludwig Hannemann	1694-02-04 1773-05-28	1694-02-04 1773-05-28
Johann Ludwig Hannemann			1640-10-25 1724-10-25	1640-10-25 1724-10-25

Očekáváno bylo 12 osob a všechny byly nalezeny. Také všechny další očekávané hodnoty byly nalezeny a shodují se s očekávanými. Přesnost i úplnost vyhledaných osob i dat je 100%. Všechny hodnoty byly získání bez potřeby zásahu uživatele, kromě úvodního vyhledání výchozího infoboxu a zadání linkovacího klíče.

Výsledná sestavována genealogie školitelů po exportu a jejím zobrazení v časové ose je na následujícím obrázku 6.1. Pro Lorenza Heistera byly data doplněny pro potřeby exportu z textu jeho stránky na Wikipedii. Jelikož ta neobsahuje infobox, nebyly tyto hodnoty očekávány.



Obrázek 6.1 Albert Einstein – genealogie školitelů v časové ose

6.2 Předkové Karla IV.

Ve druhém testovacím scénáři bude sestavován strom předků Karla IV. do roku 1200 (podle roku narození, pokud není znám, pak o jednu úroveň dále z té, kde je naposledy znám). Bude tedy od něj sestavován strom podle klíčových atributů *Mother* a *Father*.

Následující tabulka 6:2 obsahuje přehled hodnot.

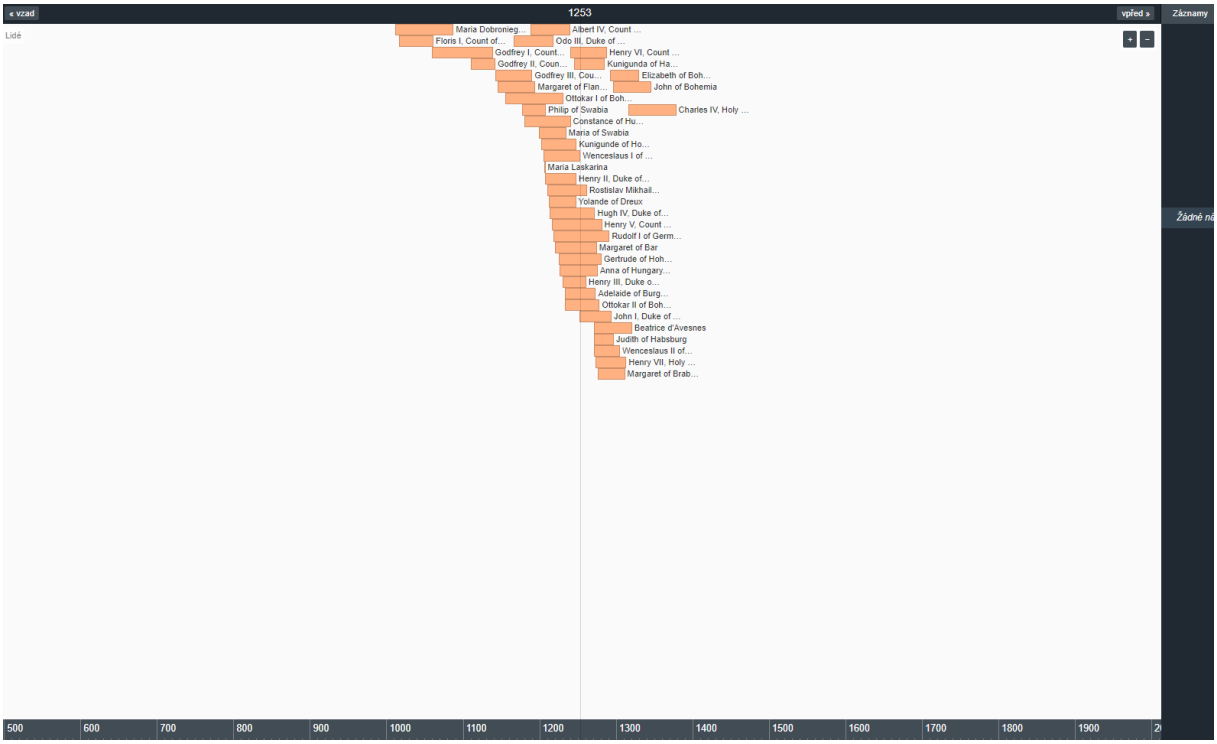
Tabulka 6:2 Předci Karla IV.

Výchozí hodnota	Očekáváná hodnota	Obdržená hodnota	Data očekávané	Data získané
Charles IV, Holy Roman Emperor	Elizabeth of Bohemia (1292–1330); John of Bohemia	Elizabeth of Bohemia (1292–1330); John of Bohemia	1316-05-14 1378-11-29	1316-05-14 1378-11-29
Elizabeth of Bohemia (1292–1330);	Judith of Habsburg; Wenceslaus II of Bohemia	Judith of Habsburg; Wenceslaus II of Bohemia	1292-01-20 1330-09-28	1292-01-20 1330-09-28
John of Bohemia	Margaret of Brabant; Henry VII, Holy Roman Emperor	Margaret of Brabant; Henry VII, Holy Roman Emperor	1296-08-10 1346-08-26	1296-08-10 1346-08-26
Judith of Habsburg	Gertrude of Hohenberg; Rudolf I of Germany	Gertrude of Hohenberg; Rudolf I of Germany	1271-03-13 1297-05-21	1271-03-13 1297-05-21
Wenceslaus II of Bohemia	Kunigunda of Slavonia; Ottokar II of Bohemia	Kunigunda of Halych (redirect z Kunigunda of Slavonia); Ottokar II of Bohemia	1271-09-27 1305-06-21	1271-09-27 1305-06-29
Margaret of Brabant	Margaret of Flanders (d. 1285); John I, Duke of Brabant	Margaret of Flanders (d. 1285); John I, Duke of Brabant	1276-10-04 1311-12-14	1276-10-04 1311-12-14
Henry VII, Holy Roman Emperor	Beatrice d'Avesnes; Henry VI of Luxembourg	Beatrice d'Avesnes; Henry VI of Luxembourg	1273-01-01 1313-08-24	1273-06-04 1313-08-24
Gertrude of Hohenberg	Matilda of Tübingen; Burkhard V, Count of Hohenberg	Matilda of Tübingen; Burkhard V, Count of Hohenberg	1225-01-01 1281-02-16	1225-01-01 1281-02-16
Rudolf I of Germany	Hedwig of Kyburg; Albert IV, Count of Habsburg	Hedwig of Kyburg; Albert IV, Count of Habsburg	1218-05-01 1291-07-15	1218-05-01 1291-07-15
Kunigunda of Halych	Anna of Hungary (b.1226); Rostislav Mikhailovich	Anna of Hungary (b.1226); Rostislav Mikhailovich	1245-01-01 1285-09-09	1245-01-01 1285-09-09
Ottokar II of Bohemia	Kunigunde of Hohenstaufen; Wenceslaus I of Bohemia	Kunigunde of Hohenstaufen; Wenceslaus I of Bohemia	1233-01-01 1278-08-26	1233-01-01 1278-08-26
Margaret of Flanders, Duchess of Brabant (redirect z Margaret of Flanders (d. 1285))	Matilda of Béthune; Guy of Dampierre	Sibylla of Anjou; Thierry, Count of Flanders	1251-01-01 1285-07-03	1145-01-01 1194-11-15
John I, Duke of Brabant	Adelaide of Burgundy (1233–1273); Henry III, Duke of Brabant	Adelaide of Burgundy (1233–1273); Henry III, Duke of Brabant	1252-01-01 1294-05-03	- 1294-05-03
Beatrice d'Avesnes	Felicitas of Coucy; Baldwin of Avesnes	- -	- 1321-02-25	- -
Henry VI of Luxembourg	Margaret of Bar; Henry V, Count of Luxembourg	Margaret of Bar; Henry V, Count of Luxembourg	1240-01-01 1288-06-05	1240-01-01 1288-06-05

Albert IV, Count of Habsburg	Agnes of Staufen; Rudolph II, Count of Habsburg	- -	1188-01-01 1239-12-13	- -
Anna of Hungary, Duchess of Macsó (redirect z Anna of Hungary (b.1226))	Maria Laskarina; Béla IV of Hungary	Maria Laskarina; Béla IV of Hungary	1226-01-01 -	1226-01-01 -
Rostislav Mikhailovich	Elena Romanovna of Halych; Mikhail Vsevolodovich	- -	1210-01-01 1262-01-01	- -
Kunigunde of Hohenstaufen	Irene Angelina; Philip of Swabia	Irene Angelina; Philip of Swabia	1202-02-01 1248-09-13	1202-02-01 1248-09-13
Wenceslaus I of Bohemia	Constance of Hungary; Ottokar I of Bohemia	Constance of Hungary; Ottokar I of Bohemia	1205-01-01 1253-09-23	1205-01-01 1253-09-23
Matilda of Béthune	Elisabeth of Morialmé; Robert VII, Lord of Béthune	- -	- 1263-11-08	- -
Guy of Dampierre	Margaret II of Flanders; William II of Dampierre	- -	1226-01-01 1305-03-07	- -
Adelaide of Burgundy, Duchess of Brabant (redirect z Adelaide of Burgundy (1233–1273))	Yolande of Dreux; Hugh IV, Duke of Burgundy	Yolande of Dreux; Hugh IV, Duke of Burgundy	1233-01-01 1273-10-23	1233-01-01 1273-10-23
Henry III, Duke of Brabant	Marie of Hohenstaufen; Henry II, Duke of Brabant	Maria of Swabia (redirect z Marie of Hohenstaufen); Henry II, Duke of Brabant	1230-01-01 1261-02-28	- 1261-02-28
Margaret of Bar	Philippa of Dreux; Henry II of Bar	- -	1220-01-01 1275-01-01	- -
Robert VII, Lord of Béthune	Mathilda of Dendermonde; William II, Lord of Béthune	- -	1201-01-01 1248-11-12	- -
Yolande of Dreux (Burgundy)	Alianor de St. Valéry; Robert III of Dreux	- -	1212-01-01 1248-01-01	- -
Hugh IV, Duke of Burgundy	Alice of Vergy; Odo III, Duke of Burgundy	Alice of Vergy; Odo III, Duke of Burgundy	1213-03-09 1272-10-27	1213-03-09 1272-10-27

Očekáváno bylo 56 osob, z toho jich bylo nalezeno 40. Odkazy všech osob směřovaly na správné stránky. Úplnost vyhledaných osob je přibližně 71,4% a přesnost je 100%. Data se očekávalo 53 a bylo jich získáno 37. Úplnost nalezených dat je tedy 69,8%. Pokud datum bylo nalezeno, pak odpovídalo očekávanému. Přesnost dat je 100%.

Vizualizace sestaveného stromu je na následujícím obrázku 6.2.



Obrázek 6.2 Předci Karla IV.

6.3 Historie verzí Microsoft Windows

Pro otestování možnosti získávat tímto nástrojem grafová data pro entity, na jejichž stránkách na Wikipedii nebude infobox typu osoba, byl vybrán testovací scénář, při kterém bude sestavována historie verzí operačního systému Microsoft Windows. Výchozím bodem bude verze MS Windows 10 a od ní bude graf sestavován zpět do historie pomocí klíče *„Preceded by“*. Jelikož se nejedná o osoby, nebude posuzován zisk dat narození a úmrtí. Přehled očekávaných a obdržených hodnot je v následující tabulce 6:3.

Tabulka 6:3 Historie verzí MS Windows

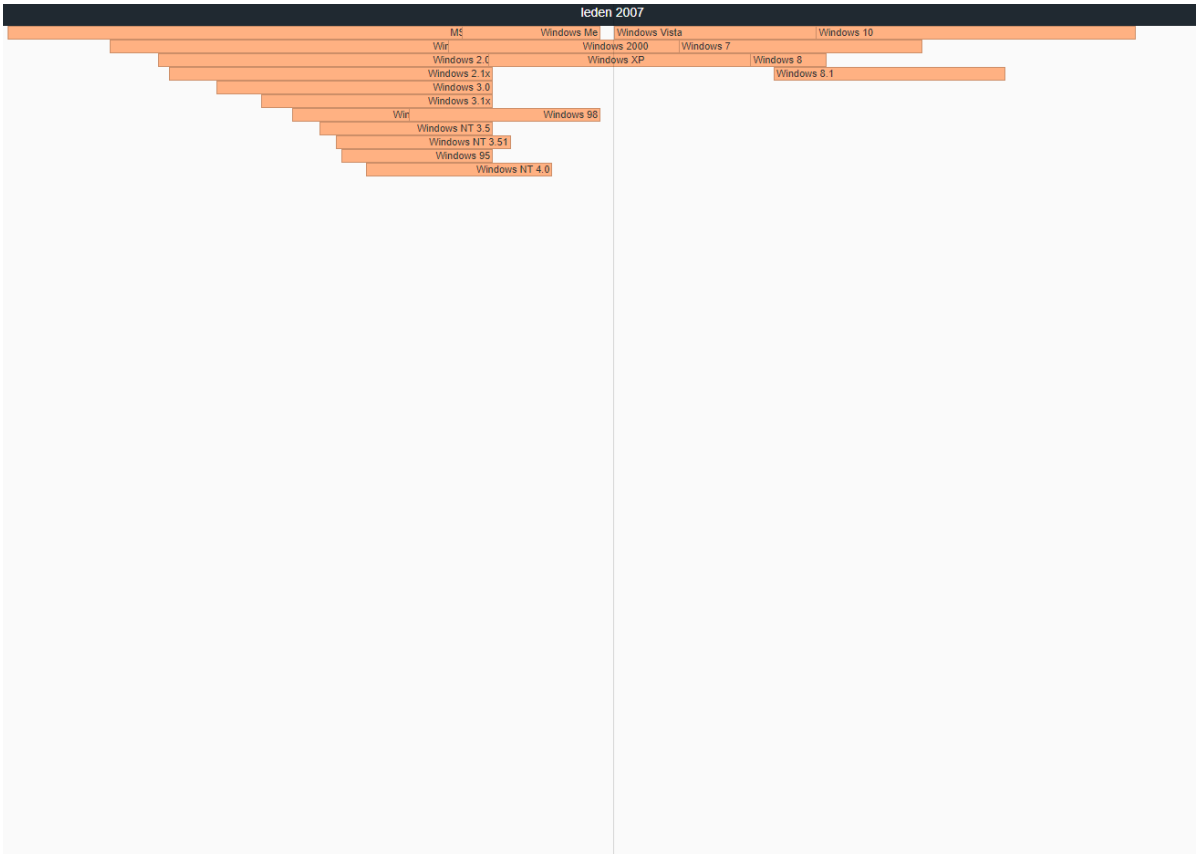
Výchozí hodnota	Očekávána hodnota	Obdržená hodnota
Windows 10	Windows 8.1	Windows 8.1
Windows 8.1	Windows 8	-
Windows 8	Windows 7	Windows 7
Windows 7	Windows Vista	Windows Vista; Buzzle
Windows Vista	Windows XP	Windows XP
Windows XP	Windows 2000; Windows Me	Windows 2000; Windows Me
Windows 2000	Windows NT 4.0	Windows NT 4.0
Windows Me	Windows 98	Windows 98

Windows NT 4.0	Windows NT 3.51	Windows NT 3.51
Windows 98	Windows 95	Windows 95
Windows NT 3.51	Windows NT 3.5	Windows NT 3.5
Windows 95	Windows 3.1x	Windows 3.1x
Windows NT 3.5	Windows NT 3.1	Windows NT 3.1
Windows 3.1x	Windows 3.0	Windows 3.0
Windows NT 3.1	Windows 3.1x	Windows 3.1x
Windows 3.0	Windows 2.1x	Windows 2.1x
Windows 2.1x	Windows 2.0	Windows 2.0
Windows 2.0	Windows 1.0	Windows 1.0
Windows 1.0	MS-DOS	MS-DOS

Pro obdržení všech devatenácti požadovaných hodnot, byl vyžadován zásah uživatele. Stránky v tomto scénáři využívají šablon typů *{{Infobox OS version}}* a *{{Infobox OS}}*. Obě podporují klíč *„preceded_by“*. Ta druhá zmíněná však dovoluje navíc i zápis klíče pro tento typ hodnoty jako *„preceded by“*. Proto bylo potřeba po nalezení stránky Windows 8.1 vyžádat vytvoření vazeb znovu s oběma variantami klíče na nově nalezených stránkách. Bez zásahu uživatele by byla úplnost dat pouze 5,3%, jelikož se druhá obdoba klíče vyskytuje hned v infoboxu nalezené stránky z té výchozí. Nicméně k dosažení 100% úplnosti, byl vyžadován pouze jeden již zásah uživatele, vyžadující tři úkony – link podle *„preceded_by“*, pak podle *„preceded by“* a opět *„preceded_by“*.

Přesnost nalezených dat byla během toho scénáře 95%. Ze stránky Windows 7 byl navíc získán odkaz, který je uveden v referencích u požadované hodnoty a odkazuje na stránku obsahující infobox. Pro získání požadovaného stromu byl tedy nutný druhý zásah uživatele spočívající ve vymazání této chybně získané hodnoty.

Pro potřeby exportu a následné vizualizace toho stromu, byla ručně vyplněna data vydání jako data narození a data ukončení rozšířené podpory jako data smrti (u Windows 10 bylo uvedeno datum ukončení rozšířené podpory verze 2019 LTSC). Výsledná vizualizace je na následujícím obrázku 6.3.



Obrázek 6.3 Historie verzí MS Windows

7 Závěr

V rámci této bakalářské práce byl vytvořen program, který uživateli umožňuje získávat data z částečně strukturovaných struktur na Wikipedii a na jejich základě mu usnadňuje z nich budovat grafové struktury. Ty mu pak umožňuje uložit ve formátu pro vizualizaci na časové ose.

Program pracuje s uspokojivou přesností, nicméně je tu prostor pro jeho vylepšení na odolnost vůči nedodržení Wikipedií předepsaných struktur, přidání možností konfigurace pro klíče podobného významu příbuzných šablon či hlubší analýzy nestrukturované části dat. Jak naznačuje poslední test, bylo by možné rozšířit program tak, aby umožňoval pracovat i s jinými typy infoboxů než osoba s podporou exportu do časové osy Timeline.

Seznam symbolů a zkratek

- **XML** - extensible markup language; univerzální značkovací jazyk
- **SGML** - standard generalized markup language; značkovací jazyk, ze kterého vychází XML
- **JSON** - JavaScript Object Notation; pro člověka jednoduše čitelný i zapisovatelný textový datový formát pro výměnu dat
- **CSV** - comma-separated values; textový datový formát oddělující jednotlivé členy čárkami či středník
- **XML** - Extensible Markup Language; obecný značkovací jazyk
- **HTML** - Hypertext Markup Language; značkovací jazyk pro tvorbu webových stránek
- **RDF** - Resource Description Framework; ontologický jazyk
- **SPARQL** - SPARQL Protocol and RDF Query Language; dotazovací jazyk pro data formátu RDF
- **WYSIWYG** - What you see is what you get; způsob editace, kdy uživatel při editací vidí do vysoké míry podobu, jakou bude mít dokument při jeho prohlížení
- **FXML** - jazyk založený na XML pro JavuFX, umožňuje deklarativní tvorbu GUI
- **GUI** - graphical user interface; grafické uživatelské prostředí

Seznam literatury a informačních zdrojů

1. Infobox. [Online] 12. 06 2020. [Citace: 17. 07 2020.] https://en.wikipedia.org/wiki/Infobox.

2. **Kacerovský, Michal.** Vizuální reprezentace precedenčního grafu. *Diplomová práce.* Plzeň : Západočeská univerzita v Plzni, Fakulta aplikovaných věd, 2015.

3. Models of Database Architecture: Hierarchical, Network and Relational Models. [Online] [Citace: 18. 07 2020.] https://www.yourarticlelibrary.com/database/models-of-database-architecture-hierarchical-network-and-relational-models/10389.

4. **Ing. Václav Vais, Ph.D.** Teoretická informatika. *3. přednáška KIV/TI.* Plzeň : Ing. Václav Vais, Ph.D., 2018.

5. **Kurzová, Helena.** Reviewed Work(s): Syntaktické struktury by Noam Chomsky. Praha : Centre for Classical Studies at the Institute of Philosophy of the Czech, 1968.

6. **Kočí, Michal.** https://www.interval.cz/clanky/co-je-xml/. [Online] 21. 02 2000. [Citace: 15. 07 2020.]

7. **Oracle and/or its affiliates.** https://docs.oracle.com/javase/tutorial/jaxb/index.html. [Online] [Citace: 30. 05 2021.]

8. https://www.json.org/json-en.html. [Online] [Citace: 15. 07 2020.]

9. **Kosek, Jiří.** http://htmlguru.cz/. [Online] [Citace: 07. 06 2021.]

10. Wikipedia. [Online] 14. 07 2020. [Citace: 16. 07 2020.] https://en.wikipedia.org/wiki/Wikipedia.

11. Wikipedia:Vandalism. [Online] 8. 07 2020. [Citace: 17. 07 2020.] https://en.wikipedia.org/wiki/Wikipedia:Vandalism.

12. Markup spec. [Online] 04. 06 2020. [Citace: 19. 07 2020.] https://www.mediawiki.org/wiki/Wikitext.

13. Wikipedia:List of infoboxes. [Online] 06. 04 2020. [Citace: 17. 07 2020.] https://en.wikipedia.org/wiki/Wikipedia:List_of_infoboxes.

14. Template:Infobox person. [Online] 28. 03 2020. [Citace: 20. 07 2020.] https://en.wikipedia.org/wiki/Template:Infobox_person.

15. Wikipedia:Database download. [Online] 14. 07 2020. [Citace: 17. 07 2020.] https://en.wikipedia.org/wiki/Wikipedia:Database_download.

16. https://wiki.dbpedia.org/about. [Online] [Citace: 15. 07 2020.]

17. https://corporate.britannica.com. [Online] [Citace: 17. 07 2020.]

18. https://www.encyclopedia.com/about. [Online] [Citace: 15. 07 2020.]

19. https://gluonhq.com/products/scene-builder/. [Online] [Citace: 18. 07 2020.]

20. https://jgrapht.org/. [Online] [Citace: 10. 06 2021.]

21. **Hessová, Gabriela.** Automatické získávání historických údajů z webových zdrojů. *Bakalářská práce.* Plzeň : Západočeská univerzita v Plzni, Fakulta aplikovaných věd, 2015.

A. Uživatelská dokumentace

Softwarové požadavky

Pro používání obou programů, je nutné mít nainstalované Java SE Runtime Environment 8[5] a databázový systém MySQL. Pro zobrazení exportovaných souborů pro časovou osu Timeline je nutný webový server. Všechny požadované služby mohou být jednoduše splněny nainstalováním WampServeru, který je dostupný online ke stažení[6]. Před instalací WampServeru je nutné mí nainstalovaný Microsoft Visual C++ Redistributable pro Visual Studio 2015, 2017 a 2019[7] a Visual C++ Redistributable for Visual Studio 2012 Update 4[8].

Příprava dat

Pro vytvoření databáze s defaultním jménem *wikipedia* a v ní vytvoření potřebných tabulek lze použít přiložený script *wikipedia-mysql-create.sql*. Skript *wikipedia-mysql-create-and-fill-test-data.sql* jí navíc naplní testovacími daty, které byly použity v rámci uvedených testů úplnosti a přesnosti.

Pro naplnění databáze daty z Wikipedia XML dumpu[9] slouží aplikace *DataMining-Wikipedia-DatabaseInserter*. Při jejím spuštění musí být jako parametry uvedeny v tomto pořadí: IP databáze, jméno databáze, uživatelské jméno a uživatelské heslo pro přístup k databázi, cesta k rozbalenému XML dumpu. Tato konfigurace také může být předána programu formou souboru *config.txt*, ve stejném pořadí, každý parametr na jednom řádku (při použití prázdného hesla je nutné uvést prázdný řádek). Konfigurační soubor musí být ve stejné složce, ze které se spustitelný JAR spouští. Zpracování XML dumpu může trvat v závislosti na hardware a velikosti konkrétního dumpu až desítky hodin.

[5] https://www.java.com/en/download/manual.jsp
[6] https://sourceforge.net/projects/wampserver/
[7] https://support.microsoft.com/cs-cz/topic/posledn%C3%AD-podporovan%C3%A1-verze-aplikace-visual-c-ke-sta%C5%BEen%C3%AD-2647da03-1eea-4433-9aff-95f26a218cc0
[8] https://www.microsoft.com/en-us/download/details.aspx?id=30679
[9] https://dumps.wikimedia.org/

Práce s daty

Po naplnění databáze, je možné přistoupit k sestavování vlastních grafů. K tomu slouží aplikace *DataMining-Wikipedia-LinkingApp*. Pro konfiguraci databázového připojení využívá podobný konfigurační soubor jako *DataMining-Wikipedia-DatabaseInserter*. Jediným rozdílem je, že nepožaduje poslední řádek, který v předchozím případě byl XML dump soubor. GUI je popsáno v podkapitole 5.2.1.