Projekt Zpřístupnění báze národních jmenných autorit v podobě propojených dat (Linked Data)

Závěrečná zpráva

Zpracovali *Vojtěch Dostál* (Wikimedia ČR), *Jiří Sedláček* (Wikimedia ČR), *Zdeněk Bartl* (Národní knihovna ČR) a *Linda Jansová* (Národní knihovna ČR)

Praha, prosinec 2019 – leden 2020

OBSAH

1. Úvod

Pilotní projekt *Zpřístupnění báze národních jmenných autorit v podobě propojených dat (Linked Data)* probíhal ve spolupráci Oddělení národních jmenných autorit Národní knihovny ČR, jež má ve své správě bázi národních jmenných autorit, a spolku Wikimedia Česká republika, který se zabývá propagací a podporou svobodné tvorby na území České republiky. Spolupráce navazuje na dlouhodobý a úspěšný projekt propojení Wikipedie s databází národních autorit, který probíhá již téměř deset let. Realizaci projektu finančně podpořilo Ministerstvo kultury ČR.

Cílem projektu *Zpřístupnění báze národních jmenných autorit v podobě propojených dat (Linked Data)* bylo:

1. vytvořit metodiku a podpůrný softwarový nástroj pro zpřístupňování autoritních záznamů, vzniklých v rámci kooperace českých knihoven, v podobě propojených dat;
2. zpřístupnit autoritní záznamy v podobě propojených dat;
3. přenést vybrané údaje z položek Wikidat (především další identifikátory typu ORCID, ISNI apod.) do záznamů v bázi národních jmenných autorit.

Poznámka: V důsledku nižší než předpokládané úrovně financování (bezmála o třetinu) bylo rozhodnuto o snížení ambicióznosti tohoto projektu. Bod (1) se omezil na vytvoření metodiky pro zpřístupňování dosavadních autoritních záznamů v podobě autoritních dat a dílčí softwarové pomůcky a bod (2) se soustředil zejména na co nejkvalitnější spárování Wikidat s bází národních jmenných autorit tak, aby báze autorit mohla být obohacena o URI příslušných položek z Wikidat.

Díky projektu se Národní knihovna ČR začlenila po bok významných zahraničních národních knihoven, které zpřístupňují alespoň část dat, jež (obvykle ve spolupráci s dalšími knihovnami z příslušných zemí) vytvářejí, v podobě propojených (a zároveň otevřených) dat. Spolupráce s projekty Wikimedia (především s Wikidaty), na které byl projekt založen, je na mezinárodní knihovnické scéně rovněž velmi aktuálním trendem. Názorně to dokládá mj. zpráva Liama Wyatta *Wikidata & Wikibase for national libraries: the inaugural meeting* zveřejněná v zářijovém čísle periodika The Signpost[1].

1 Viz https://en.wikipedia.org/wiki/Wikipedia:Wikipedia_Signpost/2019-09-30/In_focus.

2. Slovníček použitých termínů

- **autoritní záznam (autorita)** – ucelený soubor údajů o osobě, organizaci, tématu nebo například dílu;
- **báze národních autorit** – databáze Národní knihovny ČR, která obsahuje jmenné i věcné (tematické, geografické a formální) autority;
- **báze národních jmenných autorit** – dílčí část báze národních autorit, která obsahuje jmenné autority;
- **Wikidata** – sesterský projekt Wikipedie; databáze, která usiluje o zpracování veškerých poznatků lidstva formou propojených dat;
- **Wikibase –** software, na kterém fungují Wikidata a další podobné databáze; rozšiřuje základní software MediaWiki a doplňuje ho o nové funkce.

3. Výsledky projektu

3.1 Zpřístupnění autoritních záznamů formou propojených dat

Cílem této části projektu bylo zpracovat data v bázi národních jmenných autorit jako propojená data (linked data) a publikovat je pod svobodnou licencí (CC-0) v databázi, jež poběží na softwaru Wikibase. Tento software je vhodný pro zpracovávání datových sad jako propojených dat a velmi snadno může komunikovat s dalšími projekty ve světě propojených dat, jako jsou Wikidata.

Dedikovaná instalace Wikibase s jmennými autoritními daty se nachází na adrese https://autority.wikimedia.cz/. Obsahuje 773 571 položek, jednu pro každý záznam v bázi národních jmenných autorit. Každá položka obsahuje několik tvrzení odrážejících zpracovaný obsah z báze národních jmenných autorit. Obvykle se jedná o:

- Jméno v NKČR AUT;
- Datum narození;
- Datum úmrtí;
- Křestní jméno;
- Příjmení;
- Typ záznamu.

Zajímavostí je, že data narození a úmrtí byla extrahována také z pole 678, které obsahuje různé biografické informace formou jedné věty či krátkého odstavce. Nejdůležitější však jsou odkazy na identifikátory – tvrzení:

- NKČR AUT (identifikátor v bázi národních autorit);
- QID (identifikátor příslušné položky ve Wikidatech).

Díky těmto identifikátorům jsou data propojena se svými sesterskými databázemi.

3.2 Propojení záznamů v bázi národních jmenných autorit s položkami ve Wikidatech

Pro propojení obou databází je naprosto zásadní proces tzv. matchingu, při němž se identifikují položky z Wikidat pro příslušné záznamy v bázi národních jmenných autorit.

Před začátkem projektu bylo propojeno asi 231 000 záznamů z báze národních autorit, z toho většinu tvořily jmenné autority. Po dokončení projektu číslo vystoupalo na 268 200. Díky tomu má více záznamů v bázi národních jmenných autorit možnost načítat unikátní URI položek z Wikidat a poskytnout tak uživateli databáze další související informace.

Co nejkvalitnější propojení databází je zásadní také proto, aby se při přenášení dat mezi databázemi omezil vznik duplicit v aktualizované databázi, například ve Wikidatech. Počet napojených položek bude dále stoupat v čase s tím, jak se obě databáze rozšiřují. Většinu stávajících záznamů jmenných autorit s již existujícím záznamem Wikidatech jsme však propojili.

3.3 Obohacení báze národních jmenných autorit pomocí dat z Wikidat

Cílem této části projektu bylo obohatit data v bázi národních jmenných autorit o vybrané údaje z databáze Wikidata. Ačkoliv v posledních letech probíhají v Kongresové knihovně ve Washingtonu a v knihovnách, které s ní spolupracují, intenzivní práce na přizpůsobení formátu MARC 21 změnám souvisejícím s propojenými otevřenými daty, nejsou tyto diskuse dosud uspokojivě ukončeny. Na řadu podstatných otázek, které se objevují v praxi (například při snaze vhodně uvádět identifikátory včetně URI v poli 024 formátu MARC 21 pro autority), nebylo ještě ke konci roku 2019 v oficiálních dokumentech ani při konzultacích s odborníky v této oblasti možné najít jednoznačné odpovědi. Z toho důvodu byly zatím podniknuty pouze přípravné kroky směřující k obohacení autoritních záznamů o některé údaje z Wikidat (například identifikátory ORCID, ISNI apod.); vlastní import bude realizován až v okamžiku, kdy budou odpovědi na sporné otázky vyjasněny. Předpokládá se, že k potřebnému vyjasnění by mohlo dojít již v prvních měsících roku 2020, kdy by měly být k dispozici přesnější pokyny vztahující se k poli 024[2]. Je nutno podotknout, že ačkoliv není Národní knihovna ČR přímo zapojena do amerického Programu kooperativní katalogizace (Program for Cooperative Cataloging, PCC), v jehož rámci tyto pokyny vznikají, je velmi žádoucí upravovat české autoritní záznamy v souladu s nejlepší praxí zahraničních knihoven. To usnadní výměnu a sdílení autoritních záznamů (dat) na mezinárodní úrovni.

Díky práci na plnění prvních dvou cílů projektu jsou však data k obohacení záznamů již připravena a jsou dostupná pod svobodnou licencí. Národní knihovna ČR je může do své

2 Viz blíže https://www.loc.gov/aba/pcc/naco/documents/024-moratorium.pdf.

databáze kdykoliv zahrnout; z praktických důvodů to bude optimální až poté, co se otázky týkající se obohacování autoritních záznamů podaří uspokojivě vyřešit na mezinárodní úrovni.

Příklady těchto dat jsou součástí kapitoly 4 této závěrečné zprávy. Je k dispozici například více než 3 000 spárovaných identifikátorů ve vědecké databázi ORCID (před začátkem tohoto projektu jich bylo pouze cca tisíc).

Byl také zaveden systém pro hlášení pravděpodobných chyb v bázi národních autorit. V okamžiku psaní této zprávy je ve Wikidatech hlášeno 56 chyb, které budou postupně Národní knihovnou ČR prověřovány. Podrobně je problematika popsána v kapitole 4 této zprávy.

4. Metodická část

4.1 Zpřístupnění autoritních záznamů formou propojených dat

K vytvoření databáze propojených dat byly použity dva zdroje dat. Prvním byl **export dat z báze národních autorit** (soubor aut_exp.xml.gz) na FTP serveru Národní knihovny ČR.[3] Exportovaná data jsme v nástroji Catmandu[4] převedli z formátu MARCXML do formátu JSON. Druhým zdrojem dat byl **výsledek databázového dotazu ze serveru Wikidata**, který obsahoval identifikátory v bázi národních autorit (NKČR AUT) spárované s identifikátorem ve Wikidatech (QID). Spárováním dat přes identifikátor NKČR AUT jsme získali základ databáze pro import. Následně jsme z exportu báze národních autorit extrahovali následující údaje:

- Jméno v NKČR AUT, Křestní jméno, Příjmení – pole 100 $a ve formátu MARC 21;
- Datum narození / Datum úmrtí – pole 100 $d ve formátu MARC 21;
- NKČR AUT – pole 100 $7 ve formátu MARC 21;
- Poznámka – pole 678 ve formátu MARC 21.

Kromě toho jsme vytěžením obsahu pole 678 (poznámka) pomocí regulárních výrazů získali i přesnější datum narození a datum úmrtí, pokud bylo v poznámce k dispozici. Metodika této extrakce sestávala z několika kroků. Nejprve jsme standardizovali tvar uvozující data narození v bázi národních jmenných autorit, jelikož kromě běžných výrazů „narozen“ a „narozena“ se vyskytují i různé další varianty[5]. Následně byl sestaven regulární výraz pro

3 Táž data (úplné autoritní záznamy) jsou k dispozici také v rámci mezinárodního systému VIAF (Virtual International Authority File, Virtuální mezinárodní autoritní soubor, http://viaf.org/, resp. http://viaf.org/viaf/data/). Do něj se Národní knihovna ČR aktivně zapojila již v roce 2009, konkrétně jako druhá knihovna z knihoven, které se přidaly ke čtyřem zakládajícím členům systému (těmi se staly Kongresová knihovna, Německá národní knihovna, Francouzská národní knihovna a OCLC). Data v systému VIAF jsou zpřístupněna pod otevřenou licencí Open Data Commons Attribution License (ODC-By) v1.0 (https://opendatacommons.org/licenses/by/1.0/). Ta komukoliv umožňuje data sdílet (kopírovat, šířit a využívat), používat k vytváření dalších děl a upravovat je, přičemž je nutné dodržet pouze jedinou podmínku, kterou je uvedení zdroje. V projektu bylo tedy technicky možné pracovat i s daty ze systému VIAF. Z praktických důvodů (vzhledem k tomu, že projekt realizovala přímo Národní knihovna ČR a že předmětem zpracování byly výhradně autoritní záznamy z této knihovny) však byla pro vlastní zpracování dat využita data přímo z FTP serveru Národní knihovny ČR.

4 Viz https://librecat.org/Catmandu/.

5 Například pro data narození byly identifikovány tyto tvary: Narozenaaaa | Narodila se | Narodizen | Naroezena | Naroizena | Narozebna | Narozen/a | Narozena: | Narozenav | Narození: | Narodil se | Narodil se | Narizena | Narodila | Narodíla | Narodilů | Naroezen | Naroyena | Narozana | Narozemn | Narozen, | Narozen. | Narozen: | Narozena | Narozená | Narozene | Narození | Narozenl | Narozenn | Narozeno | Narozený | Narozten | Narozuen | Narozzen | Narpozen | Nasrozen | Nazozena | Nar. v r. | Naorzen | Naozena | Narodil | Naroezn | Narozeb | Narozem | Narozen | Narozev | Narozna | Narpzen | Narzena | Nazoren |

datové údaje nacházející se v textovém řetězci vždy bezprostředně za výše uvedenými výrazy[6].

Tato data jsme následně převedli do podoby trojic (tripletů) RDF (Resource Description Framework, Rámec pro popis zdrojů) a importovali do vlastní instance Wikibase na serveru spolku Wikimedia Česká republika, z.s. (dostupné na adrese http://autority.wikimedia.cz/). Veškerá autoritní data jsou zde dostupná pod licencí CC-0[7], tedy pod licencí, kterou používá i databáze Wikidata. Importováno bylo celkem 773 571 záznamů, což představuje většinu jmenných autoritních záznamů dostupných k listopadu 2019 na FTP serveru Národní knihovny ČR. Vyloučeny byly záznamy nekompletní či dosud neexportované do dostupného dumpového souboru. Import trval přibližně čtyři dny při rychlosti asi 180 importovaných položek za minutu. Úplný výpis zpracovaných dat spolek Wikimedia Česká republika poskytne na vyžádání. V příloze 1 je příklad jedné z položek (Q817472 – Falk, Quirin).

Ve výše zmíněném nástroji Catmandu byly zpracovány i další podpůrné skripty, které jsou k dispozici v příloze 2 a byly využity k efektivní práci s pseudonymy (viz též kapitolu 4.2), k získání poznámek, ale také k identifikaci typu záznamu.

4.2 Propojení záznamů v bázi národních jmenných autorit s položkami ve Wikidatech

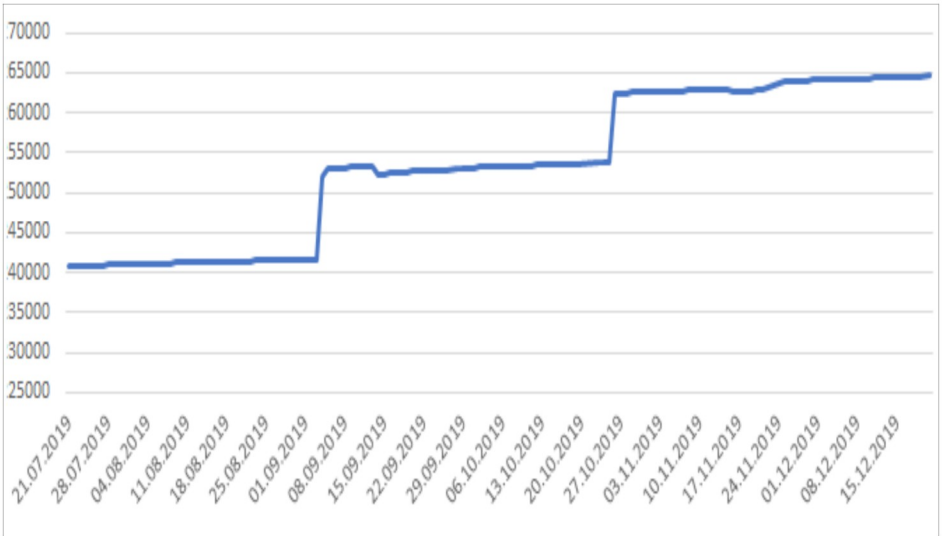
Časově nejnáročnější částí celého projektu bylo propojování záznamů v bázi národních jmenných autorit s odpovídajícími položkami ve Wikidatech. Tento proces je složitý, protože vyžaduje párování velkého množství dat stylem „každý s každým", přičemž se jedná o osoby, kde jediným spolehlivým identifikátorem bylo obvykle jméno a příjmení a dále datum narození a úmrtí. Celkem se podařilo propojit 268 200 záznamů z báze národních autorit (převodní tabulka je k dispozici na adrese https://w.wiki/EEX). Jedná se o značný vzestup oproti situaci v srpnu 2019, kdy bylo propojeno cca 240 000 položek (viz graf 1). Předpokládáme, že více než 95 % položek z Wikidat s existujícím odpovídajícím záznamem v bázi národních jmenných autorit je nyní s tímto záznamem propojeno.

Naozen | Naroen | Naroze | Narozn | Narzen | Nar. r. | Naroezn | Naroze | Naozen | Nar.

6 Například pro data narození regulární výraz: „narozen *([0-9]* *\.* *[0-9]* *\.* *[0-9]{4})".

7 Viz https://creativecommons.org/publicdomain/zero/1.0/deed.cs.

Graf 1: Vývoj počtu položek ve Wikidatech propojených na ID NKČR AUT (zdroj: Wikimedia ČR)



Ve většině případů jeden záznam v bázi národních autorit odpovídá právě jedné položce ve Wikidatech, ale existují výjimky. Pseudonymy chápe báze národních autorit jako samostatné subjekty, zatímco ve Wikidatech se zpravidla modelují v rámci položky dané osoby, která pseudonym užívá. Hlavní záznam osoby je s pseudonymy v bázi národních autorit propojen v poli 500, přičemž v upřesňovacím podpoli $w se nachází kód „v”. Tímto způsobem jsme byli schopni propojení nalézt. Někteří autoři používají více pseudonymů (výpis z Wikidat zde: https://w.wiki/EEa, rekordmanem je v tomto ohledu filozof Miroslav Petříček s 53 zaznamenanými pseudonymy). Celkový počet položek s propojením je tedy o několik tisíc nižší než celkový počet propojení, což je patrné i z grafu 1.

Pro zbylá párování (tedy pro ta párování, která nezahrnovala pseudonymy) jsme použili data narození a úmrtí. Na straně Wikidat pro to byl výchozím souborem dump Wikidat (https://www.wikidata.org/wiki/Wikidata:Database_download/cs) o velikosti více než 0,5 TB, z jehož komprimované verze byly postupným procesem extrahovány položky osob včetně dat narození a úmrtí, aby výsledný soubor obsahoval jen data, která jsme potřebovali. Data byla dále pročištěna a segmentována (například algoritmicky jsme oddělili křestní jména od příjmení tam, kde to bylo možné – jak z dat Národní knihovny ČR, tak z Wikidat) a označili jsme si ty položky, které jsou se záznamem v bázi národních autorit již propojeny (jednoduchým dotazem dostupným pomocí adresy https://w.wiki/EEe). Na straně báze národních autorit byl použit opět soubor aut_exp.xml.gz.

K dalšímu párování jsme využili skutečnosti, že v roce 2019 proběhl import identifikátorů VIAF k příslušným položkám s identifikátory autoritních záznamů různých knihoven, které jsou (podobně jako Národní knihovna ČR) zapojeny do systému VIAF. Tím se nám otevřela možnost doplnit další identifikátory NKČR AUT na základě identifikátorů VIAF. Párování proti databázi VIAF proběhlo v několika vlnách; první vlna proběhla na počátku projektu, další vlny opakovaně později. Ke konci roku 2019 proběhl další velký import z databáze VIAF, ale vyhodnotili jsme, že kvůli značnému množství chyb v databázi VIAF bude třeba nově napárované položky zkontrolovat ručně.

Nakonec jsme provedli export dat narození a úmrtí z Wikibase národních autorit do Wikidat. Tím jsme obohatili Wikidata o 90 209 dat narození (https://w.wiki/EMV) a 51 371 dat úmrtí (https://w.wiki/EMX). Všechny tyto záznamy jsou opatřeny referencí odkazující na jmenný autoritní záznam, z něhož bylo čerpáno.

Ačkoliv práce není u konce (a vzhledem k neustálému přírůstku dat na straně Národní knihovny ČR nikdy nebude), náš projekt umožnil dohnat zpoždění, které měla databáze Wikidata vůči bázi národních autorit. Na základě vnitřních pravidel Wikidat stále evidujeme různé chybové položky nebo položky, které je třeba manuálně zkontrolovat. Jejich seznam je pravidelně aktualizován na stránce dostupné na adrese https://www.wikidata.org/wiki/Wikidata:Database_reports/Constraint_violations/P691.

Dále provádíme různé doplňující monitorovací dotazy do databáze Wikidata, které umožňují zjišťovat nekonzistence v obou databázích.

4.2.1 Monitoring pravděpodobných nekonzistencí ve Wikidatech

Cílem je, aby každá položka ve Wikidatech měla nejvýše jeden identifikátor NKČR AUT. Výjimkou jsou identifikátory NKČR AUT pro pseudonymy osob nebo různé varianty názvů organizací, kde je tolerováno několik identifikátorů na jednu položku Wikidat; tyto identifikátory však musí být správně označeny pomocí vymezení „uveden jako”.

Hlavními monitorovacími dotazy sloužícími ke zjištění pravděpodobných chyb na straně Wikidat jsou:

- položky ve Wikidatech, jež nesou totožný identifikátor NKČR AUT: https://w.wiki/EMQ;

- položky ve Wikidatech, jež mají více než právě jeden identifikátor NKČR AUT s totožným vymezením typu „uveden jako”: https://w.wiki/EMS;
- položky ve Wikidatech, které se týkají osob, ale obsahují identifikátor NKČR AUT typu „aun” (tj. autor/název): https://w.wiki/DKQ.

4.2.2 Monitoring pravděpodobných nekonzistencí v bázi národních autorit

Vzhledem k pilné práci komunity Wikidat jsou téměř na denní bázi odhalovány různé drobné chyby a nekonzistence v datech báze národních autorit. Jedná se především o zamítnutí identifikátorů NKČR AUT a o zamítnutí údajů o narození a úmrtí z báze národních autorit.

Přehled zamítnutých identifikátorů NKČR AUT ve Wikidatech podává dotaz dostupný na adrese https://w.wiki/CQh. K zamítnutí dochází z různých příčin (ty jsou zpravidla ve výpisu uvedeny):

- **zrušená hodnota identifikátoru** – jde o záznamy v bázi národních autorit, jež byly v určité chvíli smazány;
- **zmatení několika konceptů** – jde o záznamy v bázi národních autorit, jež nedopatřením pojednávají o více osobách zároveň;
- **týká se jiné osoby** – jde o nesprávně přiřazené záznamy, tzn. záznamy pojednávající o jiné osobě než o té, o níž pojednává položka ve Wikidatech (to může být způsobeno chybou napojení ve Wikidatech);
- **duplicita** – jde o existenci duplicitních záznamů v bázi národních autorit (znamená to, že na dané téma existuje více než jeden záznam v autoritní bázi Národní knihovny ČR).

Zamítnutí časových údajů o narození, resp. úmrtí původem z báze národních autorit (https://w.wiki/DP7, resp. https://w.wiki/EMd) může být způsobeno:

- **chybným údajem ve zdroji** – zřejmou chybou v bázi národních autorit;
- **nižší přesností údaje ve zdroji** – báze neobsahuje dostatečně přesný údaj o době narození či úmrtí.

Právě výpis osob, pro které mají Wikidata k dispozici přesnější údaj o narození či úmrtí, představuje velmi užitečný zdroj i pro bázi autorit. Například pro data narození je přesnější údaj dostupný u více než 1 500 osob: https://w.wiki/EMg. Vedle toho existuje řada záznamů, u kterých báze národních autorit data narození/úmrtí neeviduje, ale Wikidata tyto údaje

obsahují. Například 10 857 osob, k nimž v Národní knihovně ČR existuje autoritní záznam, má ve Wikidatech uvedeno datum narození, ale v poli 100 $d v odpovídajícím autoritním záznamu tento údaj chybí.

4.3 Obohacení báze národních jmenných autorit pomocí dat z Wikidat

Výsledkem tohoto projektu je také nový potenciál k obohacení dat v bázi národních autorit.

Specifickým typem obohacení báze národních autorit je oprava chyb. Tím se zabýval závěr kap. 4.2 a dále zde již nebude toto rozebíráno.

Další možností je obohacení dat Národní knihovny ČR o různé identifikátory. Může zvolit dva způsoby jejich užívání:

1. dynamické načítání libovolných identifikátorů dostupných ve Wikidatech – pomocí dedikovaného API – například identifikátor ORCID Jiřího Zimy (identifikátor autoritního záznamu jn20000402749, QID Q51130732) lze získat dotazem https://tools.wmflabs.org/hub/P691:jn20000402749?property=P496;
2. statické uložení dat do záznamů jednotlivých osob v bázi národních autorit.

Pro druhou možnost hovoří skutečnost, že by takto vložená data mohla snáze užívat široká knihovnická komunita zvyklá pracovat s daty v klasickém formátu MARC 21. V případě zvolení této možnosti si může Národní knihovna ČR vybrat z poměrně širokého spektra dostupných identifikátorů, které se navíc neustále rozšiřuje. Co se týče identifikátorů osob v českých sekundárních databázích, v současné době existuje nejvíce spárování s databází abART (umělci), ČSFD (film) a Bibliografie dějin českých zemí (historici a historické osobnosti). Četnost různých „českých" identifikátorů v položkách s odkazem na bázi národních autorit zobrazuje dotaz dostupný prostřednictvím adresy https://w.wiki/Bdo.

Z mezinárodních databází se nabízí mnohem širší paleta možností. Národní knihovna ČR již dříve projevila neformální zájem o doplnění identifikátorů ORCID. V současné době evidujeme více než 3 000 propojení s databází ORCID, jež může báze národních autorit převzít: https://w.wiki/Dvu. Dále evidujeme kupříkladu:

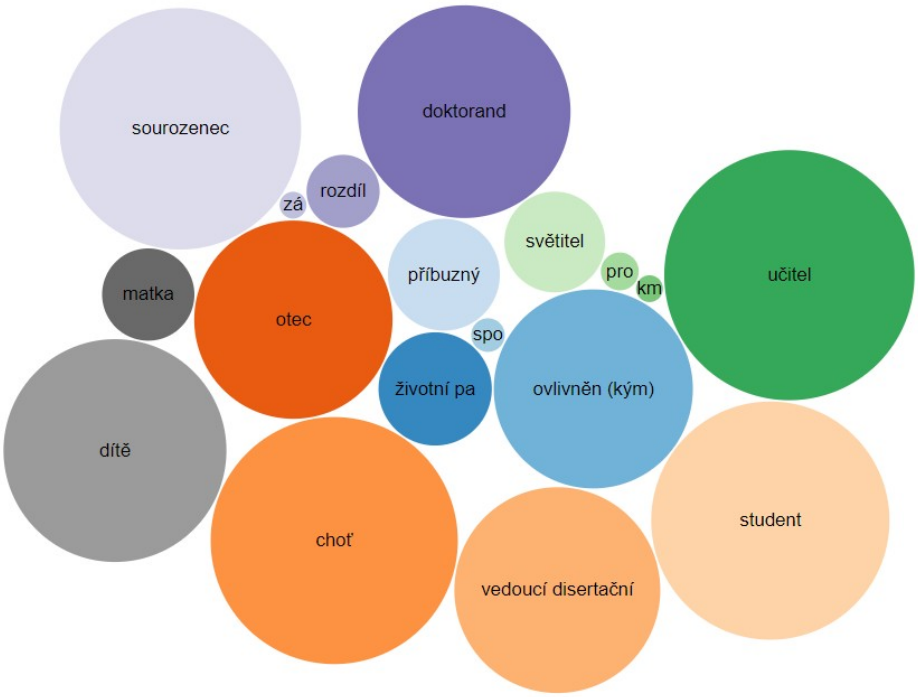
- přes 229 000 propojení na databázi ISNI: https://w.wiki/EMt;
- přes 264 000 propojení na databázi VIAF: https://w.wiki/EMu;

- přes 700 propojení na databázi ResearcherID: https://w.wiki/EMw;
- přes 2 800 propojení na databázi Scopus: https://w.wiki/EMy;
- přes 2 100 propojení na databázi poslanců Parlamentu ČR: https://w.wiki/EN2.

Další potenciál využití mají údaje o různých vztazích mezi osobami ukládané do Wikidat. Příkladem jsou vztahy typu:

- Sourozenec (https://www.wikidata.org/wiki/Property:P3373);
- Matka (https://www.wikidata.org/wiki/Property:P25);
- Otec (https://www.wikidata.org/wiki/Property:P22);
- Choť (https://www.wikidata.org/wiki/Property:P26);
- Ovlivněn (kým) (https://www.wikidata.org/wiki/Property:P737);
- Student (https://www.wikidata.org/wiki/Property:P802);
- Doktorand (https://www.wikidata.org/wiki/Property:P185);
- Světitel (https://www.wikidata.org/wiki/Property:P1598);
- Vedoucí disertační práce (https://www.wikidata.org/wiki/Property:P184).

Národní knihovna ČR či jakýkoliv další uživatel dat tato data může kdykoliv využít k moderní datové analýze či rovnou k obohacení svých dat. Nejčastější typy takto evidovaných vztahů je možné zobrazit pomocí dotazu dostupného na adrese https://w.wiki/AeN. Uvádíme je též jako bublinový diagram, viz obr. 1.



Obr. 1: Nejčastější typy evidovaných vztahů (zdroj: https://w.wiki/AeN)

5. Závěr

Projekt prokázal užitečnost spolupráce mezi Národní knihovnou ČR a spolkem Wikimedia ČR. Zatímco Národní knihovna ČR poskytla data ke zpracování, spolek Wikimedia ČR zajistil vlastní zpracování těchto dat a přípravu a zprovoznění instalace Wikibase jako místa pro uložení a zpřístupnění zpracovaných dat.

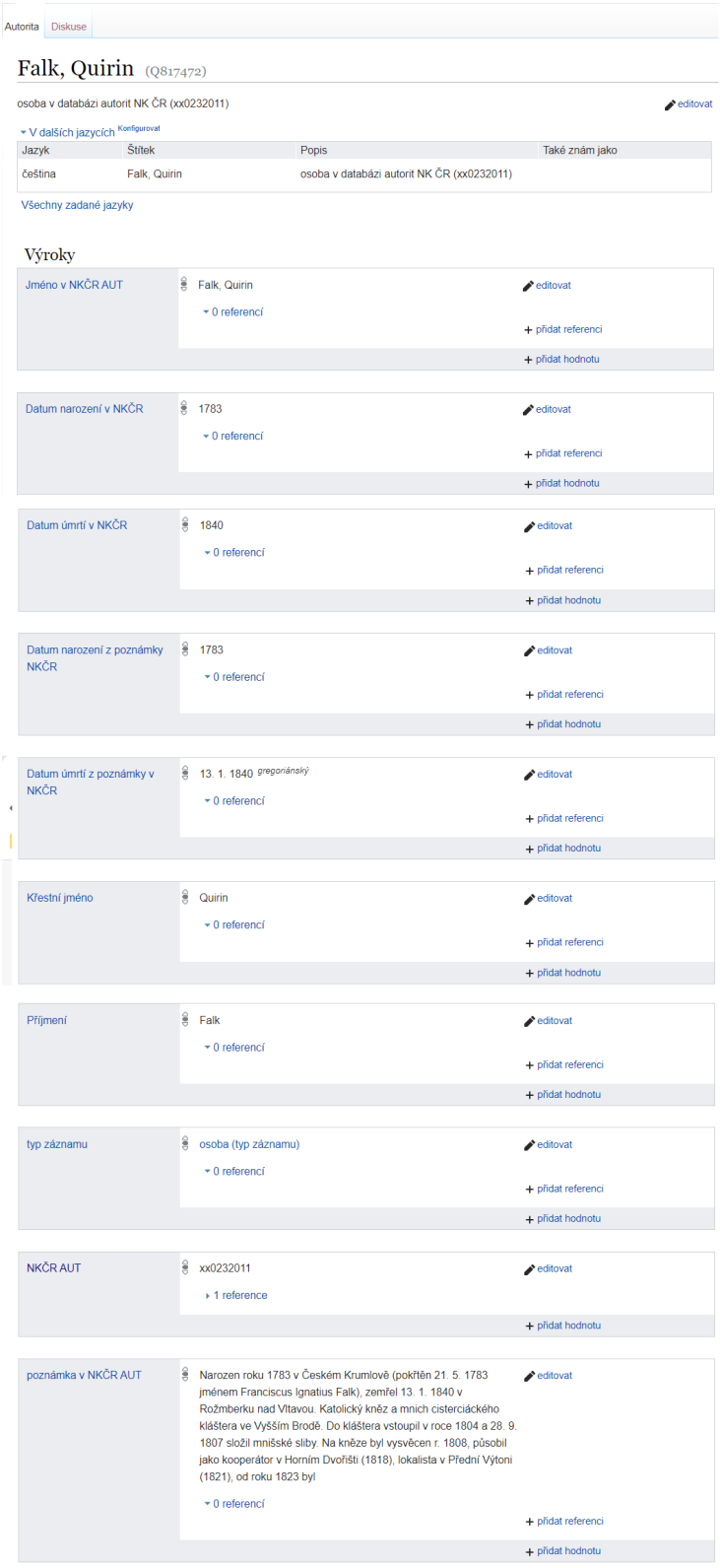
První cíl projektu, tj. vytvoření metodiky a podpůrného softwarového nástroje pro zpřístupňování autoritních záznamů v podobě propojených dat, byl splněn s tím, že se nejedná pouze o jeden konkrétní podpůrný nástroj, ale spíše o sadu dílčích nástrojů.

Druhého cíle projektu, tj. zpřístupnění autoritních záznamů v podobě propojených dat, bylo dosaženo instalací Wikibase dostupné na adrese https://autority.wikimedia.cz/ a importu údajů z autoritních záznamů Národní knihovny ČR do této instalace.

Třetí cíl projektu, tedy přenesení vybraných položek z Wikidat do záznamů v bázi národních jmenných autorit, byl realizován pouze zčásti, neboť v době realizace projektu ještě nebyla dostupná potřebná doporučení týkající se zapracování těchto údajů do autoritních záznamů ve formátu MARC 21. S ohledem na využití autoritních záznamů nejen v českém prostředí, ale také v mezinárodním kontextu dospěl tým odborníků podílejících se na projektu k závěru, že je vhodnější data zatím pouze předpřipravit a vlastní import provést až v okamžiku, kdy bude jasné, jaká řešení v tomto případě zvolí další (především americké) knihovny.

Díky realizaci projektu se Národní knihovna ČR přiblížila národním (a dalším významným) knihovnám v zahraničí, které svá data dnes běžně zpřístupňují v podobě propojených otevřených dat, a to pod licencí CC-0.

Příloha 1: Příklad záznamu (položky) ve Wikibase pro národní autority



Zdroj: https://autority.wikimedia.cz/auto/Autorita:Q817472

Příloha 2: Podpůrné skripty – Catmandu

Získání pseudonymů a jejich napojení na ostatní položky

```
if marc_match(110w,"p")

marc_map(500wa7, pseudonym, split:1, nested_arrays:1);
marc_map(510, pseudonym_org, split:1, nested_arrays:1);
marc_map(001,id);
marc_map(100a,name);
marc_map(110a,name_org);
retain(id,pseudonym,name,name_org,pseudonym_org);

else
reject()
end
```

Získání poznámek k jednotlivým záznamům

```
if marc_has(100)

marc_map(678a, pozn);
marc_map(1007,nkcr);
retain(pozn,nkcr);

else
reject()
end
```

Získání druhu záznamu (osoba, organizace, událost)

```
if marc_has(100)
   marc_map(001,id);
   add_field(type,'person');
   retain(id,type);
else
   if marc_has(110)
      marc_map(001,id);
      add_field(type,'org');
      retain(id,type);
   else
      if marc_has(111)
         marc_map(001,id);
         add_field(type,'event');
         retain(id,type);
      else
         reject();
      end
```

```
    end
end
```

Příloha 3: Prezentace výsledků projektu

Září 2019

- BARTL, Zdeněk. Soubory národních jmenných autorit a propojená data (Linked Data). In: *Konference Knihovny současnosti 2019* [online]. Praha: Sdružení knihoven ČR, 2019 [cit. 2020-01-20]. 18 snímků. Dostupné z: http://sdruk.mlp.cz/data/xinha/sdruk/Bartl_AUT%20a%20propojen%C3%A1%20data-def.pdf
- BARTL, Zdeněk. Soubory národních jmenných autorit apropojená data (linked data). In: *Knihovny současnosti 2019: sborník z 27. ročníku konference, konané 10.–12. září 2019 na Pedagogické fakultě Univerzity Palackého v Olomouci* [online]. Praha: Sdružení knihoven ČR, 2019 [cit. 2020-01-30], s. 66–70. ISBN 978-80-86249-89-6 a 978-80-7051-278-4. Dostupné z: http://sdruk.mlp.cz/data/xinha/sdruk/2019/KKS2019/Sbornik_KKS19.pdf
- DOSTÁL, Vojtěch. Databáze všeho, na níž se mohou podílet i knihovny. In: *Konference Knihovny současnosti 2019* [online]. Praha: Sdružení knihoven ČR, 2019 [cit. 2020-01-20]. 18 snímků. Dostupné z: http://sdruk.mlp.cz/data/xinha/sdruk/Dostal_knihovny-soucasnosti-2019.pdf

Listopad 2019

- BARTL, Zdeněk. Národní jmenné autority jako propojená data. In: *12. výroční seminář SK ČR* [online]. Praha: Národní knihovna ČR, 22. 11. 2019 [cit. 2020-01-30]. 18 snímků. Dostupné z: https://www.caslin.cz/caslin/dokumenty/rok-2019/bartl-autority-jako-propojena-data

Prosinec 2019

- BARTL, Zdeněk. Soubory národních jmenných autorit a propojená data aneb Šance pro autority, šance pro Wikidata, šance pro paměťové instituce! In: *Archivy, knihovny, muzea v digitálním světě 2019* [online]. Praha: Svaz knihovníků a informačních pracovníků ČR, Národní archiv, Národní knihovna ČR a Národní muzeum, 2019 [cit. 2020-01-30]. 20 snímků. Dostupné z: https://bulletin.skipcr.cz/prezentace/archivy-2019/1/Bartl.pdf
- BARTL, Zdeněk. Informace o projektu přednesená na setkání lokálních supervizorů projektu Kooperativní tvorba a využívání souborů národních autorit (Praha, Národní knihovna ČR – Klementinum, 16. prosince 2019).

Leden 2020

- DOSTÁL, Vojtěch. 773 571 osob pod licencí CC-0: databázi jmenných autorit Národní knihovny jsme propojili s Wikidaty. *Wikimedia Česká republika: blog* [online]. 2. 1. 2020 [cit. 2020-01-30]. Dostupné z: https://blog.wikimedia.cz/2020/01/773-571-osob-pod-licenci-cc-0-databaze-jmennych-autorit-narodni-knihovny-byla-integrovana-s-wikidaty/

- SEDLÁK, Jan. Wikimedia ČR a Národní knihovna uvolňují 750 tisíc záznamů o osobách. *Lupa.cz* [online]. 7. 1. 2020 [cit. 2020-01-30]. Dostupné z: https://www.lupa.cz/aktuality/wikimedia-cr-a-narodni-knihovna-uvolnuji-750-tisic-zaznamu-o-osobach/