

Soubory národních jmenných autorit a propojená data

aneb

Šance pro authority, šance pro Wikidata, šance pro paměťové instituce!



National Authority Files/ Names

ÚVOD

Propojená data (Linked Data) je název metodiky pro publikování **strukturovaných dat** způsobem, který umožní jejich **vzájemné propojování** a **opakované využití** v rámci **sémantického dotazování** (**sémantického webu**)

(Sémantický web je rozšířením současného webu, v němž **informace mají přidělen definovaný význam**, který dovoluje **automatické zpracování** počítačem. Sémantický web představuje pohled na dokumenty nejen jako na soubory textů a grafiky, ale **umožňuje znát také jejich obsah**)

Z technického hlediska se jedná o data publikovaná na webu způsobem, který je **strojově čitelný**, jejich **smysl** (význam) je jasně vyjádřen, jsou **propojena na jiná data** a je možné se na ně **odkazovat**

LinkedData

Cílem je umožnit přímou a snadnou integraci dat a vytváření odkazů mezi daty napříč celým webem.

Propojená data představují nejvhodnější způsob zveřejňování, sdílení a propojování dat (informací a poznatků) v prostředí tzv. sémantického webu. Využívají jednotné **identifikátory zdrojů** (uniform resource identifiers, **URIs**) a rámec pro popis zdrojů (angl. Resource Description Framework, **RDF**)

Často je používán termín **propojená otevřená data** (angl. linked open data). Jedná se o propojená data, která jsou zároveň daty zpřístupněnými pod otevřenou licencí, jež nebrání jejich dalšímu využívání – **nutno vyřešit v českém prostředí!!!**

V podobě propojených otevřených dat jsou zpřístupňovány zdroje pocházející z knihoven. V přehledu tzv. LOD cloud nechybí např. následující zdroje:

- předmětový heslář Kongresové knihovny ve Washingtonu
- německá národní bibliografie
- britská národní bibliografie
- virtuální mezinárodní soubor autoritních záznamů VIAF.

Příklad

DBPedia – projekt Lipské univerzity, Mannheimské univerzity, Institutu Hasso Platnera a společnosti OpenLink Software, jehož cílem je těžít z Wikipedie a Wikidat strukturovaná data a zpřístupnit je na webu jako součást sémantického webu.

Zpřístupnila obsah Wikipedie ve formátu RDF. Umožňuje tak nad daty z Wikipedie sestavovat složité dotazy a zároveň na tato data napojovat další datové zdroje dostupné online

Projekt

(rok 2019)

„Soubory národních jmenných
autorit jako propojená data“

Báze národních (jmenných i věcných) autorit



- **Nezastupitelný racionalizační nástroj zpracování sbírek/dokumentů** (jednou zpracováno, mnohonásobně využito – př. Městské veřejné knihovny)
- **Silný selekční nástroj pro národní i mezinárodní prostředí** (př. J.Čapek, VIAF)
- **Propojená data** (boří zeď mezi knihovními katalogy a webovým prostředím, použití obecných vyhledávačů typu Google pro vyhledávání sbírek)

Cíle projektu jsou:

Doplnit URI identifikátor do záznamů národních autorit (personálních) a vytvořit metodiku pro zpřístupňování autoritních záznamů, vzniklých v rámci kooperace českých knihoven, v podobě propojených dat a popsat postupy nutné pro vznik propojených dat.

Propojení existujících záznamů (personálních autorit) v databázi Národních autorit s existujícími záznamy ve Wikidatech. Cílem je spárovat obě databáze, aby ze sebe vzájemně mohly těžit. Přenést vybrané údaje z položek Wikidat (výhledově především další identifikátory typu ORCID, ISNI apod.) do záznamů v bázi národních jmenných autorit

Zpracování celé databáze Národních autorit (personálních) do podoby Wikibase licencované pod CC-0. Wikibase bude následně synchronizována s Wikidaty (import personálních záznamů, které na Wikidatech chybí, do Wikidat).

Postup prací: snaha spárovat záznamy autorit s položkami ve Wikidatech

- Spárováno 262 761 záznamů (k 20.11.2019)
 - Několik tisíc položek se podařilo spárovat přes data narození a úmrtí
 - Čeká se na import pseudonymů (jedna osoba ve Wikidatech má více záznamů v databázi národních autorit – pro každý pseudonym jeden)
 - Postupné ruční párování = časově pracné, ale bude nadále pokračovat (nikdy nekončící práce) ... Každý den se podaří ručně propojit cca 10 záznamů

Co s nespárovanými záznamy

- V průběhu prosince budou záznamy, které se s Wikidaty spárovat nepovedlo, naimportovány jako nové položky do Wikidat. Další párování se tak bude provádět formou slučování položek na Wikidatech.

Technické výzvy

- ▶ Definice záznamu v databázi autorit a ve Wikidatech není shodná. Wikidata obvykle nevytváří záznamy pro každý pseudonym ... Je tedy potřeba k některých položkám na Wikidatech přiřadit více identifikátorů a jeden označit jako „hlavní“
 - ▶ Speciální případ – skupinové pseudonymy – řešeno ve Wikidatech polomanuálně
- ▶ Při párování pomocí dat narození nalezeny osoby, které se jmenují stejně (např. John Walter) a narodily se i ve stejném roce, přesto jsou to odlišné osoby ... řešit polomanuální kontrolou.
- ▶ Některé záznamy se asi nikdy nepodaří spárovat – chybí u nich například data narození a další biografické údaje, takže nedokážeme konkrétní osobu identifikovat

Nahlašování výjimek a možných chyb

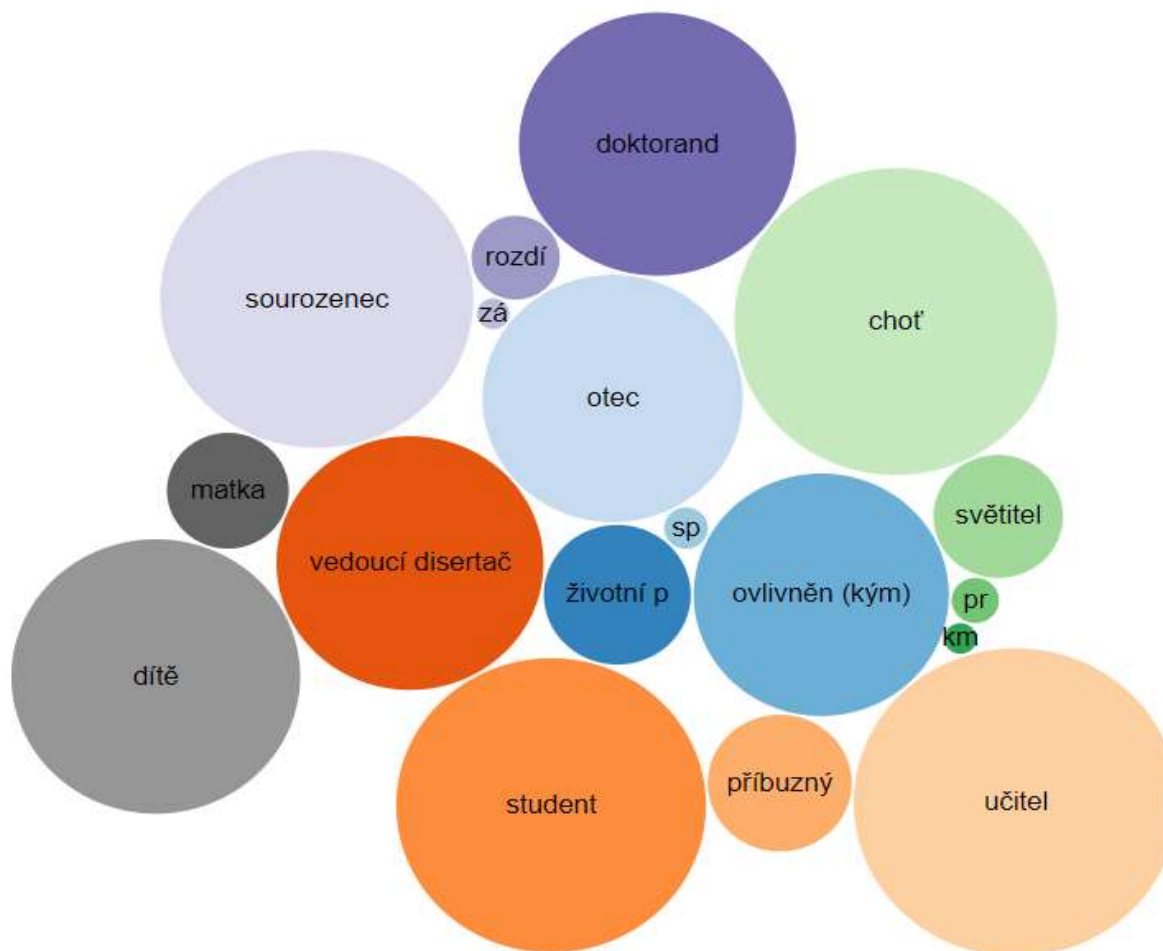
- ▶ Položky ve Wikidatech se stejným identifikátorem NK ČR: <https://w.wiki/CQg>
 - ▶ Příčina: nejčastěji buď vágní definice pojmu v databázi autorit, nebo chybné spárování ve Wikidatech
- ▶ Identifikátory NK ČR importované do Wikidat, ale z různého důvodu zavržené: <https://w.wiki/CQh>
 - ▶ Příčina: Různá (Zrušené záznamy v NK ČR, nesprávné datum narození v databázi NK ČR, duplicita v NK ČR)

Nahlašování chyb je teprve v začátcích, bude jich postupně přibývat

Zajímavé možnosti využití: Wikidata umožňují hledat vztahy mezi záznamy v databázi autorit

- Různé typy meziosobních vztahů – v diagramu

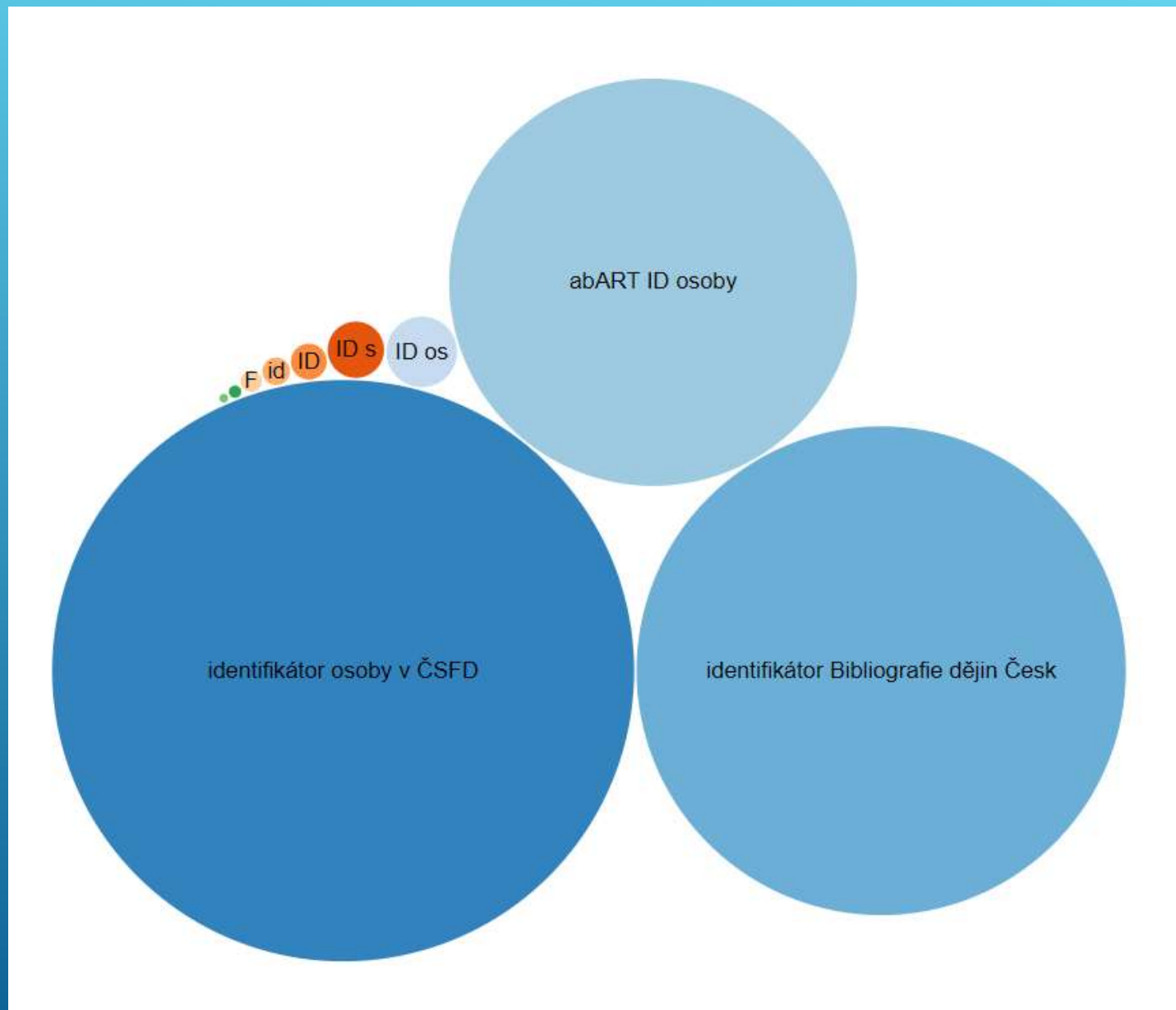
<https://w.wiki/AeN>



Wikidata umožňují prolinkování s dalšími databázemi osob

- ▶ Diagram nejčastějších českých databází, které jsou s personálními autoritami propojeny na Wikidatech

<https://w.wiki/Bdo>



**Wikidata CS**

Skupina

132 členů

Jirka Sedlacek sdílel odkaz.

12. května

Library of Congress po přidání LOC id do WD samo propojí WD položku k sobě.

To už taky pár let s NK ČR umíme, ne? (Aspoň Wikipedii)

Inspirace pro SNA - Česká terminologická databáze knihovnictví a informační vědy (TDKIV)

Termín	metadata
Termín anglicky	metadata
Výklad termínu	Strukturovaná data, která nesou informace o primárních datech. Pojem metadat je používán především v souvislosti s elektronickými zdroji a vztahuje se k datům v nejširším smyslu slova (datové soubory, textové informace, obrazové informace, hudba aj.). Funkce metadat je popisná, selekční a archivační. V souvislosti s těmito funkcemi se rozlišují metadata pro účely popisu, správy, právních nároků, technické funkčnosti, užití a archivace. Údaje se obvykle vkládají přímo do zdroje (umísťují se např. v záhlaví dokumentu HTML).
Normativní výklad	Data, která jsou začleněna do ucelené jednotky nebo se vztahují k ucelené jednotce jejím popisem a která umožňují její výběr. [ČSN ISO 8459-5]
Zdroj výkladu	SKLENÁK-2001A:331-335 GLOSS-DCMI GLOSS VOJTÁŠEK-2002
Zdroj norm.výkladu	ČSN ISO 8459-5
Další informace	<u>Wikipedie</u> / <u>Wikidata</u> <u>Wikidata</u>
Autor hesla	Celbová, Ludmila
Lektor hesla	Tkačiková, Daniela
Redaktor hesla	Schwarz, Josef Burgetová, Jarmila
Systém. číslo	000000543

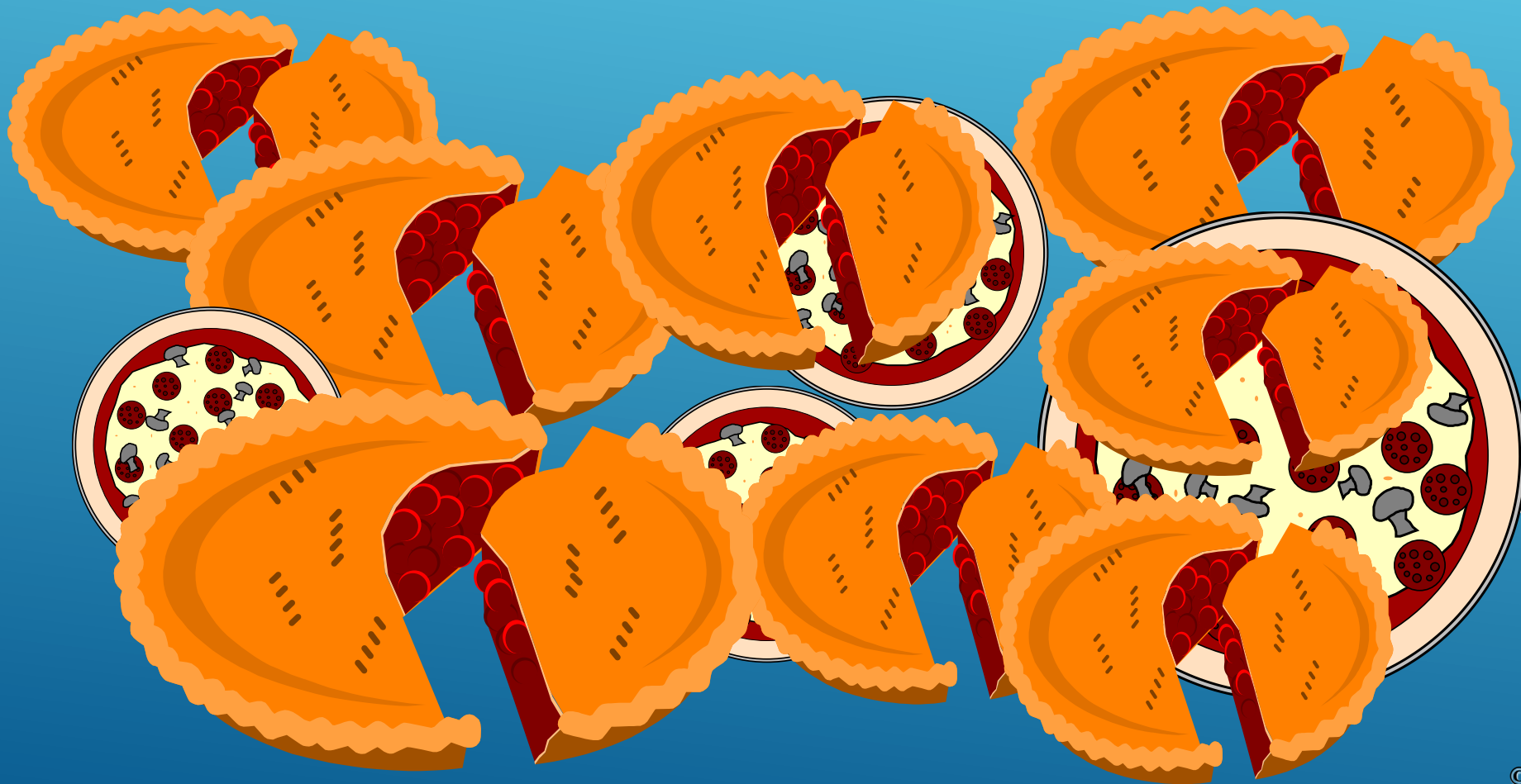
Trocha utopie ☺

OCLC společně s 16 dalšími knihovnami se v rámci pilotního Project Passage věnovala možnostem využití Wikibase (a dalších produktů Wiki) v rámci experimentálního workflow katalogizace knihovních dat do podoby Linked Data

Hlavním přínosem studie :

- Wikibase lze použít pro tvorbu strukturovaných propojených dat (Přesnost a možnosti popisu propojených dat převyšují současné knihovní standardy)
- Využití Linked Data při katalogizaci je obrovská změna katalogizačního paradigmatu
- Zajímavá byla též zmínka o tom, že přechod na Linked Data workflow bylo jednodušší, než při výměně knihovního systému

*Bez (spolu)práce nejsou
koláče/authority !*



NEJVĚTŠÍ PŘÍNOS





Neslibujeme zázraky,

ale reálné usnadnění (katalogizačních) prací

Kontakt, informace

PhDr. Zdeněk Bartl
Národní knihovna ČR
Centrální depozitář
Sodomkova 2/1146
102 00 Praha 15 - Hostivař
tel.: 21 663 322
e-mail: zdenek.bartl@nkp.cz