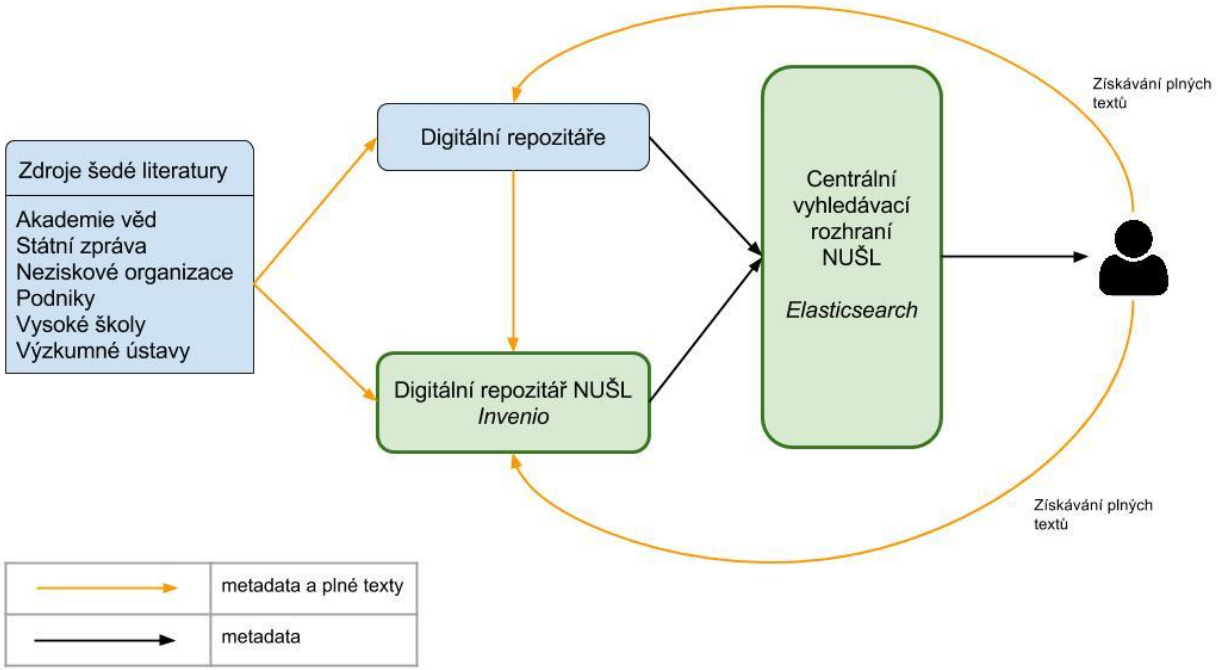
Sjednocování věcného popisu agregovaných záznamů v repozitáři NUŠL

Michaela Charvátová

IKI 2016

21. říjen 2016

NUŠL


Zdroje šedé literatury
Akademie věd
Státní zpráva
Neziskové organizace
Podniky
Vysoké školy
Výzkumné ústavy
Digitální repozitáře
Digitální repozitář NUŠL
Invenio
Centrální vyhledávací rozhraní NUŠL
Elasticsearch
Získávání plných textů
Získávání plných textů
metadata a plné texty
metadata

Důvody - obecně

- vyhledávání
- dělení podle témat
- požadavek jiných systémů

Metody sjednocování věcného popisu v NUŠL

- automatická indexace
  - na základě plných textů
  - na základě záznamů
- mapování klasifikačních schémat
  - schéma Konspektu
  - MeSH
  - klíčová slova?

Polytematický strukturovaný heslář

- produkuje a využívá NTK
- pod CC licencí

- 14 000 hesel
- 42 stromů
- každé heslo je ve struktuře pouze 1

Automatická indexace - plné texty

- 2009 - 2011
- asistent pro vkladatele
- open source Maui indexer
- strojové učení

- Full-text u méně než 2% záznamů v Inveniu
- omezení vytvořeného modelu

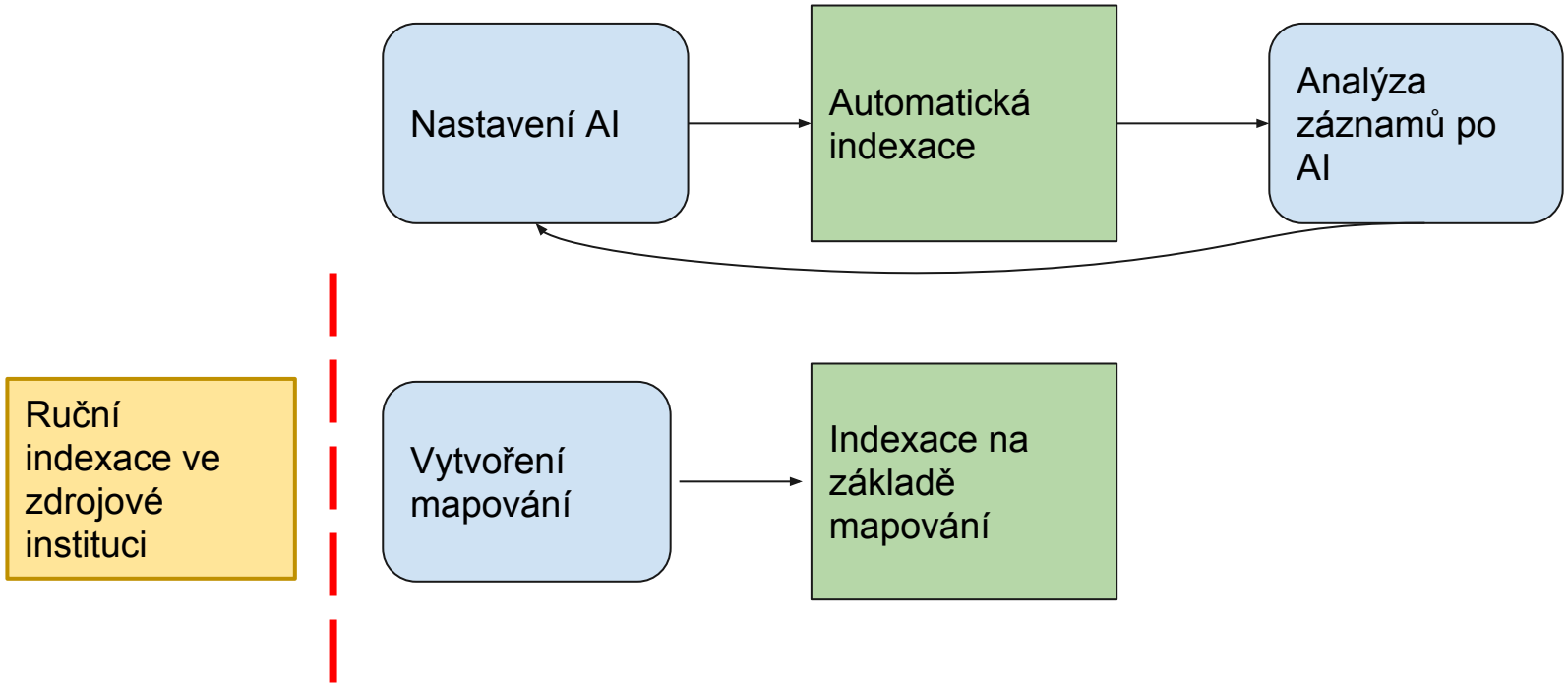
Automatická indexace - záznamy

- projekt 2012 - 2013
- používáno i v současnosti
- pole: název, abstrakt, klíčová slova, konference, instituce
- shoda: slov, skupin slov (i jejich tvarů)
- bodování míry shody a váhy parametrů

Automatická indexace - záznamy

- primárně pro předávání do OpenGrey
  - mapování PSH - klasifikace SIGLE
- přidělená hesla skryta v záznamech
- úspěšnost?
  - Česká republika, party...

Mapování?


Nastavení AI
Automatická indexace
Analýza záznamů po AI
Ruční indexace ve zdrojové instituci
Vytvoření mapování
Indexace na základě mapování

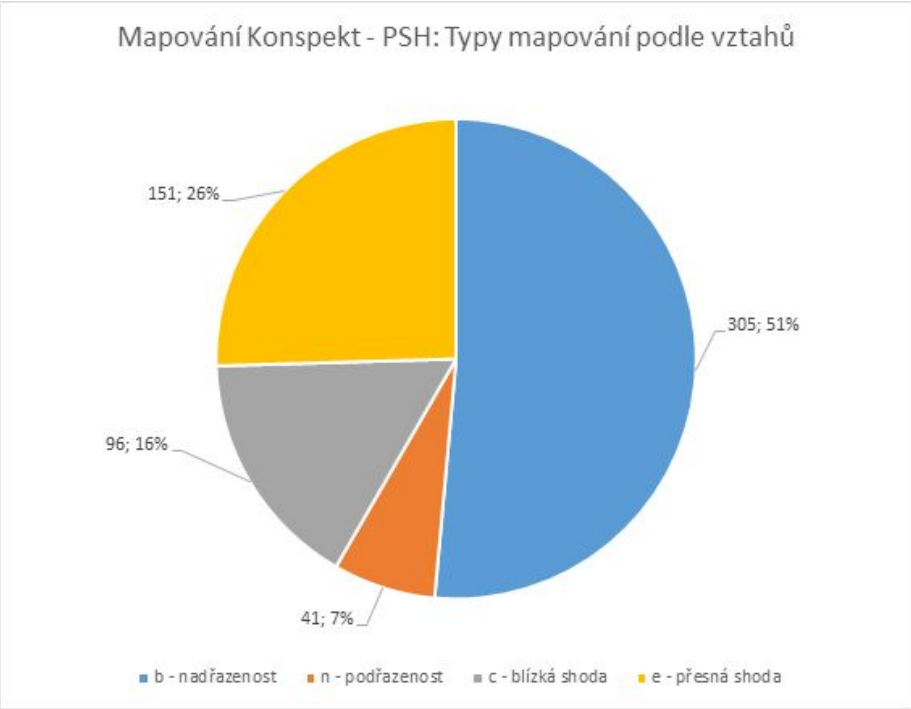
Mapování

- ISO 25964-2:2013
- obohacování záznamů v Inveniu
- cílový slovník - PSH
- mapování 1:1 a 1:2
- vyznačování vztahů (specifikace SKOS)
  - exact, close, broad nebo near match

Mapování Konspekt

- česká verze: 26 kategorií, 593 skupin
- jednoduchá tabulka
  - skupina Konspektu
  - shoda
  - ID hesla PSH (max 2)
- modul BibKnowledge v Inveniu

Mapování Konspekt



- 590 mapování 1:1
- 3 mapování 1:2

Mapování MeSH

- verze 2015: 27 455 deskriptorů
- 16 + 1 kategorie
- 1 deskriptor může být v několika tematických kategoriích
- notace
- komplikovanější zpracování, částečná automatizace

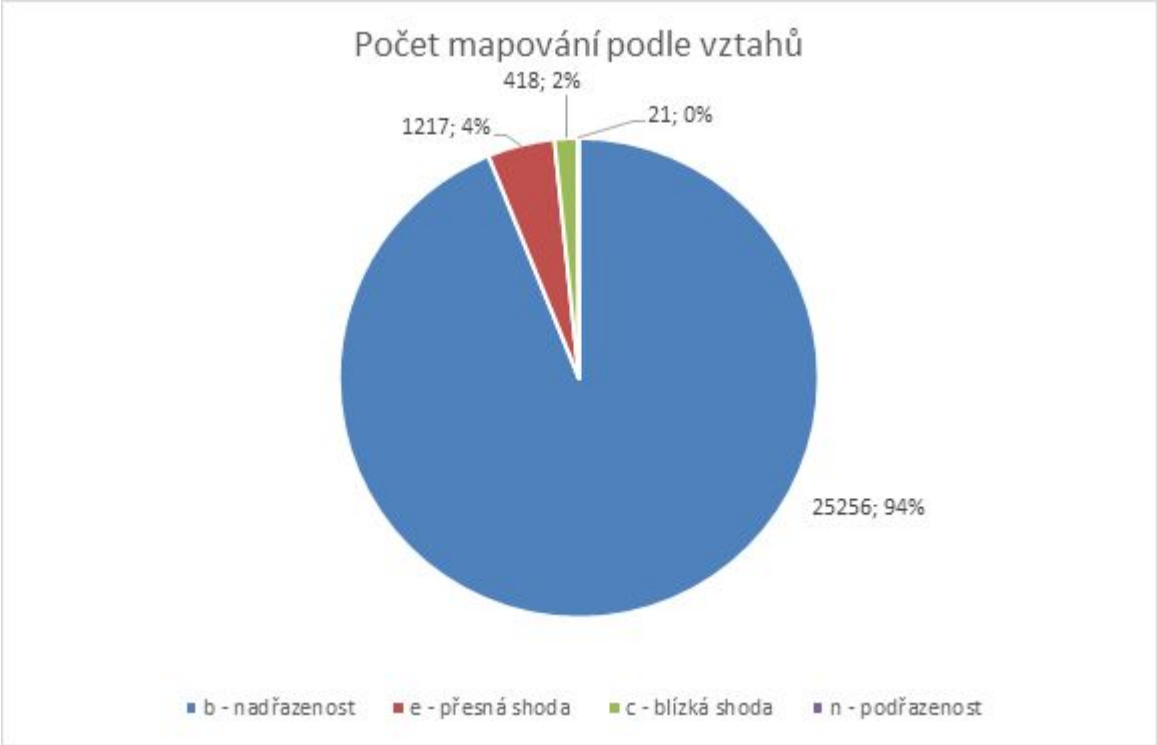
Mapování MeSH

- postup v rámci stromu kategorie
- označení posledního mapovatelného deskriptoru
- hlouběji umístěným deskriptorům přiděleno nadřazené heslo PSH s označením “broad match”
- většina deskriptorů ve více větvích MeSH ->

Mapování MeSH

- většina deskriptorů ve více větvích MeSH -> více navrhovaných hesel PSH
- priorita podle typu shody (e>c>b>n)
- v další úrovni podle hloubky umístění hesla PSH

Mapování MeSH



- 25 998 mapování 1:1
- 914 mapování 1:2

Mapování × automatická indexace

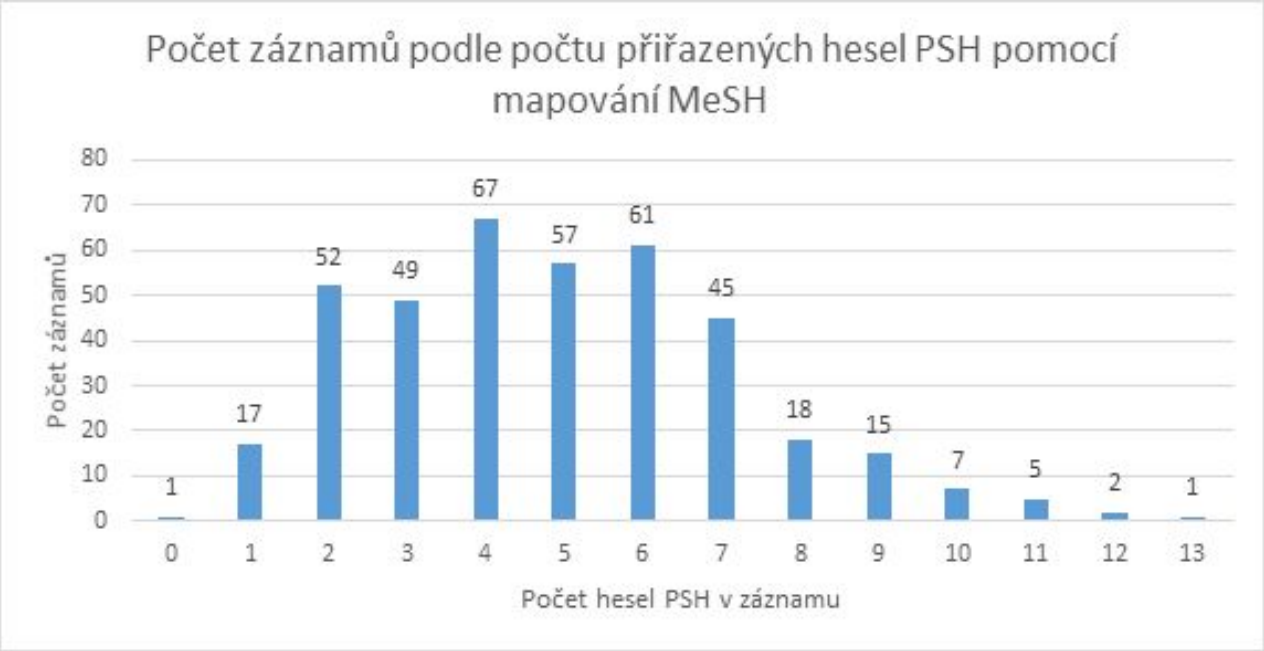
- celkem 3 802 záznamů z NLK
- vzorek 397 záznamů
  - výběr podle témat (poměr podle Konspektu)
  - uvnitř tematických skupin podle roku
- hodnocení: chybné přiřazení, chybějící hesla, počty a hloubky přiřazených hesel
  - vztaženo k ručně tvořenému popisu z NLK (MeSH, Konspekt, obor NLK)

Mapování × automatická indexace

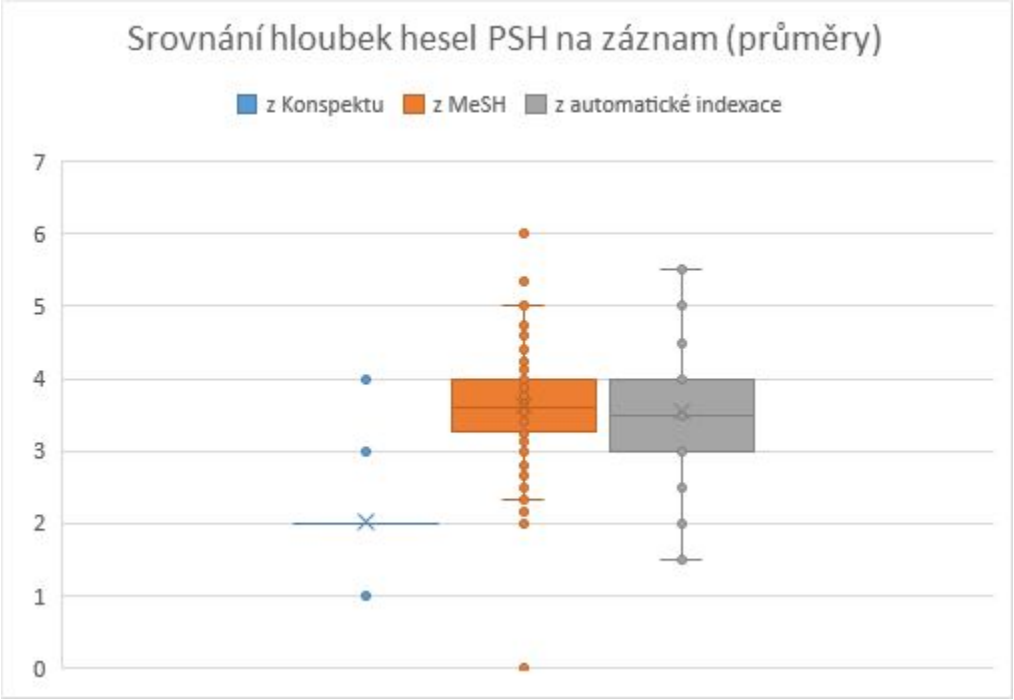
- Problematické přiřazené heslo pomocí AI:
  - 59 záznamů (14,86 % vzorku)
    - chybná větev, formální znak, obecnost

- Kvalita automatická indexace záznamů NLK
  - obor?
  - zpracování?

Mapování × automatická indexace



Mapování × automatická indexace



Mapování klíčových slov?

	ČZU	JČU	VŠE
Počet záznamů	19 731	24 747	42 349
Počet jedinečných klíčových slov	45 098	49 267	63 705
Počet jedinečných klíčových slov na záznam (počet záznamů/počet klíčových slov)	2,29	1,99	1,50
Celkový počet použití klíčových slov	130 851	72342	181 112
Počet použití na klíčové slovo (počet použití/počet klíčových slov)	2,90	1,47	2,84
Počet klíčových slov majících alespoň 10 opakování	1 936	1 350	2 598
Množství jedinečných klíčových slov majících 10 a více opakování (v procentech)	4,29%	2,74%	4,08%
Celkový počet použití klíčových slov s 10 a více opakováními	61 740	36 640	86 585
... procentuální vyjádření vůči celkovému počtu použití klíčových slov	47,18%	50,65%	47,81%
Počet záznamů s klíčovými slovy majícími 10 a více opakování	13 778	14 423	21 371
... procentní zastoupení takových záznamů v celé množině záznamů	69,83%	58,28%	50,46%
Počet klíčových slov majících 9 a méně opakování	43 162	47 917	61 107
Množství jedinečných klíčových slov majících 9 a méně opakování (v procentech)	96%	97%	96%
Celkový počet použití klíčových slov s 9 a méně opakováními	69 111	72 332	94 527
Počet klíčových slov majících 5 a méně opakování	41 632	46 578	59 179
Množství jedinečných klíčových slov majících 5 a méně opakování (v procentech)	92,31%	94,54%	92,90%
Celkový počet použití klíčových slov s 5 a méně opakováními	58 106	62 742	80736

Dotazy?