

Slovenská Webová Encyklopédia

Patrik Gajdošík

Vedúci: Michal Holub

Wikipédia

- okolo 195 000 slovenských článkov
- využitie dát: odporúčanie, personalizácia, vyhľadávanie
- neštruktúrované dáta
- dáta uvádzané v infoboxoch

Púchov	
mesto	
	
Múzeum	
Štát	 Slovensko
Kraj	Trenčiansky kraj
Okres	Púchov
Rieka	Váh
Nadmorská výška	265 m n. m.
Súradnice	 49°07′12″S 18°19′50″V
Rozloha	41,5 km² (4 150 ha) ^[1]

- projekt zameraný na transformáciu Wikipédie do štruktúrovanej podoby



- projekt zameraný na transformáciu Wikipédie do štruktúrovanej podoby
- potrebné mapovania (Aprosio, et. al, 2013)



Ciel' a hypotéza

- cieľom je vytvoriť potrebné mapovania pre slovenský jazyk automaticky
- hypotéza:

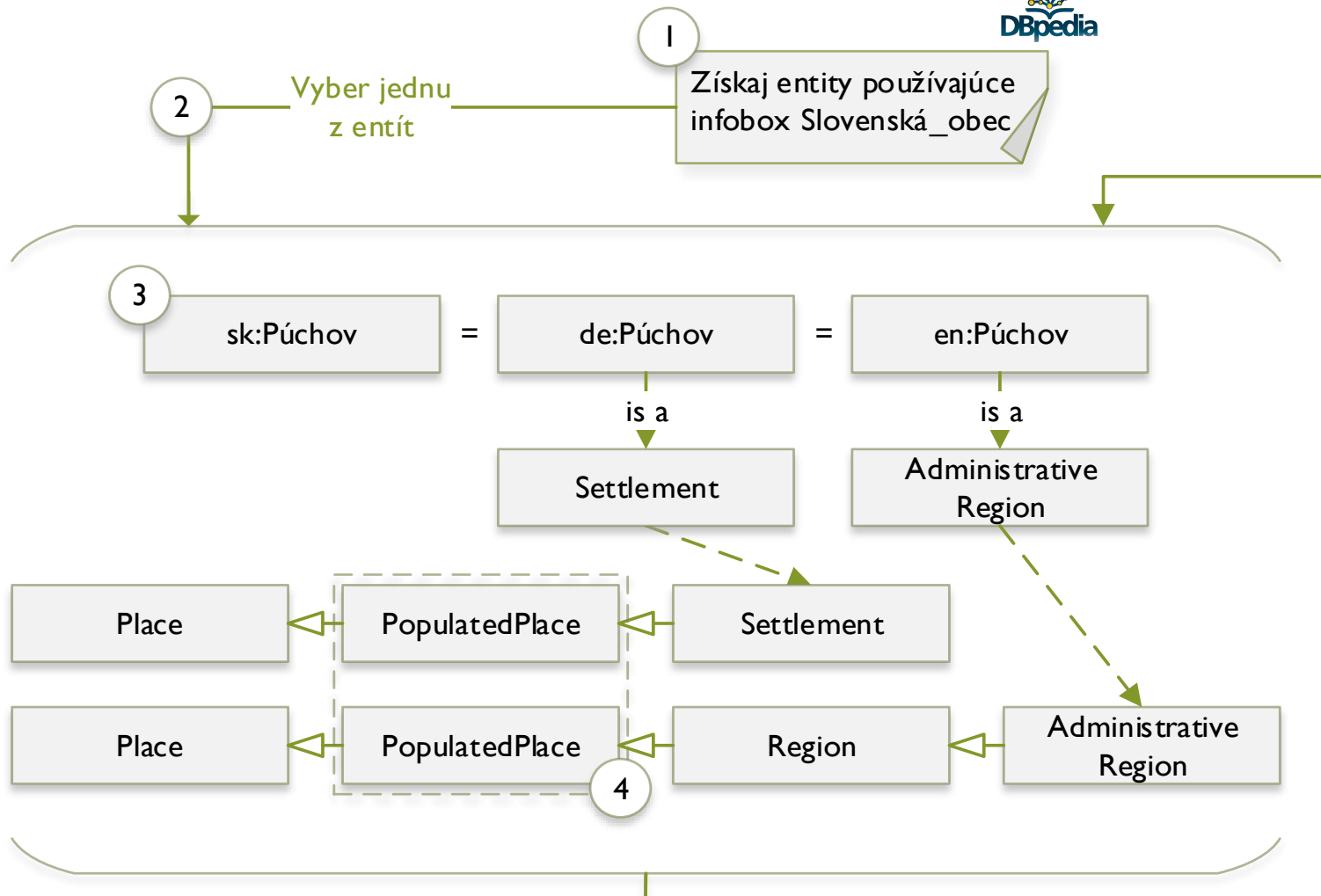
„Už existujúce mapovania v DBpédii sa dajú použiť pre automatické vytvorenie mapovaní pre jazyky, ktoré ich nemajú.“

Návrh

1. mapovanie infoboxov na triedy
 - 3 metriky návrhu mapovaní
2. mapovanie atribútov na predikáty

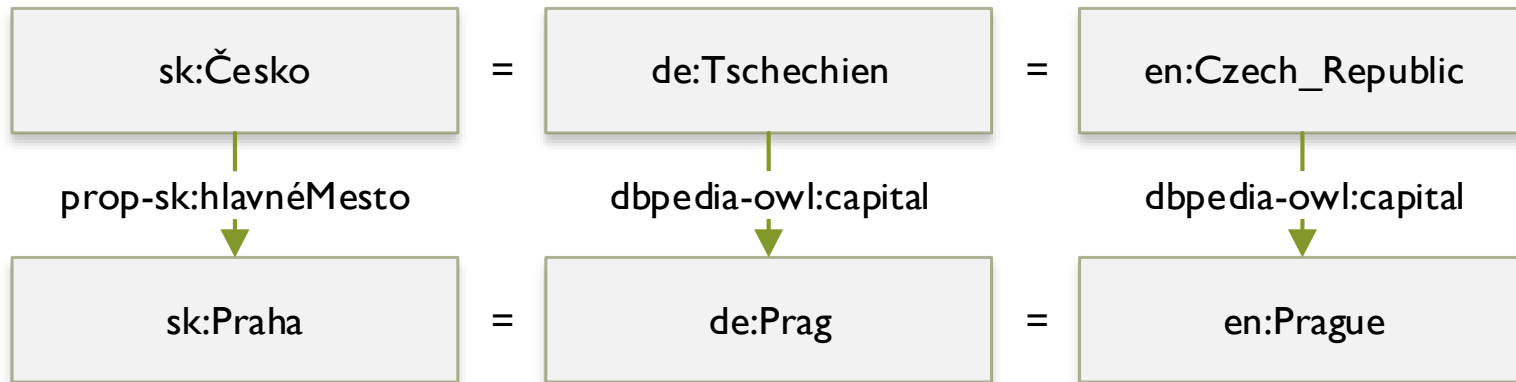
Mapovanie infoboxov na triedy

- priradenie triedy z ontológie DBpédie k infoboxu
- 3 metriky
 - základná (Aprosio, et al., 2013)
 - špecializujúca sa
 - zovšeobecňujúca





Mapovanie atribútov na predikáty



Implementácia

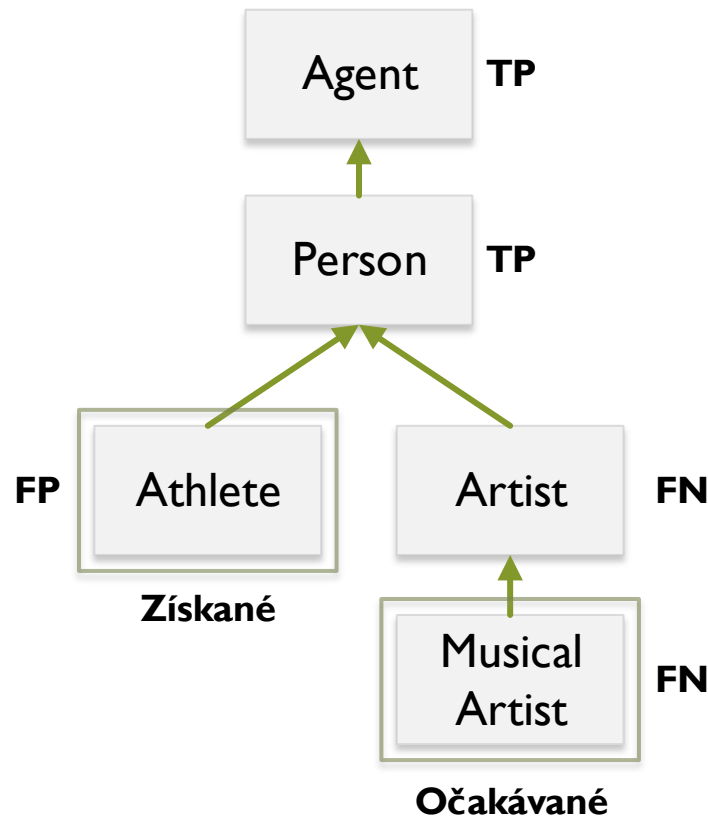
- nástroj *parmenides* poskytujúci funkcionality potrebnú pre automatické vytváranie mapovaní
 - konzolová aplikácia
- vyextrahované trojice pomocou našich mapovaní

Experimenty

- overenie mapovania infoboxov na triedy
 - presnosť a návratnosť
 - skóre správnosti
- overenie mapovania atribútov na predikáty

Overenie mapovania infoboxov na triedy

- metóda presnosti a návratnosti (Aprosio, et al., 2013)



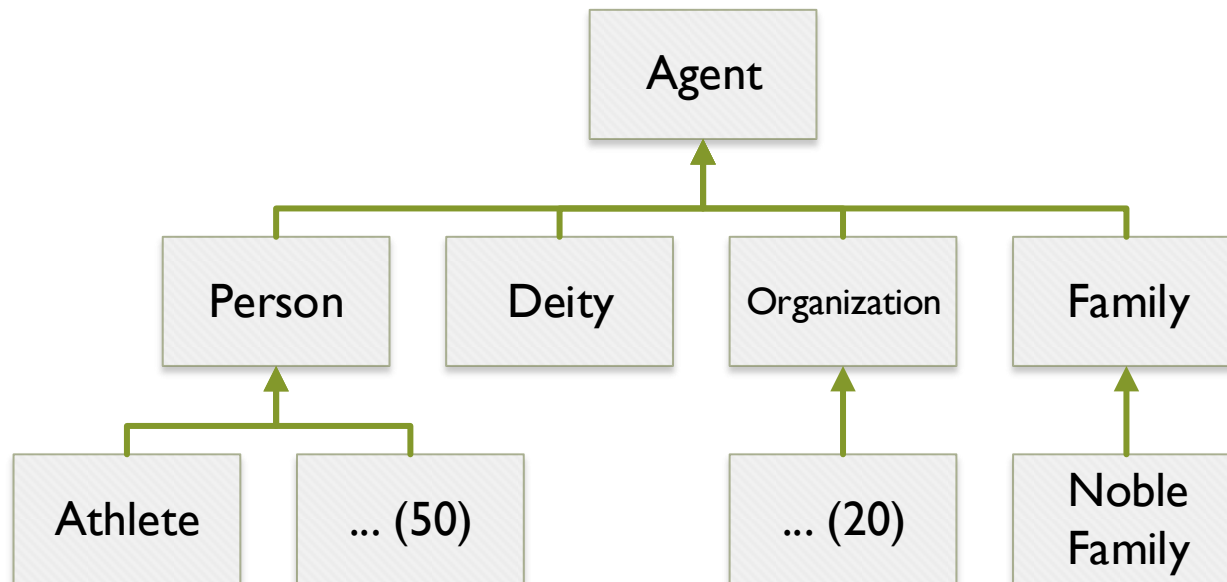
Overenie mapovania infoboxov na triedy

Hlavný jazyk	Zdrojový jazyk	Základná metrika			Špecializujúca sa metrika			Zovšeobecňujúca metrika		
		P	N	F1	P	N	F1	P	N	F1
SK	DE, EN, ES	0,97	0,81	0,88	0,86	0,90	0,88	0,98	0,57	0,72
	EN	0,92	0,83	0,87	0,81	0,85	0,83	0,98	0,57	0,72
	FR, IT, PT	0,94	0,86	0,90	0,87	0,92	0,89	0,94	0,64	0,76
CS	DE, EN, ES	0,96	0,92	0,94	0,76	0,96	0,85	1,00	0,42	0,59
	FR, EN, PT								0,23	0,42

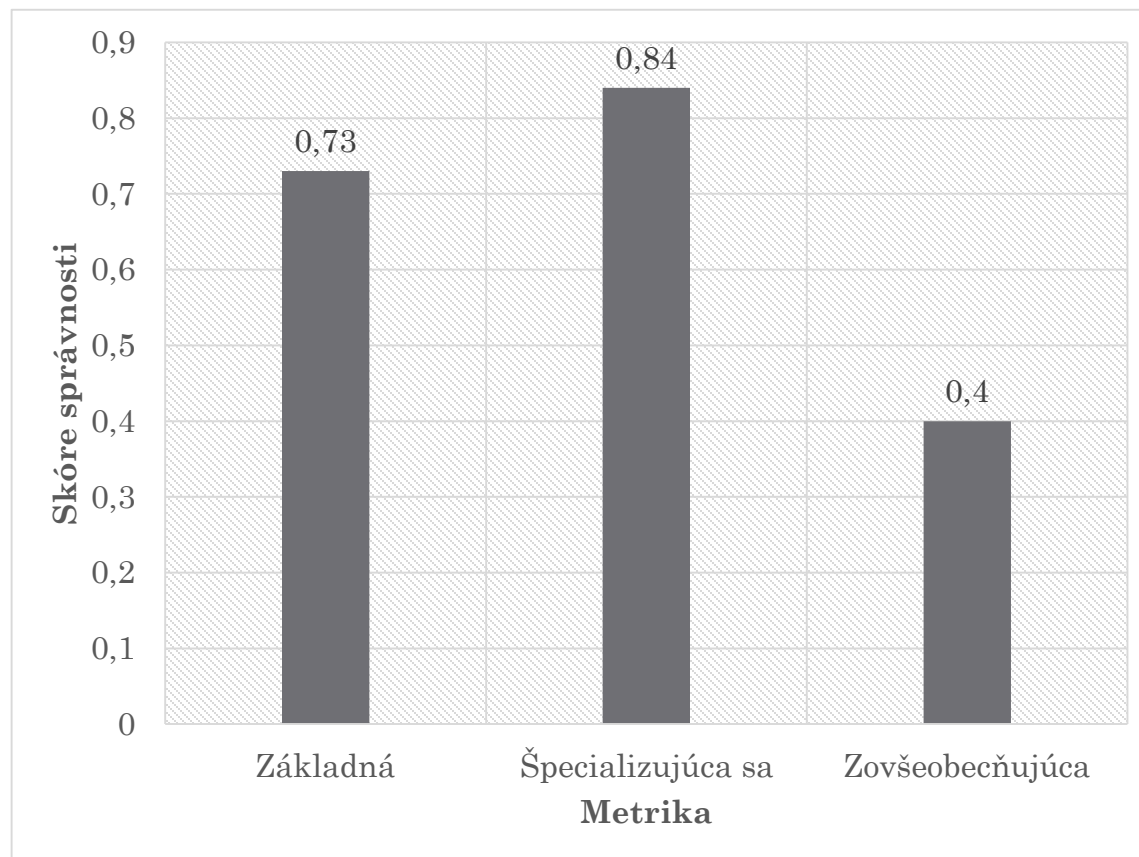
P – presnosť, N – návratnosť

Overenie mapovania infoboxov na triedy

- skóre správnosti (Melamed, et al., 2000)

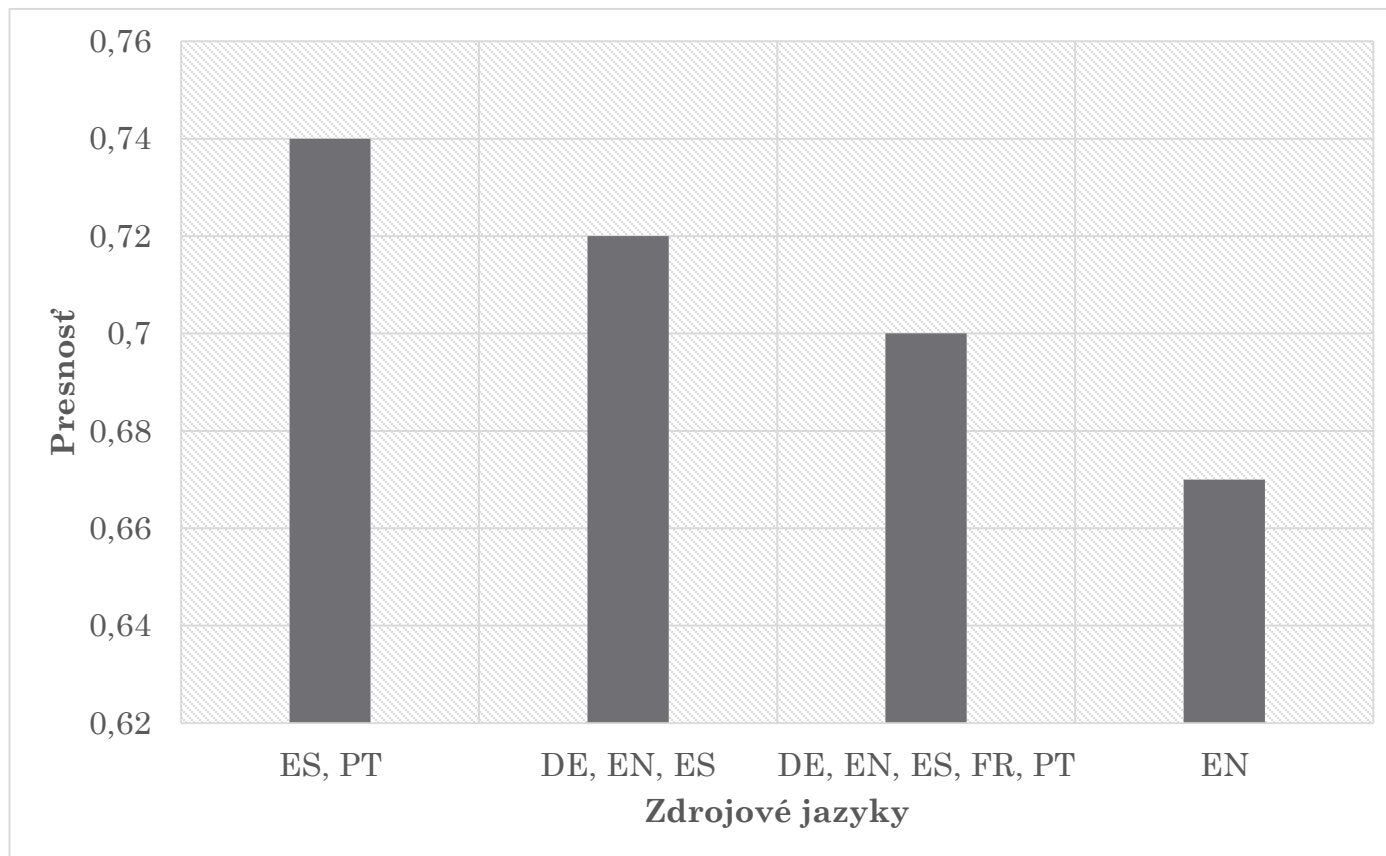


Overenie mapovania infoboxov na triedy



Overenie mapovania atribútov na predikáty

- zlatý štandard z (Aprosio, et al., 2013)



Zhodnotenie

- navrhnuté metódy pre automatické vytváranie mapovaní
- vytvorené prostredie pre budúce zlepšovanie metód návrhu mapovaní (bez obmedzenia na cieľový jazyk)
- overenie nášho riešenia dvoma spôsobmi
- vyextrahované trojice pre slovenčinu

Spojenie dvoch častí metódy

- atribúty majú rovnaké mapovanie pre rôzne typy infoboxov
- typ infoboxu ako vodítko pre vylúčenie určitých možností mapovania atribútov
- atribúty infoboxu ako vodítko pre určenie jeho typu

Overenie mapovania infoboxov na triedy

Hlavný jazyk	Zdrojový jazyk	Základná metrika			Špecializujúca sa metrika			Zovšeobecňujúca metrika		
		P	N	F1	P	N	F1	P	N	F1
SK	DE, EN, ES	0,97	0,81	0,88	0,86	0,90	0,88	0,98	0,57	0,72
	EN	0,92	0,83	0,87	0,81	0,85	0,83	0,98	0,57	0,72
	FR, IT, PT	0,94	0,86	0,90	0,87	0,92	0,89	0,94	0,64	0,76
CS	DE, EN, ES	0,96	0,92	0,94	0,76	0,96	0,85	1,00	0,42	0,59
	FR, EN, PT								0,23	0,42

P – presnosť, N – návratnosť

Overenie mapovania atribútov na predikáty

Zdrojové jazyky	Presnosť	Návratnosť	F1	Neoverených
ES, PT	0,74	0,020	0,039	149
DE, EN, ES	0,72	0,021	0,041	217
DE, EN, ES, FR, PT	0,70	0,023	0,045	274
EN	0,67	0,018	0,035	184