Extrakce informací z českého Wikislovníku

Autor: Bc. Martin Lukáš

Vedoucí: Ing. Václav Zeman

Oponent: prof. Ing. Vojtěch Svátek, Dr.

Vysoká škola ekonomická

Fakulta informatiky a statistiky

Katedra informačního a znalostního inženýrství

Cíle práce

Extrakce především morfologických informací
z českého Wikislovníku do RDF:

* prozkoumat předchozí extraktory Wiktionary
* navrhnout a implementovat extraktor pro český Wikislovník
* otestovat výsledný RDF model

Předchozí extraktory

* JWKTL – Java API k Wikislovníku (EN, DE, RU)
* wikokit – také Java API (EN, RU)
* Wiktionary2RDF – extrakce do RDF (6 verzí)
* DBnary – také do RDF, pro více než 20 verzí

Problém:

* žádné z nich se nesoustředí na morfologické informace (DBnary nově pro DE a FR)

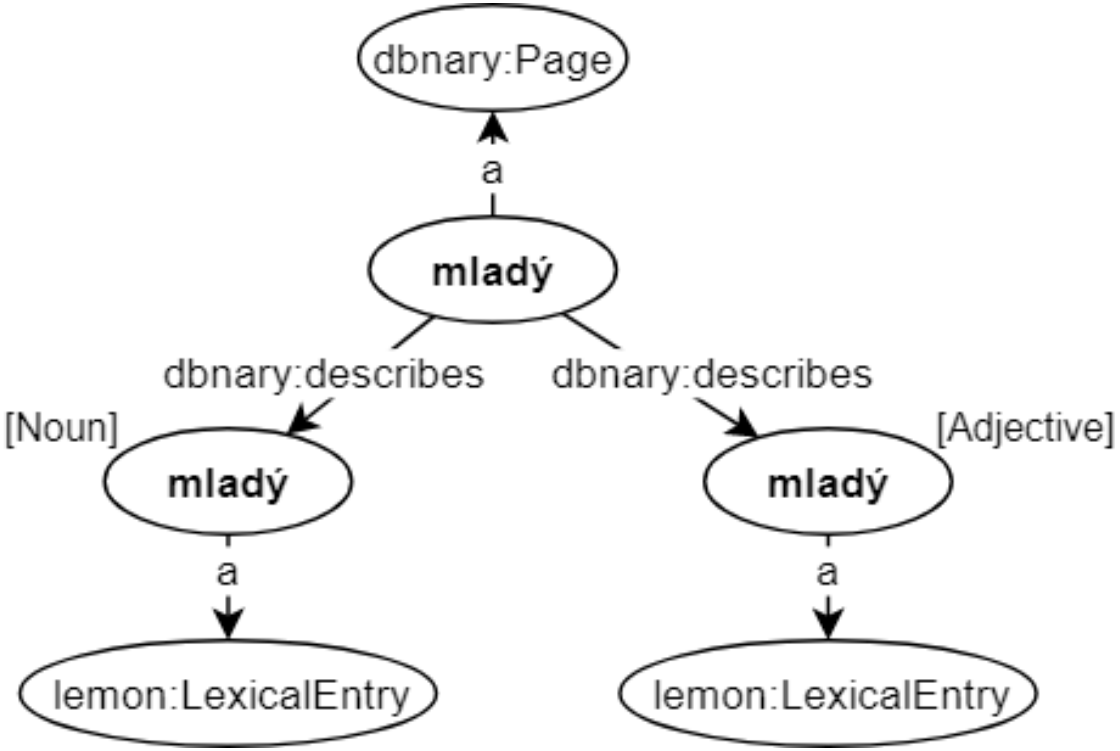
Struktura Wikislovníku

* jazykové sekce
  * výslovnost
  * sekce slovních druhů
    * vlastnosti slovního druhu
    * skloňování / stupňování / časování
    * …
  * …
* sekce externích odkazů
* …

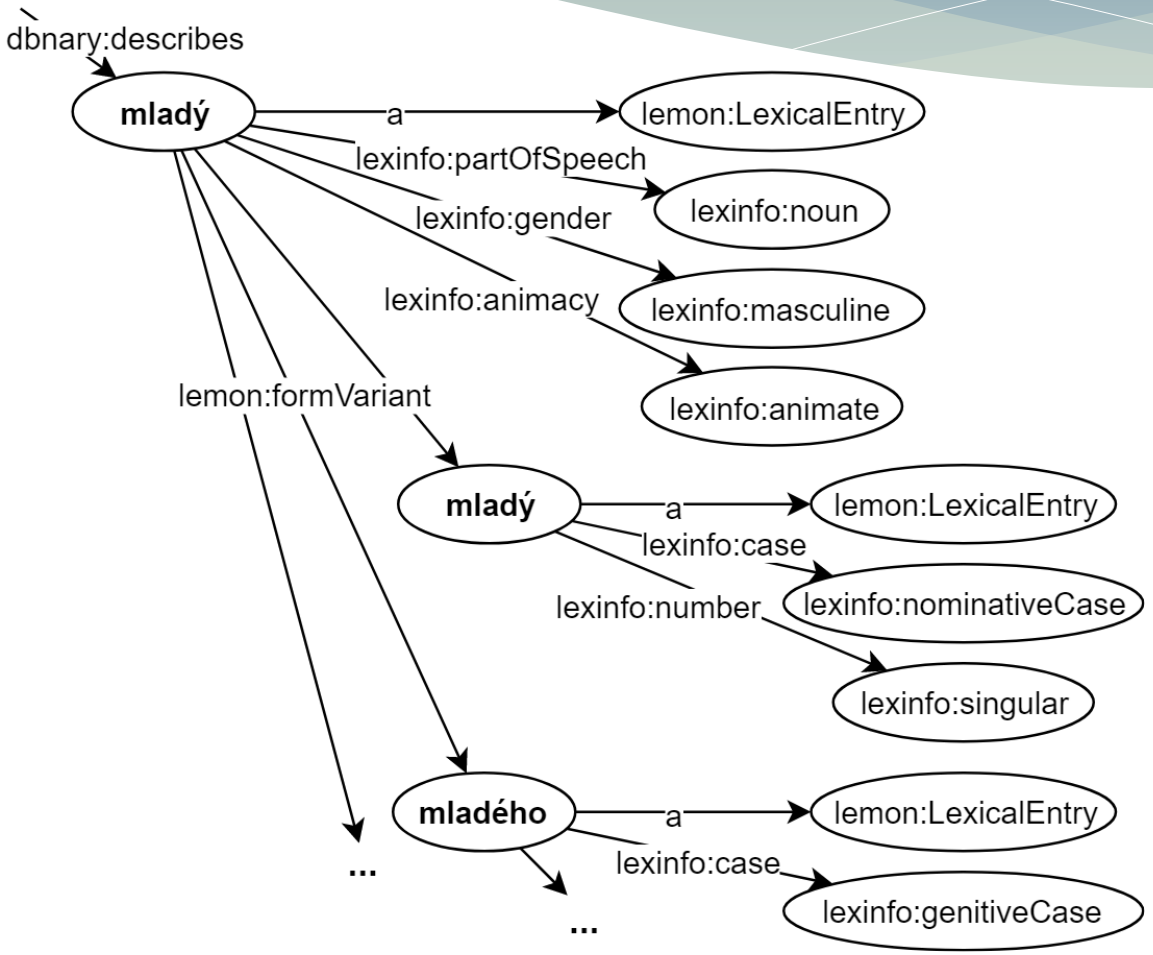
Implementace extraktoru

* extraktor v jazyce Java
* knihovna pro práci s RDF – Apache Jena
* ontologie na modelování struktury:
  * DBnary
  * Lemon
* ontologie na modelování morfologických informací:
  * LexInfo
  * OLiA-MTE

Struktura RDF modelu I.



Struktura RDF modelu II.



Struktura RDF modelu III.

* kromě morfologických informací také extrahován externí odkaz na článek na Wikipedii
* tento článek převeden na odkaz na **DBpedii**
* navázán na zdroj stránky pomocí *rdfs:seeAlso*

Analýza XML dumpu

Vlastnosti XML dumpu:

* celkový počet stránek:	143 399
* počet stránek o pojmech:	114 599
* počet stránek s českou jaz. sekcí:	46 151
* počet extrahovaných stránek:	**42 734**
* s externím odkazem na Wikipedii:	9 464

Analýza RDF datasetu

Vlastnosti RDF datasetu:

* počet trojic v modelu:	5 632 208
* počet stránek (*dbnary:Page*):	42 734
* počet zdrojů slovních druhů:	46 156
* počet slovníkových záznamů (*lemon:LexicalEntry*):	748 457
* počet unikátních slovních záznamů:	269 212

Uplatnění datasetu I.

* oprava informací na Wikislovníku
  * např. počet podstatných jmen s chybějícím rodem:

```
SELECT (count(?le) as ?no_of_nouns)
WHERE {
  ?page  a                    dbnary:Page ;
         dbnary:describes     ?le .
  ?le    a                    lemon:LexicalEntry ;
         lexinfo:partOfSpeech lexinfo:noun .
  filter not exists {?le lexinfo:gender ?gen}
}
```

Uplatnění datasetu II.A

* zjištění sémantiky slov pomocí DBpedie
* RDF model využit ke zjištění základního tvaru slova, a ten dosazen do SPARQL dotazu na DBpedii

Uplatnění datasetu II.B

* př. SPARQL dotaz umožňující vyhledat zdroje z DBpedie pro slovo „Sněžku“ – i když existuje pouze zdroj pro „Sněžka“

```
select distinct ?thing WHERE {
  ?word         rdfs:label         "Sněžku"@cs .
  {
    ?word       a                  lemon:LexicalEntry .
    ?word_pos   lemon:formVariant  ?word .
    ?word_page  dbnary:describes   ?word_pos ;
                rdfs:seeAlso       ?dbp_res .
    ?dbp_res    a                  ?thing .
  } UNION {
    ?word       a                  lemon:LexicalEntry .
    ?word_page  dbnary:describes   ?word ;
                rdfs:seeAlso       ?dbp_res .
    ?dbp        a                  ?thing .
  } UNION {
    ?word       a                  dbnary:Page ;
                rdfs:seeAlso       ?dbp_res .
    ?dbp        a                  ?thing .
  }
}
```

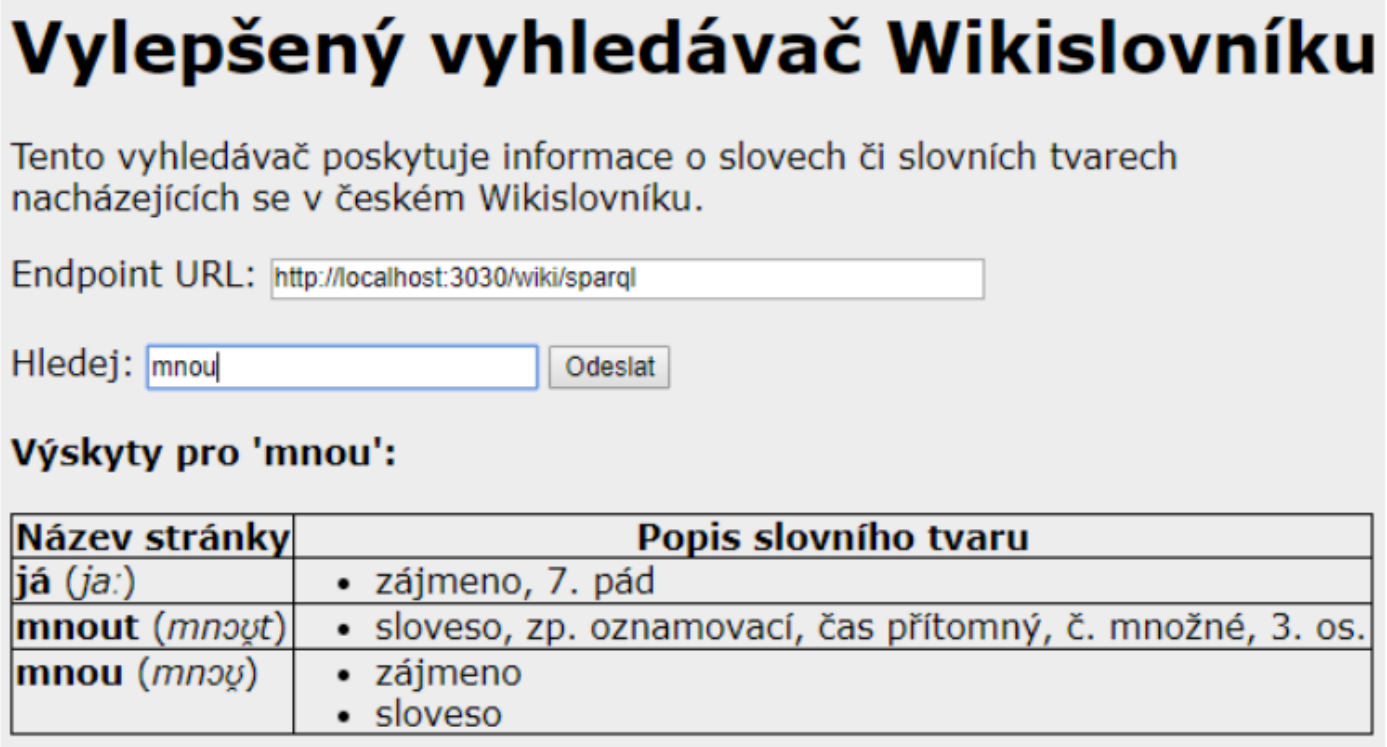
Uplatnění datasetu II.C

* výsledky pro SPARQL dotaz:

thing
http://www.w3.org/2002/07/owl#Thing
http://www.w3.org/2003/01/geo/wgs84_pos#SpatialThing
http://www.wikidata.org/entity/Q8502
http://schema.org/Mountain
http://schema.org/Place
http://dbpedia.org/ontology/Location
http://dbpedia.org/ontology/NaturalPlace
http://dbpedia.org/ontology/Place
http://dbpedia.org/ontology/Mountain

Uplatnění datasetu III.

* webová aplikace pro vylepšené vyhledávání:



Vylepšený vyhledávač Wikislovníku

Tento vyhledávač poskytuje informace o slovech či slovních tvarech nacházejících se v českém Wikislovníku.

Endpoint URL: http://localhost:3030/wiki/sparql

Hledej: mnou Odeslat

Výskyty pro 'mnou':

Název stránky	Popis slovního tvaru
já (*ja:*)	• zájmeno, 7. pád
mnout (*mnɔʊ̯t*)	• sloveso, zp. oznamovací, čas přítomný, č. množné, 3. os.
mnou (*mnɔʊ̯*)	• zájmeno • sloveso

Rozvoj v budoucnu

* doplnit program o extraktory dalších sekcí (dělení, etymologie, významy, překlady atd.)
* propojit RDF model s ostatními RDF modely Wikislovníku (např. DBnary či Wiktionary2RDF)