DBPEDIA

Václav Zeman

December 2015

KIZI - VŠE

CO JE DBPEDIA?

CO JE DBPEDIA?

- DBpedia je **komunita** lidí zaměřující se na získávání informací a znalostí z Wikipedie.
- DBpedia je **sada nástrojů**, které extrahují informace z Wikipedie.
- DBpedia je **báze propojených a strojově čitelných dat** získaných z Wikipedie.

CO JE DBPEDIA?

- DBpedia je **služba**, která dovoluje uživateli vyhledávat informace z Wikipedie sofistikovanějším způsobem.
- DBpedia je Wikipedie přizpůsobena ke strojovému **zpracování informací**.

CO JE DBPEDIA?

Wikipedia

Sdílená tvorba obsahu, **Web 2.0**

DBpedia

Sémantický web, propojená data, **Web 3.0**

K ČEMU JE DBPEDIA?

K ČEMU JE DBPEDIA?

- **Strojové čtení** a zpracování informací obsažených na Wikipedii.
- Sofistikované **vyhledávání** informací.
- Využití pro **objevování** nových znalostí.
- Jednoduché a přímé **odpovídání** na otázky, které uživatele zajímají:

Otázka
Nejvyšší hora v Česku?

Odpověď
Sněžka

K ČEMU JE DBPEDIA?

Nejvyšší hora v Česku?

Otázka v podobě sémantického dotazu (SPARQL)

```
SELECT ?hora {
  <http://cs.dbpedia.org/resource/Česko>
  <http://dbpedia.org/ontology/highestMountain>
  ?hora
}
```

Odpověď

```
http://cs.dbpedia.org/resource/Sněžka
```

K ČEMU JE DBPEDIA?

Všechny filmy, které režíroval Jan Svěrák?

Otázka v podobě sémantického dotazu (SPARQL)

```
SELECT ?film {
  ?film
  <http://dbpedia.org/ontology/director>
  <http://cs.dbpedia.org/resource/Jan_Svěrák>
}
```

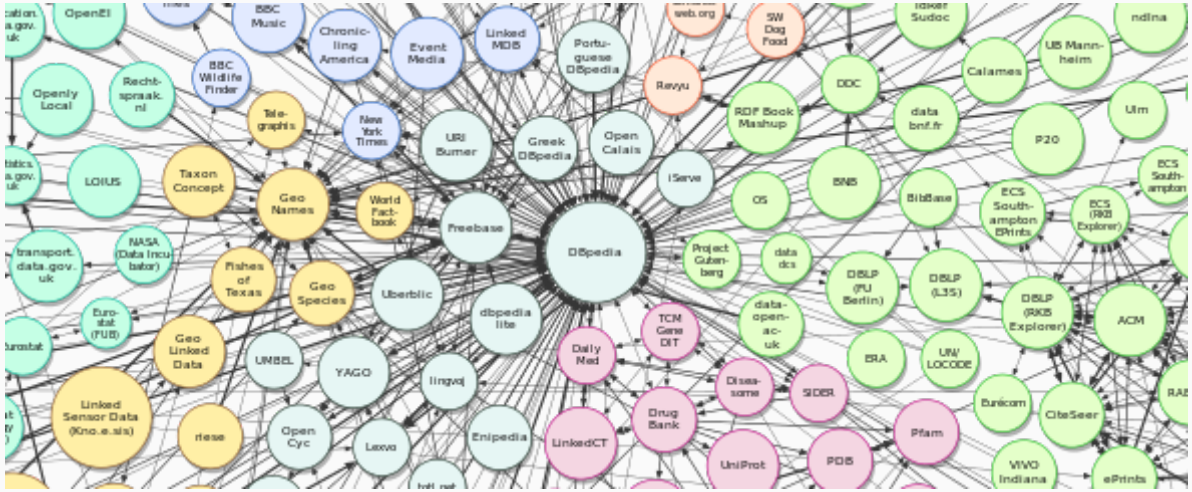
Odpověď

```
http://cs.dbpedia.org/resource/Jízda_(film)
http://cs.dbpedia.org/resource/Akumulátor_1
http://cs.dbpedia.org/resource/Kolja
http://cs.dbpedia.org/resource/Kuky_se_vrací
http://cs.dbpedia.org/resource/Obecná_škola_(film)
http://cs.dbpedia.org/resource/Tmavomodrý_svět_(film)
http://cs.dbpedia.org/resource/Tři_bratři_(film)
http://cs.dbpedia.org/resource/Vratné_lahve
```

K ČEMU JE DBPEDIA?

Technicky, DBpedia nabízí:

- Vyextrahované informace z Wikipedie v podobě **RDF** datasetů.
- **SPARQL** endpoint pro sémantické dotazování.
- Vlastní rozsáhlou **ontologii** pro zajištění konzistence vyextrahovaných dat.
- Možnost **propojení** s jinými informačními bázemi.



JAK VZNIKÁ DBPEDIA?

JAK VZNIKÁ DBPEDIA?

Prague

From Wikipedia, the free encyclopedia

Coordinates: 50°05′N 14°25′E

This article is about the capital of the Czech Republic. For other uses, see Prague (disambiguation).

Prague (/ˈprɑːɡ/; Czech: *Praha* pronounced [ˈpraɦa] (listen)) is the capital and largest city of the Czech Republic.[4] Situated in the north-west of the country on the Vltava river, the city is home to about 1.3 million people, while its metropolitan area is estimated to have a population of over 2.3 million.[5] The city has a temperate oceanic climate with warm summers and chilly winters.

Prague has been a political, cultural and economic centre of Europe[citation needed] and particularly central Europe[citation needed] during its 1,100 year existence. For centuries, during the Gothic and Renaissance eras, Prague was the permanent seat of two Holy Roman Emperors and thus was also the capital of the Holy Roman Empire.[citation needed] Later it was an important city in the Habsburg Monarchy and the Austro-Hungarian Empire,[citation needed] and after World War I became the capital of Czechoslovakia. The city played major roles in the Protestant Reformation, the Thirty Years' War, and in 20th-century history, during both World Wars and the post-war Communist era.

Prague is home to a number of famous cultural attractions, many of which ...ved the violence and destruction of twentieth century Europe. Main attractions inc... the following: Prague Castle, the Charles Bridge, Old Town Square, the Jewish Qua... the Lennon Wall, and Petřín hill. Since 1992, the extensive historic centre of Prague has been included in the UNESCO list of World Heritage Sites.

Prague boasts more than ten major museums, along with countless theatres, galleries, cinemas, and other historical exhibits. Also, Prague is home to a wide range of public and private schools, including the famous Charles University. Its rich history makes it a popular tourist destination, and the city receives more than 4.1 million international visitors annually, as of 2009.[6][7] Prague is classified as a global city.[citation needed]

A modern public transportation system connects the city. Prague is also accessible by road, train, and air.

Contents [hide]

1 History
- 1.1 Ancient age
- 1.2 The era of Charles IV
- 1.3 Habsburg era

Prague
Czech: *Praha*

About: Prague

An Entity of Type : populated place, from Named Graph : http://dbpedia.org, within Data Space : dbpedia.org

Prague is the capital and largest city of the Czech Republic. Situated in the north-west of the country on the Vltava river, the city is home to about 1.3 million people, while its metropolitan area is estimated to have a population of over 2.3 million. The city has a temperate oceanic climate with warm summers and chilly winters. Prague has been a political, cultural and economic centre of Europe and particularly central Europe during its 1,100 year existence.

dbpedia-owl:areaTotal	496000000.000000 (xsd:double)
dbpedia-owl:country	dbpedia:Czech_Republic
dbpedia-owl:leaderName	dbpedia:Bohuslav_Svoboda
dbpedia-owl:leaderParty	dbpedia:Civic_Democratic_Party_(Czech_Republic)
dbpedia-owl:leaderTitle	Mayor
dbpedia-owl:maximumElevation	399.000000 (xsd:double)
dbpedia-owl:motto	(Prague, Head of the State; Latin)
dbpedia-owl:populationAsOf	2011-01-14 (xsd:date)
dbpedia-owl:populationMetro	2300000 (xsd:integer)
dbpedia-owl:populationTotal	1290846 (xsd:integer)
dbpedia-owl:postalCode	1xx xx
dbpedia-owl:thumbnail	http://upload.wikimedia.org/wikipedia/commons/thumb
dbpedia-owl:timeZone	dbpedia:Central_European_Time

ZDROJ DAT

Pro extrakci informací z Wikipedie je nejprve nutné stáhnout všechny Wikipedia stránky. Jak toho docílit?

- Wikipedia **dump** = Jeden soubor obsahující všechny wiki stránky
- Dump je **veřejně dostupný** na adrese dumps.wikimedia.org
- Aktualizace dumpu probíhá jednou za měsíc.
- Velikost dumpu obsahující všechny stránky české Wikipedie je po rozbalení cca 2.15 GB

EXTRAKČNÍ FRAMEWORK

Jednotlivé informace z Wikipedia stránek jsou získávány pomocí tzv. extraktorů.

- Primární **extrakční framework**:
  - je open source, veřejně dostupný na githubu
  - obsahuje sadu extraktorů implementovaných v jazyce Scala/Java
- Každý extraktor extrahuje právě jeden typ informací z Wikipedie.
- Informace se extrahují hlavně pomocí regulárních výrazů, ale také s použitím metod strojového učení (pokročilejší extraktory).
- Možnost zapojení vlastních extraktorů do DBpedie.
- Linked Hypernym Dataset[1] = extraktor vyvinutý na VŠE, součástí DBpedie od roku 2015.

[1] http://ner.vse.cz/datasets/linkedhypernyms/

EXTRAKČNÍ WORKFLOW

Source: Zdrojem je Wikipedia stránka ve wiki formátu.

```
'''Prague''' ({{IPAc-en|'|p|r|a:|g}};
{{lang-cs|Praha}}, {{IPA-cs|'praha||Cs-Praha.ogg}})
is the capital and [[List of cities in the
Czech Republic|largest city]] of the [[Czech Republic]]
```

WikiParser: Převádí obsah stránky v podobě prostého textu na vlastní datovou strukturu.

Extractor: Z načtené Wikipedia stránky extrahuje informace v podobě trojic.

Destination: Finální uložení trojic do RDF datasetů.

Klíčové DBpedia extraktory

LabelExtractor

Extrahuje názvy Wikipedia stránek.

PageLinksExtractor

Extrahuje interní linky mezi Wikipedia stránkami.

CategoryLabelExtractor

Extrahuje kategorie Wikipedia stránek.

DisambiguationExtractor

Extrahuje rozcestníky.

RedirectExtractor

Extrahuje synonyma názvů Wikipedia stránek.

InfoboxExtractor

Extrahuje informace z takzvaných infoboxů (tabulky v pravé části článků na Wikipedii).

- **Problém:** Názvy jednotlivých vlastností uvnitř infoboxů nejsou konzistentní. Různé názvy pro různé jazykové verze a typy infoboxů.
- **Řešení:** Mapování vlastností z infoboxů na DBpedia vlastnosti definované v rámci jedné konzistentní ontologie.

```
<http://cs.dbpedia.org/resource/Česko>
<http://cs.dbpedia.org/property/nejvyššíHora>
<http://cs.dbpedia.org/resource/Sněžka> .
```

Česká republika

Vlajka

Znak

Hymna: *Kde domov můj*

Geografie

Poloha Česka

Hlavní město:	Praha
Rozloha:	78 866[1] km² (113. na světě) z toho 2 % vodní plochy
Nejvyšší bod:	Sněžka (1603 m n. m.)
Časové pásmo:	+1 +2 (letní čas)
Poloha:	50°0′ s. š., 16°0′ v. d.

MappingExtractor

Mapuje vlastnosti z infoboxů na vlastnosti z DBpedia ontologie.

- Extraktor využívá tzv. mapovací pravidla.
- Mapovací pravidla se vytvářejí ručně na stránce mappings.dbpedia.org
- Namapované vlastnosti jsou konzistentní v rámci všech jazykových verzí DBpedie.

```
<http://cs.dbpedia.org/resource/Česko>
<http://dbpedia.org/ontology/highestMountain>
<http://cs.dbpedia.org/resource/Sněžka> .
```

Mapping cs:Infobox stát

Template Mapping (help)	
map to class	Country

Mappings

Property Mapping (help)	
template property	úřední název
ontology property	longName

Property Mapping (help)	
template property	hymna
ontology property	anthem

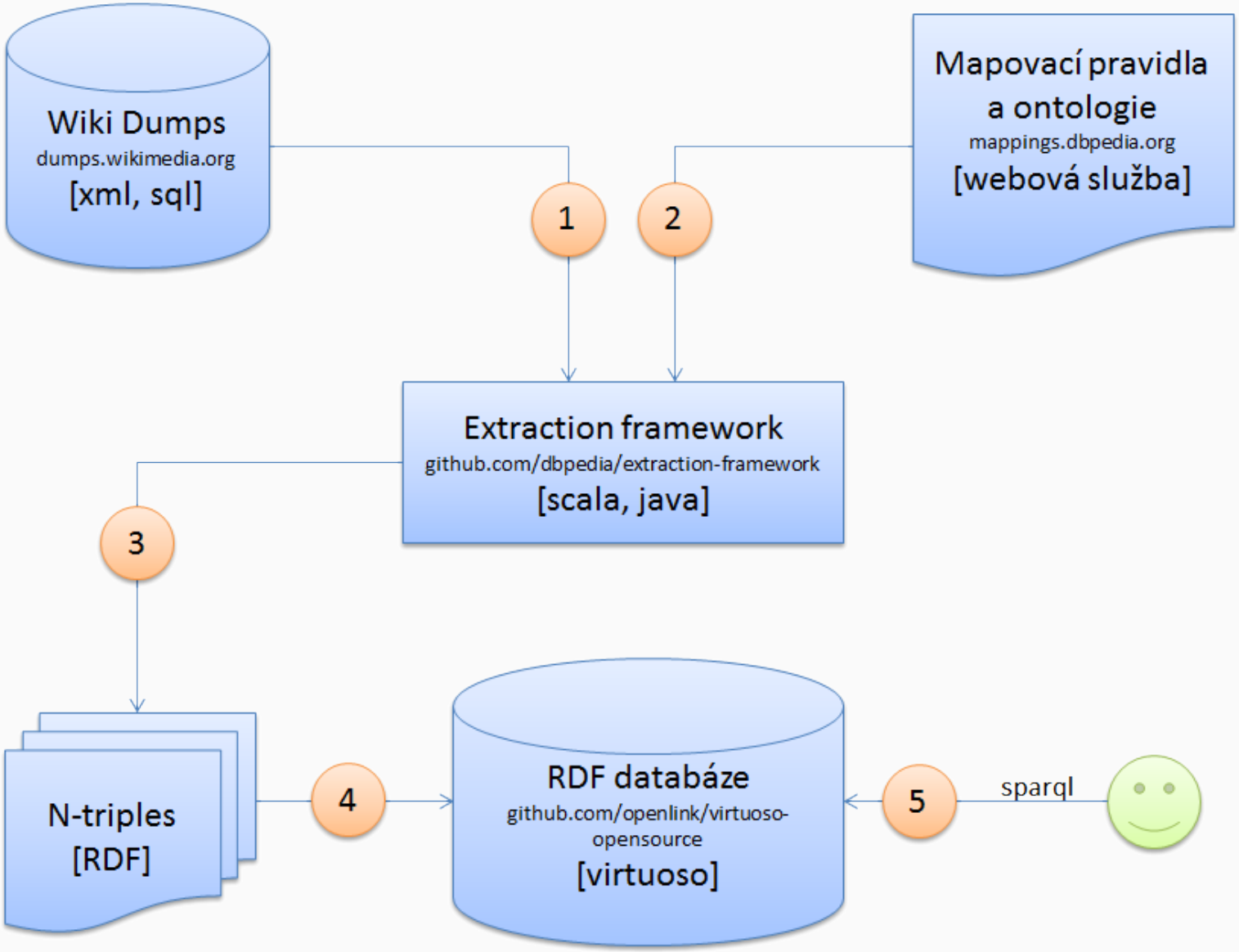
Property Mapping (help)	
template property	hlavní město
ontology property	capital

Property Mapping (help)	
template property	procent vody
ontology property	percentageOfAreaWater

Property Mapping (help)	
template property	nejvyšší hora
ontology property	highestMountain

Property Mapping (help)	
template property	hustota
ontology property	populationDensity
unit	inhabitantsPerSquareKilometre

KOMPLETNÍ WORKFLOW



PUBLIKOVÁNÍ VYEXTRAHOVANÝCH INFORMACÍ

- Jako **množina souborů** obsahující strojově čitelná a propojená data dle specifikace RDF (N-Triples, RDF/XML, JSON-LD, CSV aj.).
- Jako **služba**, ve které je možné vyhledávat informace dle sémantických dotazů (SPARQL endpoint).
- Jako **webové stránky**, které vizualizují veškeré vyextrahované informace.

Česko

Country, Place, PopulatedPlace

Česko, úředním názvem Česká republika, je stát ve střední Evropě. Vznikl 1. ledna 1969 jako formálně svrchovaný národní stát pod názvem Česká socialistická republika v rámci federalizace Československa, od 6. března 1990 nese název Česká republika, 1.

cs.wikipedia.org/wiki/Česko

Property:	Value:
dbpedia-owl:PopulatedPlace/populationDensity :	133.0
dbpedia-owl:areaCode :	+420 (xsd:string)
dbpedia-owl:areaTotal :	7.8866e+10 (xsd:double)
dbpedia-owl:capital :	dbpedia-cs:Praha
dbpedia-owl:currency :	dbpedia-cs:Koruna_česká
dbpedia-owl:governmentType :	dbpedia-cs:Parlamentní_republika
dbpedia-owl:highestMountain :	dbpedia-cs:Sněžka
dbpedia-owl:leaderName :	dbpedia-cs:Bohuslav_Sobotka dbpedia-cs:Miloš_Zeman
dbpedia-owl:leaderTitle :	Prezident @cs

DBPEDIA VS WIKIDATA

- Akademický decentralizovaný projekt
- Pouze extrahuje informace z Wikipedia stránek
- Důraz je kladen na **kvantitu** informací
- Aktualizováno 2x za rok

- Spravuje přímo Wikimedia Foundation
- Informace jsou ručně vytvářeny komunitou, stejně jako Wikipedia
- Důraz je kladen na **kvalitu** informací
- Aktualizováno v reálném čase

ČESKÁ DBPEDIA

ČESKÁ DBPEDIA

Obsahuje více než 29,5 miliónů výroků (trojic).

K dispozici je:

- Kompletní množina vyextrahovaných dat (volně ke stažení ∼6GB)
- Služba pro sémantické dotazování (SPARQL endpoint)
- Náhledy na vyextrahované informace pro jednotlivé Wikipedia stránky

http://cs.dbpedia.org

VYUŽITÍ DBPEDIE

Česká DBpedie je momentálně spravována katedrou informačního a znalostního inženýrství na VŠE.

Školní projekty postavené na DBpedii:

- **Targeted Hypernym Discovery:**[2] Automatická sémantická anotace textu.
- **Linked Hypernym Dataset:**[3] Nástroj využívající metod strojového učení pro odvození typu Wikipedie stránky dle první věty abstraktu.
- **DB-quiz:**[4] Vědomostní hra odvozená od populární televizní soutěže AZ-kvíz. Otázky jsou automaticky generovány z české a anglické DBpedie.

[2] http://ner.vse.cz/thd/
[3] http://ner.vse.cz/datasets/linkedhypernyms/
[4] http://mynarz.net/db-quiz/

DĚKUJI ZA POZORNOST