

4IZ440 Propojená data na webu

Organizační a „kontextový“ úvod

Vyučující: Doc. Ing. Vojtěch Svátek, Dr.

Zimní semestr 2016

<http://nb.vse.cz/~svatek/rzzw.html>

Náplň předmětu

- Technologie **propojených dat** (linked data, LD) na (sémantickém) webu
- Cílem je naučit se data na sémantickém webu
 - **vytvářet a vystavovat**, včetně
 - návrhu schémat / slovníků
 - extrakce dat, která nativně v RDF nejsou
 - **zpracovávat** (především propojovat)
 - **využívat** (v aplikacích)
- Důraz je na aktivní přístup a interakci
 - Nebude to „nalejvárna“...

Co předmět je a není?

- Nejde o desítky let ustálenou disciplínu, pro kterou by existovaly „nepohnutelné“ osnovy a učebnice (i když dobré knihy a příručky už vyšly!)
- Obsahem výuky není jedna přesně vymezená technologie (jazyk, metodika, software...)

ALE

- Jde o relativně volné propojení jazykových standardů, softwarových nástrojů, výzkumných iniciativ...
- „Správné odpovědi“ na mnohé klíčové otázky v oboru zatím nikdo na 100% nezná
- Vzhledem k tomu je u studentů velmi vítána vlastní iniciativa, zvědavost, zkoumání toho, co se nově objevuje na webu, na konferencích apod.!

Zdroje pro studium

- Základní literatura:
 - **kniha**: *Linked Data: Evolving the Web into a Global Data Space* od T. Heatha a C. Bizera, online <http://linkeddatabook.com>
 - Tutoriál L. Feigenbaum, E. Prud'hommeaux: *SPARQL by Example*, online <http://www.cambridgesemantics.com/semantic-university/sparql-by-example>
 - Další tutoriály ze Semantic University na <http://www.cambridgesemantics.com>
 - Slidy k přednáškám, budou postupně vystavovány na webu
 - Kapitoly z knihy **Umělá inteligence 6**, Academia (knihovna/studovna/zakoupit?)
 - učební text *Aktuální problémy a perspektivy sémantického webu* přístupný z webu předmětu
- Doplňková literatura:
 - Viz web; zvl. B. DuCharme: *Learning SPARQL* (knihovna/studovna)
 - Specifikace používaných **jazyků a formátů** (RDF, Turtle, SPARQL, RDFa) - jsou dostupné na WWW
 - V případě zájmu mnoho další literatury u vyučujícího, články na WWW a v digitálních knihovnách

Kontrolní testy

- Teoretické znalosti budou ověřeny třemi písemnými **testy** (**11+15+14 = 40 bodů**)
 - psaní „kódu“, grafy, odpovědi volným textem, zaškrtačky
- Bodové minimum z každého testu je 5; v jednom testu lze minimum nesplnit bez dalších sankcí
- V případě nedosažení minima ve dvou testech lze o absolvování předmětu uvažovat jen pokud budou splněny všechny ostatní povinnosti a součet bodů dosáhne 60
 - následuje ústní přezkoušení, a pokud je úspěšné, student získá známku „vyhověl“ bez ohledu na celkový počet bodů
- *Testy budou zřejmě ve 4., 8. a 12. týdnu semestru, kvůli kolizi s jiným předmětem budou na cvičeních*

Praktické úkoly

- Zpracovat samostatně **semestrální projekt** (max. **40 bodů**)
 - *Detaily a termíny odevzdání budou upřesněny*
- Zpracovat samostatně ústní a písemný **minireferát** (max. **15 bodů**)
 - Ústní prezentace 28.11. nebo 5.12.
 - Z odborného článku z hlavní vědecké konference v oboru - ISWC 2016 (vysoce doporučeno, po dohodě lze i z jiné tematicky vhodné konference, případně časopisu), přístup k článkům bude zajištěn
- Průběžné úkoly na cvičeních (celkem až **15 bodů**), předběžně
 - De/serializace RDF, zkracování Turtle
 - Tvorba dotazů SPARQL na různých úrovních náročnosti (dotazování, kontrola a transformace dat, propojování...)
 - Případová studie extrakce, čištění, propojení a využití fiskálních dat
 - Návrh jednoduchého datového slovníku
 - Průběžná prezentace semestrálního projektu

Předběžný harmonogram cvičení

1. RDF, Turtle
2. Turtle, SPARQL ukázky
3. SPARQL – základy, seznámení s cvičebnicí
4. SPARQL pokročilejší, *Test 1*
5. SPARQL pokročilejší
6. Extrakce RDF z tabulkových fiskálních dat
7. Čištění a propojování fiskálních dat pomocí SPARQL
8. Využití fiskálních dat v aplikaci, *Test 2*
9. Návrh datového slovníku - úvod
10. Návrh datového slovníku – pokročilé metody
11. Extrakce dat z Wikipedie
12. (Rezerva,) *Test 3*
13. Prezentace vznikajících semestrálních projektů

A k čemu Vám to dále může být?

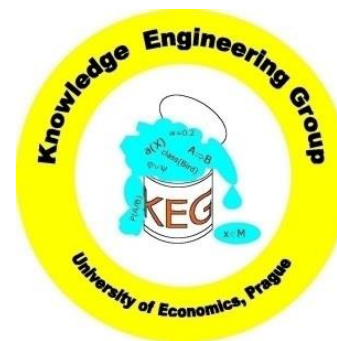
- Pro studenty bezprostředně směřující do praxe
 - Nový pohled na techniky, které se v praxi běžně využívají (integrace datových zdrojů, textová a webová analytika, architektury orientované na služby, metamodelování IS, groupware, ...)
 - Připravenost na novinky, které do praxe zřejmě proniknou v příštích 5-10 letech
- Pro studenty se zájmem o prozkoumávání neprozkoumaného
 - Přehled významné části výzkumných témat, která jsou na KIZI řešena a mohou být předmětem doktorských disertací, ale i diplomových prací

Související předměty

- 5FI430 – Znalosti a ontologické inženýrství
 - M. Vacura, KFIL, povinné pro KI
 - Problematika ontologických modelů jako sofistikovanější varianty datových slovníků, žijící „svým životem“ – formální odvozování nad koncepty, filozofické ukotvení atd.
- 4IZ470 – Dolování znalostí z webu
 - V. Svátek, povinné pro ZWT (zpravidla již absolvovali...)
 - Součástí „web miningu“ je extrakce informací z webových textů – komplementární k převodu již strukturovaných dat do RDF
- 4IZ530 – Logické programování
 - Š. Sem, volitelný předmět
 - Programování v jazyce Prolog; má řadu společných rysů s psaním dotazů ve SPARQL, lze v něm psát sémantické aplikace

Pro hlubší zájemce

- Pracovní skupina KIZI **SWOE**
 - „Semantic web and ontological engineering“, viz <http://kizi.vse.cz/swoe>
 - účastní se i studenti
- Neformální výzkumný seminář **KEG**
„Knowledge Engineering Group“
 - některé čtvrtky od 16 hodin
zpravidla na 473NB (zasedačka FIS)
 - možno přijít bez předchozího přihlášení
 - viz <http://keg.vse.cz/seminars>
 - kdo chcete dostávat oznámení, napište!
- Aktivity mezi-institucionální iniciativy **OpenData.cz**: <http://opendata.cz>
 - vč. společného týmu KIT a KIZI, <http://opendata.vse.cz/>



Pro hlubší zájemce

- Možnost zapojení do **vědeckých projektů** KIZI a fakulty s touto problematikou – na evropské, národní i školní úrovni
 - Vývojářské (Java, Scala, PHP, JS, Python aj.), ev. výzkumné činnosti vč. možné spoluúčasti na publikacích a cest na konference
 - Honorováno přes mimořádná stipendia nebo DPP
 - Možnost zpracovat DP na aktuální témata řešená světovým výzkumem
- Zejména evropský projekt v programu Horizont 2020:

