4IZ440 Propojená data na webu

Organizační a „kontextový“ úvod

Vyučující: Doc. Ing. Vojtěch Svátek, Dr.

Zimní semestr 2015

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, zaškrtávačky
- 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*

Praktické úkoly

- Zpracovat samostatně semestrální projekt (max. 40 bodů)
  - *Detaily a termíny odevzdání budou upřesněny*
- Zpracovat samostatně ústní a písemný minireferát (max. 15 bodů) - v průběhu semestru (prezentace 30.11.)
  - První možnost: z odborného článku z hlavní vědecké konference v oboru - ISWC 2015, přístup k článkům bude zajištěn
  - Druhá možnost: zpracovat s využitím širokého okruhu materiálů podrobnou odpověď na otázku položenou na portálu http://answers.semanticweb.com/
- Průběžné úkoly na cvičeních (celkem až 15 bodů), předběžně
  - De/serializace RDF, zkracování Turtle (3 b)
  - Tvorba dotazů SPARQL na různých úrovních náročnosti a pro různé úlohy (dotazování, kontrola a transformace dat, propojování…)
  - Převod dat do RDF (OpenRefine, mapování pro Wikipedii)
  - Průběžná prezentace semestrálního projektu

Předběžný harmonogram cvičení

1. RDF, Turtle
2. *(odpadá)*
3. SPARQL ukázky
4. SPARQL 1.1 - základy
5. SPARQL pokročilejší
6. Tvorba RDF z tabulkových dat
7. Kontroly dat pomocí SPARQL
8. SPARQL (UPDATE)
9. *(odpadá)*
10. Propojování dat
11. Programový přístup k RDF přes JSON-LD
12. Extrakce z Wikipedie
13. Prezentace vznikajících semestrálních projektů

Hosté na cvičeních

- Jindřich Mynarz
  - pokročilý SPARQL, tvorba datových slovníků, data veřejné správy (rozpočty, veřejné zakázky, číselníky…) vč. modelu DataCube, herní aplikace LD
- Marek Dudáš
  - webový mark-up (RDFa), slovníky pro e-commerce, vizualizace dat a slovníků (LODSight), datová rozhraní JSON
- Václav Zeman
  - (anglická i česká) DBpedia, data mining nad LD
- Ondřej Zamazal
  - programový přístup k RDF a slovníkům, software pro ontologické inženýrství

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 možná 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

Navazující 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 ZTW
  - Součástí „web miningu“ je extrakce informací z webových textů – komplementární k převodu již strukturovaných dat do RDF

Pro hlubší zájemce

- Pracovní skupina KIZI **SWOE**
  - „Semantic web and ontological engineering“, viz http://kizi.vse.cz/swoe
  - účastní se i studenti (Horáková, Hanzal, Hazuza)

…weeding the semantic web garden

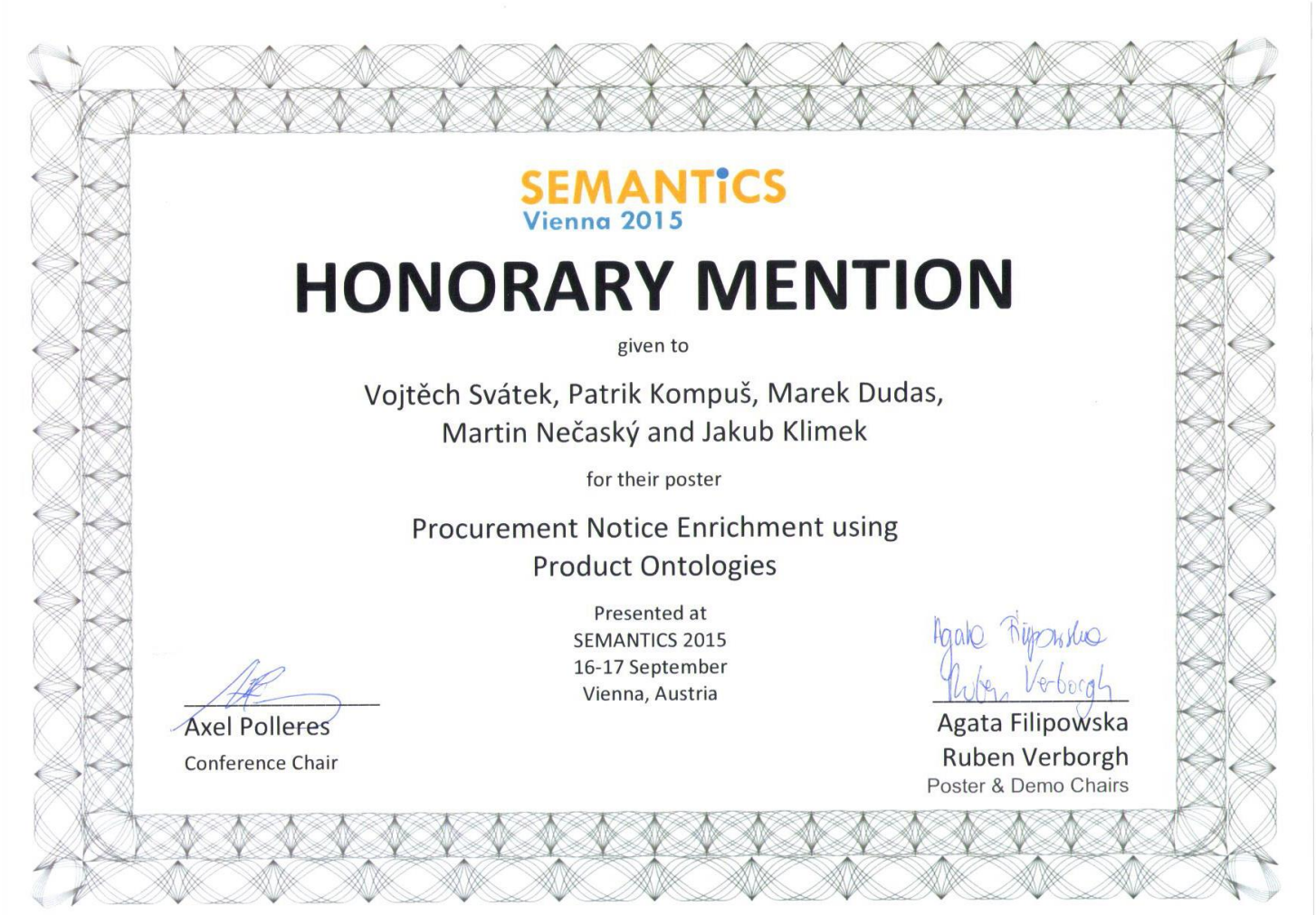
- Neformální výzkumný seminář **KEG** „Knowledge Engineering Group“
  - některé čtvrtky od 10.30 do cca 12.00 zpravidla na 473NB (zasedačka FIS)
  - možno přijít bez předchozího přihlášení
  - viz http://keg.vse.cz/seminars
  - kdo chcete dostávat oznámení, napište!
- Aktivity mezi-institucionální iniciativy **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

- Konference **Data+Znalosti** 1.-2.10.2015
  - Viz http://dataznalosti.cz
  - Mj. příspěvky odvozené ze 2 DP na KIZI (P. Hazuza, Š. Turečková)

Aplikace využívající sémantické modely v oblasti veřejných zakázek (spolupráce s MFF UK a FIT ČVUT) získala čestné uznání na konferenci SEMANTiCS 2015



SEMANTiCS
Vienna 2015

HONORARY MENTION

given to

Vojtěch Svátek, Patrik Kompuš, Marek Dudas,
Martin Nečaský and Jakub Klimek

for their poster

Procurement Notice Enrichment using
Product Ontologies

Presented at
SEMANTICS 2015
16-17 September
Vienna, Austria

Axel Polleres
Conference Chair

Agata Filipowska
Ruben Verborgh
Poster & Demo Chairs