



Propojená data na webu (motivační úvod)

prof. Ing. Vojtěch Svátek, Dr.
Katedra informačního a znalostního inženýrství

ZS 2018



Propojená data na webu

- Jeden z frekventovaných termínů, které označují téměř totéž, např. v angličtině:
 - Linked (open?) data (on the web)
 - Web of data
 - Web of entities
 - Semantic web
 - Semantic technology ☹️
 - Semantics ☹️☹️



„Běžný“ vs. sémantický web



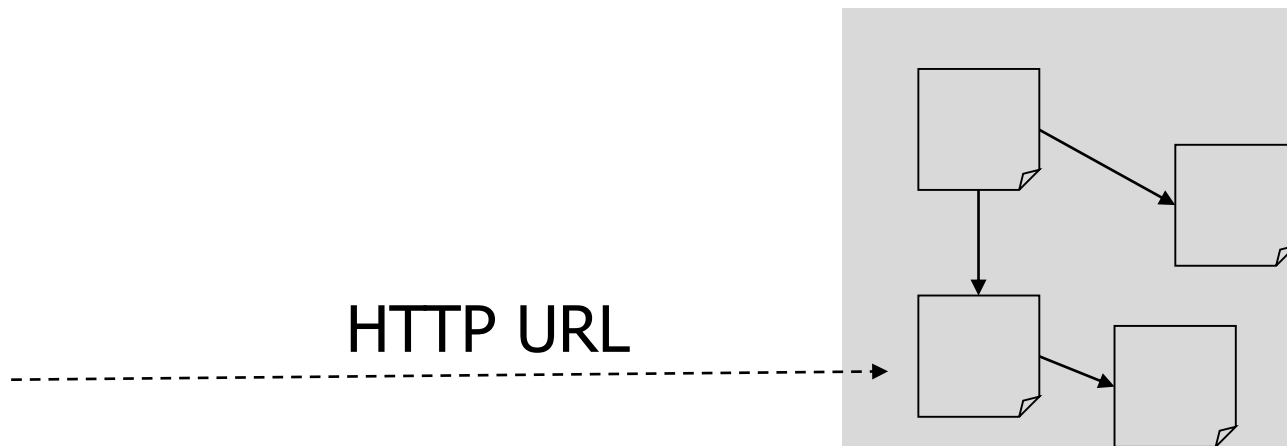
„Běžný“ vs. sémantický web

- „Běžný“ web
 - složený z dokumentů určených pro lidské uživatele
 - dokumenty jsou propojeny netypovanými odkazy

„Běžný“ vs. sémantický web

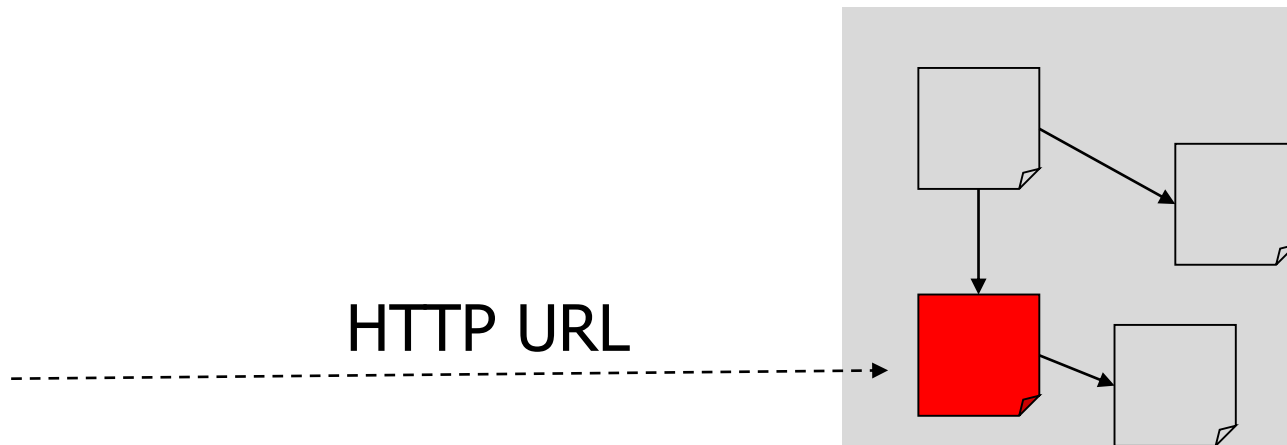
- „Běžný“ web

- složený z dokumentů určených pro lidské uživatele
- dokumenty jsou propojeny netypovanými odkazy
- požadavek HTTP jde na server, ten vrátí dokument



„Běžný“ vs. sémantický web

- „Běžný“ web
 - složený z dokumentů určených pro lidské uživatele
 - dokumenty jsou propojeny netypovanými odkazy
 - požadavek HTTP jde na server, ten vrátí dokument





„Běžný“ vs. sémantický web

- Idea **sémantického webu** (T. Berners Lee)



„Běžný“ vs. sémantický web

- Idea **sémantického webu** (T. Berners Lee)
 - propojit nejen dokumenty, ale i **strukturovaná data**



„Běžný“ vs. sémantický web

- Idea **sémantického webu** (T. Berners Lee)
 - propojit nejen dokumenty, ale i **strukturovaná data**
 - opatřit data **strojově čitelným významem**



„Běžný“ vs. sémantický web

- Idea **sémantického webu** (T. Berners Lee)
 - propojit nejen dokumenty, ale i **strukturovaná data**
 - opatřit data **strojově čitelným významem**
 - propojit je **typovanými odkazy**



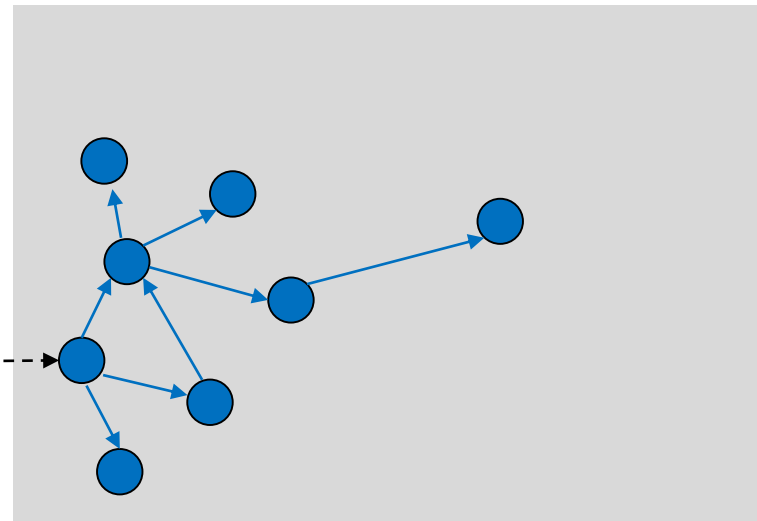
„Běžný“ vs. sémantický web

- Idea **sémantického webu** (T. Berners Lee)
 - propojit nejen dokumenty, ale i **strukturovaná data**
 - opatřit data **strojově čitelným významem**
 - propojit je **typovanými odkazy**
 - požadavek HTTP bude vracet **množinu dat**

„Běžný“ vs. sémantický web

- Idea **sémantického webu** (T. Berners Lee)
 - propojit nejen dokumenty, ale i **strukturovaná data**
 - opatřit data **strojově čitelným významem**
 - propojit je **typovanými odkazy**
 - požadavek HTTP bude vracet **množinu dat**

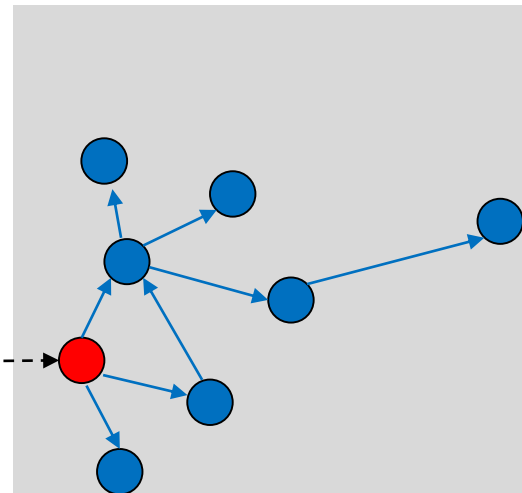
HTTP IRI



„Běžný“ vs. sémantický web

- Idea **sémantického webu** (T. Berners Lee)
 - propojit nejen dokumenty, ale i **strukturovaná data**
 - opatřit data **strojově čitelným významem**
 - propojit je **typovanými odkazy**
 - požadavek HTTP bude vracet **množinu dat**

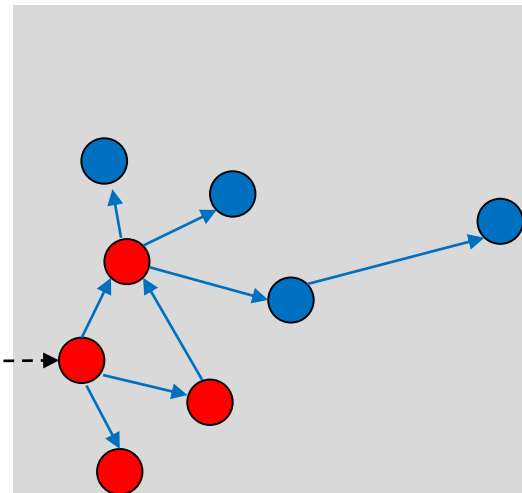
HTTP IRI



„Běžný“ vs. sémantický web

- Idea **sémantického webu** (T. Berners Lee)
 - propojit nejen dokumenty, ale i **strukturovaná data**
 - opatřit data **strojově čitelným významem**
 - propojit je **typovanými odkazy**
 - požadavek HTTP bude vracet **množinu dat**

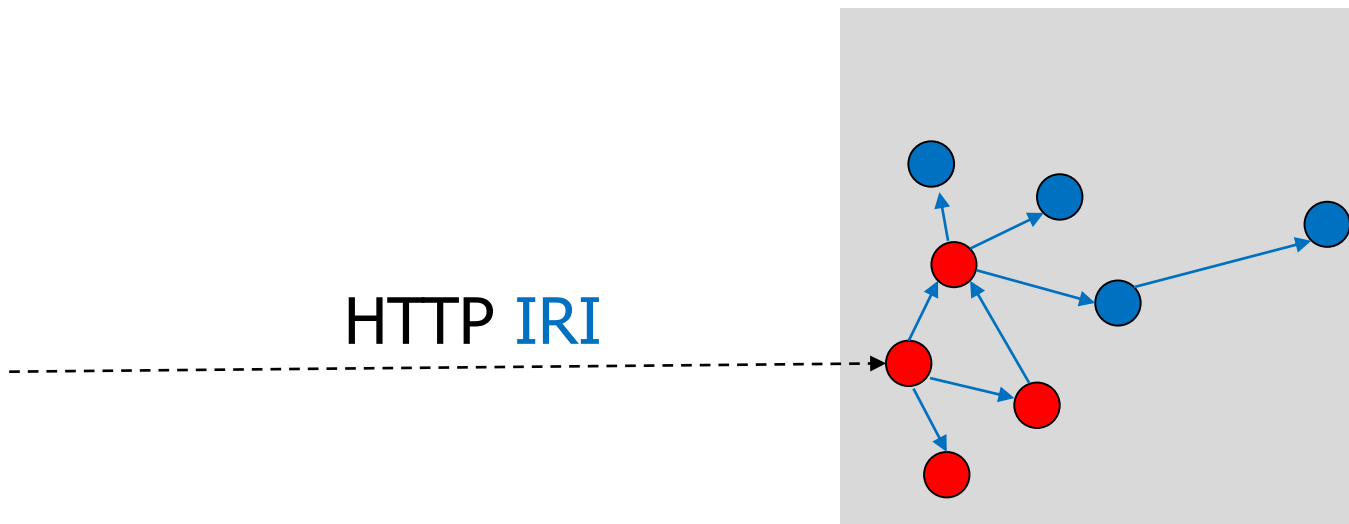
HTTP IRI



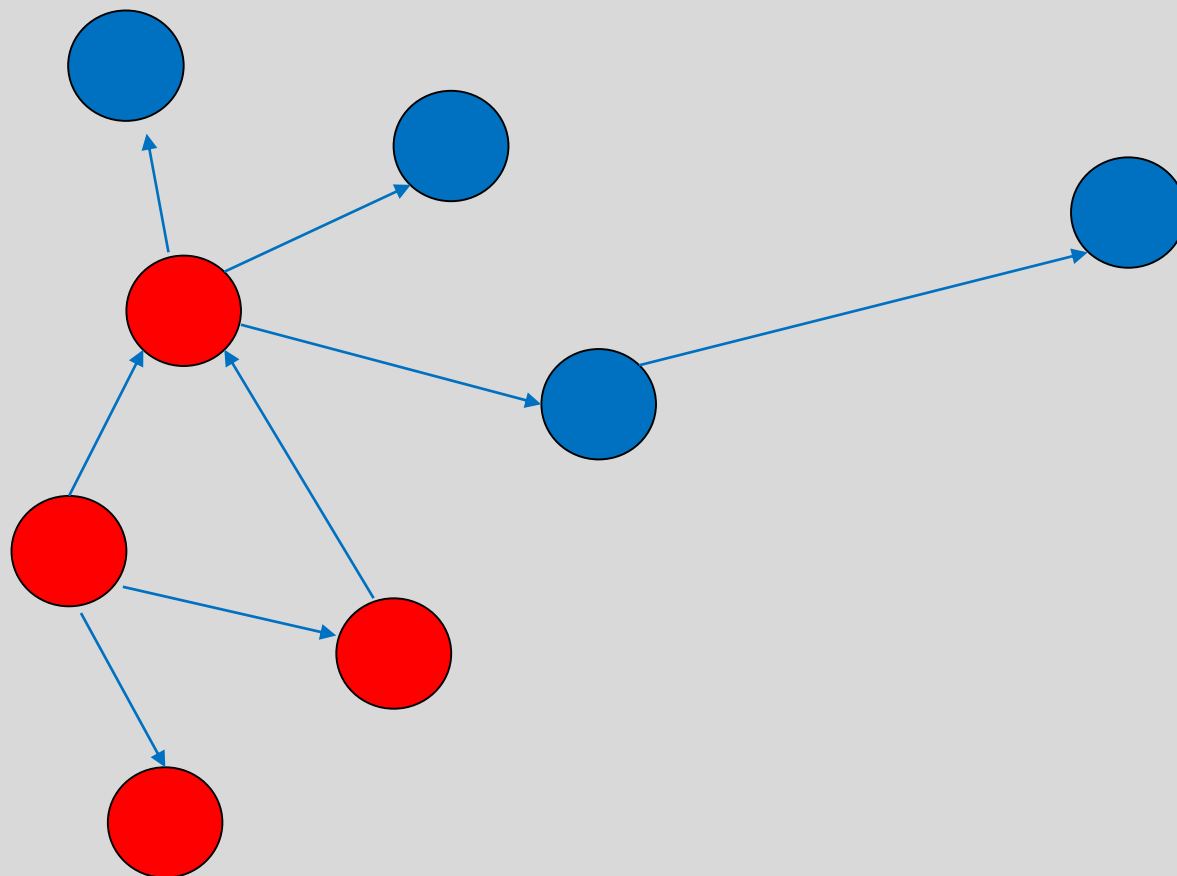


Strojově čitelný význam

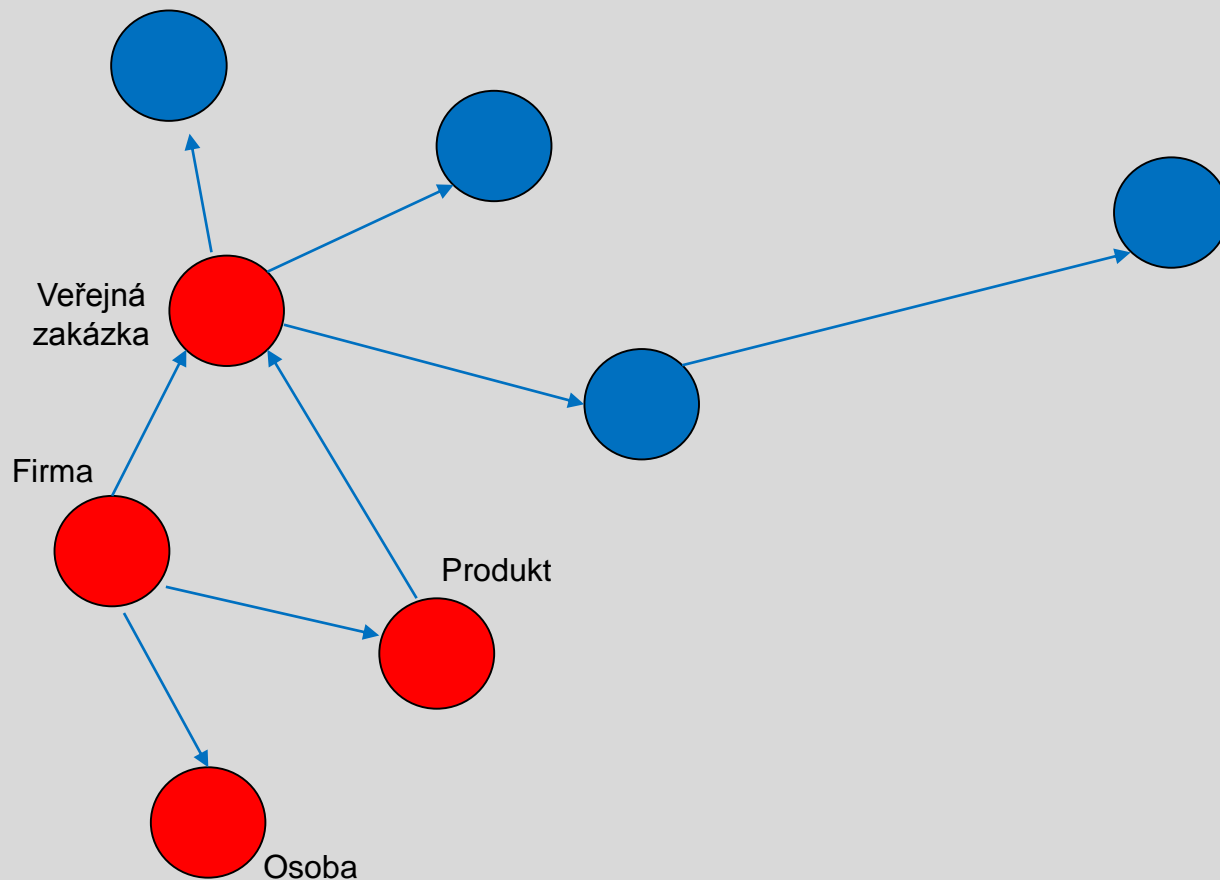
HTTP IRI



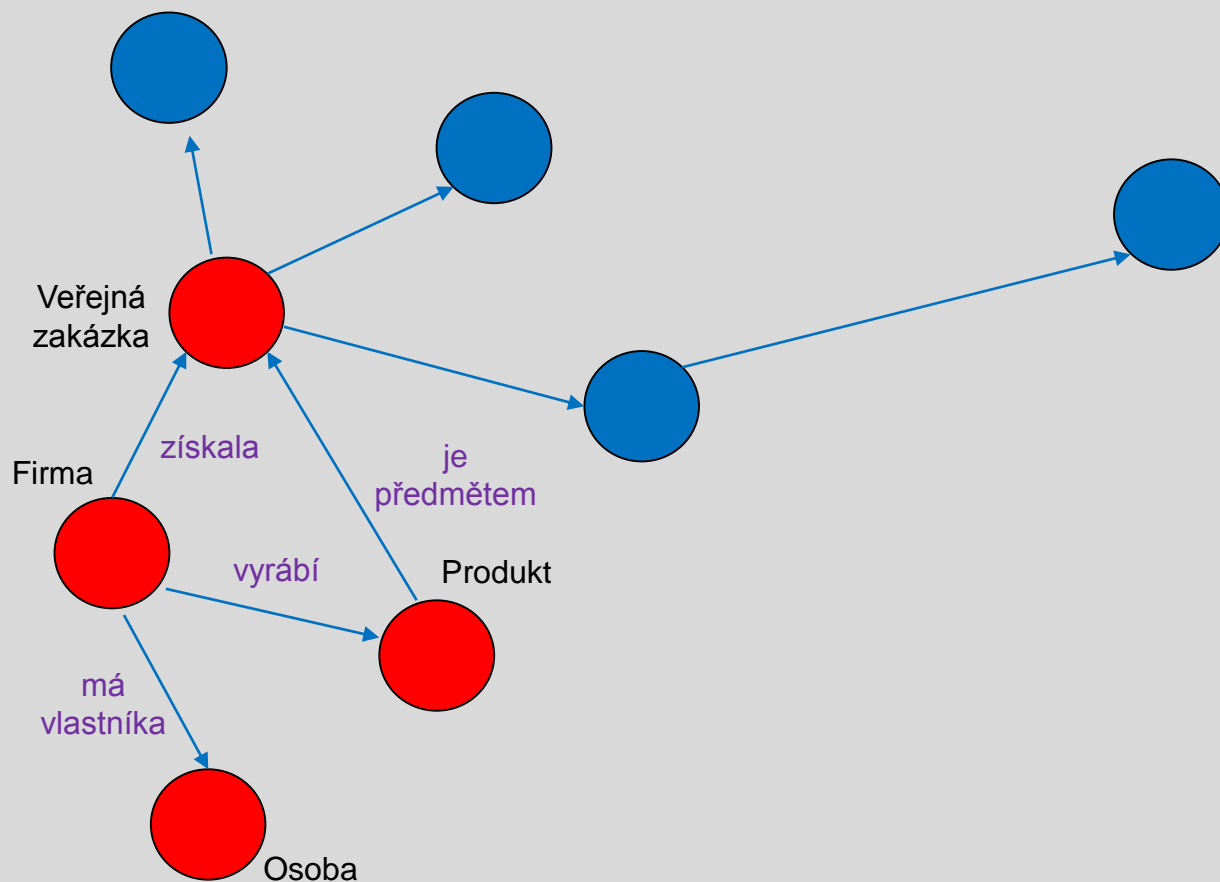
Strojově čitelný význam



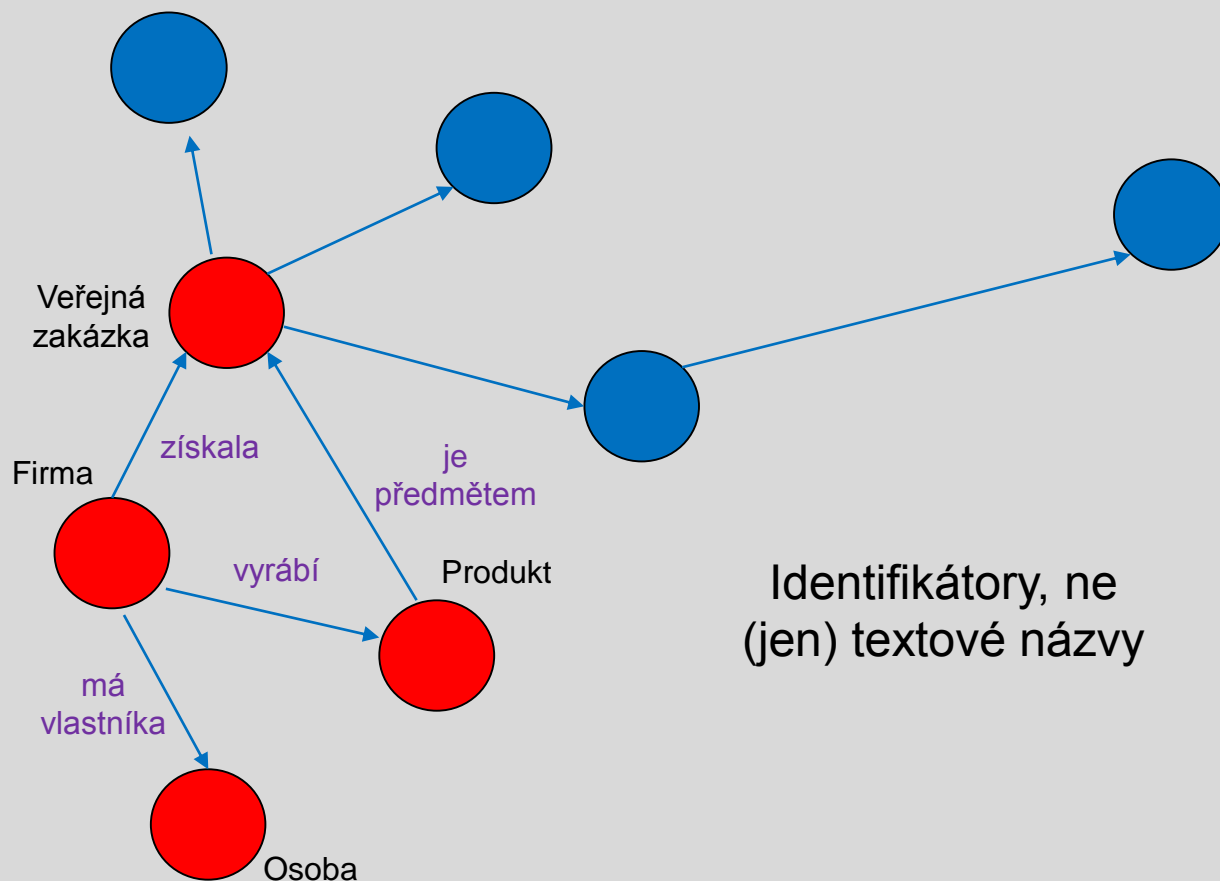
Strojově čitelný význam



Strojově čitelný význam



Strojově čitelný význam





Cíl předmětu

- Naučit se data na sémantickém webu
 - **vytvářet a vystavovat**, včetně
 - návrhu datových schémat a doménových slovníků
 - extrakce dat, která nativně v RDF nejsou
 - **zpracovávat** (čistit, transformovat a propojovat)
 - **využívat** (v jednoduchých aplikacích)



Probíraná témata

- **RDF**: jazyk pro reprezentaci propojitelných dat na webu v obecné grafové struktuře
- **SPARQL**: dotazovací jazyk pro RDF
- **Datové slovníky** (ontologie): popis významu RDF dat existujícími slovníky, i tvorba nových slovníků
- Proces **zpracování dat** („ETL“) vedoucí k jejich vystavení v RDF: extrakce, transformace (včetně čištění a propojování) a výsledné publikování
- Využívání RDF dat v **softwarových aplikacích**
- **Byznys modely** podporující využití RDF dat na webu



Dimenze „uvažování o tématu“

- Životní cyklus dat
 - zdrojová data – RDF schémata a slovníky – ETL –
– dotazování pomocí SPARQL – využití v aplikaci
- Věcné domény dat
 - encyklopedická; statisticko-ekonomická
(jiné: biomedicína, knihovnictví, produkty/služby...)
- Myšlenková prostředí
 - Věda: výzkumné projekty, experimenty a publikace
 - Inženýrství: reprezentační jazyky a SW nástroje
 - Byznys: modely výnosů a nákladů



Co předmět je a není?

- Nejde o desítky let ustálenou disciplínu, pro kterou by existovaly „nepohnutelné“ osnovy (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 projektů a iniciativ...
- „Správné odpovědi“ na mnohé klíčové otázky zatím nikdo na 100% nezná, obor se rok od roku rychle vyvíjí
 - metafora „lesní školky“ – něco vyroste, něco padne a pohnejí
- Vzhledem k tomu je u studentů velmi vítána vlastní iniciativa, zvědavost, zkoumání toho, co se nově objevuje jak na akademických konferencích, tak i např. v blogpostech



Prof. Anja Feldmann

Pražský infromatický seminář, 27.4.2017

The Internet: A fascinating (research) object

Témata přednášky:

detekce výpadků sítě, predikce výkonu distribuovaných systémů, apod.



Prof. Anja Feldmann

Pražský informatický seminář, 27.4.2017

The Internet: A fascinating (research) object

Témata přednášky:

detekce výpadků sítě, predikce výkonu distribuovaných systémů, apod.

Internet, ale i web jako jeho nejviditelnější součást, jsou **fascinujícími** tématy výzkumu.



Prof. Anja Feldmann

Pražský infromatický seminář, 27.4.2017

The Internet: A fascinating (research) object

Témata přednášky:

detekce výpadků sítě, predikce výkonu distribuovaných systémů, apod.

Internet, ale i web jako jeho nejviditelnější součást, jsou fascinujícími tématy výzkumu.

Neplatí to ještě více o spojení **internetu/webu** s možností předávat si přes něj strukturovaná data s popsáním **významem**?



Možnosti uplatnění

- Univerzity
 - Dělá se v nějaké podobě skoro na všech
- Veřejná správa
 - Otevírání datových zdrojů státní správy, měst a obcí
- Výzkumná centra zahraničních korporací
 - 2 noví part-time doktorandi KIZI pro tuto oblast
- SW startupy – domácí, ale zejména zahraniční
 - práce na dálku - Praha je „kompetenčním centrem“ pro PDW
- „Běžné“ (velké) firmy
 - **Ano** search engines, business info; pharma, automotive, ...
 - **Ne** banky nebo SW firmy - spíš „mainstream“ technologie



PDW, přednášející a katedra

- PDW spadá do tématiky pracovní skupiny **SWOE** („Semantic Web and Ontological Engineering“)
 - Jedna z pracovních skupin KIZI, viz <https://kizi.vse.cz/english/research-groups/>
 - Vedoucí V. Svátek a O. Zamazal
 - Skupina zaměřená primárně *vědecky*...
 - ...spíše sekundárně na *spolupráci s praxí* a na (magisterskou) *výuku*; PDW je jediný takto zaměřený předmět povinný pro ZWT



PDW, přednášející a katedra

- **SWOE** v kontextu ostatních pracovních skupin
 - Využití webu jako platformy pro sdílení dat
 - viz webové inženýrství - skupina **WELT**
 - Využití strojového odvozování
 - viz umělá inteligence - skupina **IIS**
 - Příprava „vysokých“ (sémanticky popsaných) a „širokých“ (propojených) dat pro analytické využití
 - viz data mining/science - skupina **DMKD**
- Tj. skupina do značné míry průniková, lze propojit s kteroukoli z ostatních



PDW, přednášející a katedra

- **SWOE** v kontextu ostatních pracovních skupin
 - Využití webu jako platformy pro sdílení dat
 - viz webové inženýrství - skupina **WELT**
 - Využití strojového odvozování
 - viz umělá inteligence - skupina **IIS**
 - Příprava „vysokých“ (sémanticky popsaných) a „širokých“ (propojených) dat pro analytické využití
 - viz data mining/science - skupina **DMKD**
- *Kdo je tady spíš webový vývojář, nadšenec do AI, nebo datový analytik?*



PDW, přednášející a katedra

- (Jen na okraj...) hlavní oblasti odborného zájmu přednášejícího
 - Ontologické inženýrství
 - úzká nika, v ČR převážně jen akademický výzkum
 - skupina SWOE zde patří k širší světové špičce oboru
 - Publikování a využívání propojených dat
 - i v ČR už řadu let populární téma, zvl. ve veřejné správě
 - zapojení do projektů EU, hlavní zdroj financování skupiny
 - **hlavní téma předmětu 4iz440**
 - Data mining (historicky – mj. DP a DisP)
 - Business informatika („zainteresovaný laik“)
 - plyne z pozice garanta doktorského oboru – sledování témat doktorandů všech 4 inženýrských kateder
 - hledání prostoru pro vědecká témata v této oblasti



Související předměty

- 5FI430 – Znalosti a ontologické inženýrství
 - M. Vacura, KFIL, povinný pro KI
 - Problematika ontologických modelů jako bohatší varianty datových slovníků – formální odvozování nad koncepty, filozofické ukotvení atd.
- 4IZ470 – Dolování znalostí z webu
 - V. Svátek, povinný 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, oborově volitelný pro ZWT a KI
 - 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**
 - účastní se i studenti, zvl. v rámci DP
- Neformální výzkumný seminář **KEG**
„Knowledge Engineering Group“
 - některé (1./2.) čtvrtky v semestru 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/>



...weeding the semantic web garden



OPENDATA CZ

Příklady zajímavých DP na SWOE

- Řízení kvality otevřených dat veřejné správy
 - ETL a využívání OD města Děčín
- Zpřístupnění dat o VŠ kvalifikačních pracích z NUŠL v podobě LD
- Ontologie přístupnosti budov
 - ... pro handicapované
- Modeling Events on the Semantic Web
- Využití propojených dat na webu ke tvorbě strategické znalostní hry

