Linked Data a jazyk RDF

prof. Ing. Vojtěch Svátek, Dr.

Zimní semestr 2019

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

Osnova přednášky

- Strukturovaná data na webu - přehled
- Principy Linked Data
- Jazyk RDF
  - Hlavní syntaxe
  - „Slovníkový“ jazyk RDFS, nejpoužívanější slovníky
  - Dotazovací jazyk SPARQL
- Podrobněji http://linkeddatabook.com
  - kap. 1 a 2

Jak využít potenciál webu?

- Vystavit otevřená, strukturovaná, propojená data!
- Současnost:
  - Běžné dokumenty (zvl. psané volným textem) jsou otevřené a propojené, ale nestrukturované
    - HTML sice může dodat určitou strukturu, ale prezentační aspekt se v ní mísí se sémantickým (trochu se to zlepší využitím mikroformátů a zejména mikrodat)
  - Některá data (zvl. veřejné správy) mohou být na webu otevřená, ale jsou nepropojená, a jejich struktura dokonce často není „rozumně“ strojově zpracovatelná (např. v PDF)
  - Webová API poskytují data strukturovaná, ale nepropojená, a otevřená jen v míře, jakou dovoluje omezené rozhraní… *nově jsou populární i grafové jazyky*

Jak strukturovaně reprezentovat data?

- Tabulková reprezentace
  - RDB, XLS, tabulky v HTML apod.
- Stromová reprezentace
  - Hierarchické DB (historicky), XML, OODB?
- Síťová reprezentace
  - Síťové DB (historicky), cizí klíče v RDB, konceptuální grafy (aj. umělá inteligence), grafové DB (např. Neo4J), **RDF**
- *Jsou na sebe (částečně) převoditelné*
- *Nelze paušálně preferovat jednu z nich!*

Tabulky

- + Efektivně indexovatelné
- + Přehledné
- - Rigidní
  - schéma je náročné měnit, data je obtížné slučovat
- - Neúsporné

- *Objekty popsané pevnou, vnitřně nečleněnou množinou vlastností*

Stromy

- Stále ještě přehledné
- + Flexibilnější
- - Hůře indexovatelné

- *Objekty popsané vnitřně členěnou množinou vlastností, některé mohou být nepovinné*
- *Objekty složené z jiných objektů*

Sítě (orientované grafy)

- + Nejflexibilnější, nejbližší reálnému světu
- + Lze snadno slučovat
- - Nepřehledné (obtížná vizualizace)
- - Nejednoznačně zapisovatelné (odkud začít?)
- - Špatně indexovatelné – horší škálovatelnost

- *Vztahy mezi reálnými objekty a vlastnosti těchto objektů (pokud možno vnitřně nečleněné…)*

Které požadavky jsou na webu nejdůležitější?

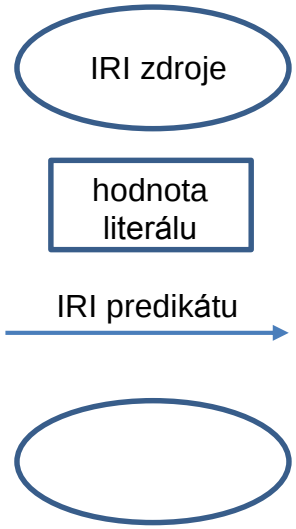
- Flexibilita a možnost data slučovat – web se dynamicky mění!
- Odrážet realitu – různé subjekty a skupiny mohou mít různé konvence, jak zapisovat data, ale realita za tím je (často) stejná
- Škálovatelnost – web je obrovský
  - Ale lze ho rozložit na menší subdomény
- Srozumitelnost, přehlednost – web určen pro lidi (?)
  - Ale k „datovému“ webu budou primárně přistupovat především aplikace, které uživatele od reprezentace odstíní

RDF jako datový model

- Nastudovat z knihy, Kap. 2.4
- Vhodné projít i http://www.w3.org/TR/rdf-primer/
- V předmětu se soustředíme na zápis ve formátu (serializaci) Turtle, ev. grafický zápis
- Česká terminologie:
  - Trojice (triplet): subjekt, predikát, objekt
  - Zdroje s identifikátorem vs. literály; anonymní zdroje (s lokálním identifikátorem, „prázdné uzly“ – blank nodes)

Grafická notace RDF

- Existují variace, v rámci předmětu budeme používat tuto:
  - zdroj s identifikátorem IRI
  - literál
  - predikát
  - zdroj s lokálním identifikátorem (blank node)



Původ „blank nodes“
(informativně)

- Objekt, který nemá ustálený identifikátor
  - „Existenčně kvantifikovaná proměnná“, např.
    něco, co má hodnotu vlastnosti foaf:name rovnu literálu “Ora Lassila”
  - Z hlediska *publikování* dat (ať už veřejně, nebo dlouhodobě v rámci interního repozitáře) je takový pohled překonaný: je lepší zavést nový globální identifikátor než nemít žádný (resp. jen lokální v dokumentu)
  - Smysl mají *interní*, *dočasné* anonymní zdroje – v rámci dotazu SPARQL, nebo v průběhu procesu přípravy dat (ETL) před publikováním
- Shluk hodnot určité vlastnosti objektu
  - Např. cena výrobku (zahrnující hodnotu, měnu, časovou platnost) nebo adresa (zahrnující město, ulici, …)
  - Anonymní zdroj zde zastupuje vnitřní uzel *stromu* hodnot
  - V tomto případě lze použití anonymního zdroje akceptovat – indikuje, že shluk je určen „na jedno použití“ a vydavatel dat nepředpokládá jeho odkazování zvenku (otázka ovšem je, nakolik toto lze předjímat)

Principy Linked Data

- Používat IRI jako globální identifikátory objektů
- Používat IRI dohledatelná (dereferencovatelná) pomocí HTTP
- Na dané adrese poskytnout užitečné informace o objektu („resource descriptions“)
  - Zejména okolí objektu z hlediska grafu RDF – hodnoty datových vlastností a (typované) vztahy k jiným objektům
- Informace o objektu budou zahrnovat i jeho vztahy k jiným objektům
  - Tj. i možnost navigace dál, včetně jiných datasetů

Další technologické postupy

- Negociace obsahu – HTML vs. RDF
  - Pro lidi vs. aplikace, pomocí hlavičky v HTTP
- Odlišení „neinformačních zdrojů“ (reálných, nebo i abstraktních objektů „nevisících“ na webu) a (webových) informačních zdrojů při dohledávání
  - Návratový kód *303 See Other*, nutnost druhého požadavku
  - Identifikátor fragmentu v rámci dokumentu (tzv. „# IRI“)
    - Pro kompaktní zdroje, např. slovníky RDFS/OWL
- „Cool IRIs“ – vše je poznat „na první pohled“
    - Osoba: http://www.example.com/id/alice
    - Homepage: http://www.example.com/people/alice
    - Data v RDF: http://www.example.com/data/alice

Serializace RDF

- Turtle
- RDF/XML
- RDFa
- N-Triples
- JSON-LD
- *Notation3 (N3) – velká vyjadřovací síla, spíš „znalostní“ než datový jazyk*

Serializace RDF

- **Turtle**
  - lidsky čitelné, SPARQL ruční tvorba dat (zvl. slovníků)
- RDF/XML
  - první standardizovaná serializace, vhodná jen pro strojové zpracování (zvl. pokud se používají XML nástroje)
- RDFa
  - možnost zanořování RDF do atributů HTML/XML
  - populární popis (vč. základů RDF, FOAF, atd., s drobnými neaktuálnostmi) viz https://www.youtube.com/watch?v=ldl0m-5zLz4
- **N-Triples**
  - nejjednodušší syntaxe (→ parsing), neúsporné, ale pro přenos lze zkomprimovat
- **JSON-LD**
  - na vzestupu, oblíbený u programátorů, populární úvod viz https://www.youtube.com/watch?v=vioCbTo3C-4

Příklad v Turtle (veřejná zakázka)

deklarace použitých slovníků

```
@prefix dc: <http://purl.org/dc/terms/> .
@prefix pc: <http://purl.org/procurement/public-contracts#> .
…
<http://www.city.com/PC/12345678> a pc:Contract;
  dc:title "Whips for police"@en;
  dc:description "Detailed description of a public contract for police supplies."@en;
  pc:procedureType pc:Open;
  pc:kind pc:Supplies;
  pc:referenceNumber "1234567890";
  pccz:limit pccz:UnderLimit;
  pc:lot <http://www.city.com/PC/12345678-1>;
  pc:attachment <http://www.city.com/PC/12345678.pdf>;
  pc:publicationDate "2012-01-01"^^xsd:date;
  pc:tenderDeadline "2012-01-10"^^xsd:date;
  pc:startDate "2012-03-01"^^xsd:date;
  pc:durationDays "31"^^xsd:positiveInteger;
  pc:estimatedEndDate "2010-03-31"^^xsd:date;
  pc:numberOfTenders "2"^^xsd:nonNegativeInteger;
  pc:awardDate "2012-02-01"^^xsd:date;
  pc:actualEndDate "2012-05-31"^^xsd:date;
  pc:cancellationDate "2012-04-30"^^xsd:date.
```

Příklad v Turtle (veřejná zakázka)

```
@prefix dc: <http://purl.org/dc/terms/> .
@prefix pc: <http://purl.org/procurement/public-contracts#> .
…
<http://www.city.com/PC/12345678> a pc:Contract;
  dc:title "Whips for police"@en;
  dc:description "Detailed description of a public contract for police supplies."@en;
  pc:procedureType pc:Open;
  pc:kind pc:Supplies;
  pc:referenceNumber "1234567890";
  pccz:limit pccz:UnderLimit;
  pc:lot <http://www.city.com/PC/1234
  pc:attachment <http://www.city.com/PC/12345678.pdf>;
  pc:publicationDate "2012-01-01"^^xsd:date;
  pc:tenderDeadline "2012-01-10"^^xsd:date;
     03-01"^^xsd:date;
     l"^^xsd:positiveInteger;
  pc:estimatedEndDate "2010-03-31"^^xsd:date;
  pc:numberOfTenders "2"^^xsd:nonNegativeInteger;
  pc:awardDate "2012-02-01"^^xsd:date;
  pc:actualEndDate "2012-05-31"^^xsd:date;
  pc:cancellationDate "2012-04-30"^^xsd:date.
```

Slovníková třída

Literály s jazykovým tagem

Slovníkové instance nepřímo vyjadřující typ

Literál interpretovaný defaultně jako xsd:string

Slovníková vlastnost + instance z jiného jmenného prostoru (deklarace prefixu zde vynechána) – přídatný modul slovníku pro české prostředí

Vztah mezi objekty (zakázka k podzakázce)

IRI (tj. URL) „informačního zdroje“ vystaveného na webu, dereferencuje se přímo

Slovníky (schémata, ontologie) RDF

- Definují vlastnosti (predikáty), třídy zdrojů a vyčleněná individua pro určitou tématickou oblast
- Příklady populárních slovníků používaných *napříč* datasety
  - **Dublin Core** – vlastnosti dokumentů
  - **FOAF** – vlastnosti osob a jejich vzájemné vztahy
  - SIOC – zapojení lidí do online komunit
  - DOAP – softwarové projekty
  - GoodRelations – e-commerce
  - **SKOS** – terminologické stromy
  - **schema.org** – „všezahrnující“ slovník původně vyvinutý pro značkování webových stránek s ohledem na vyhledávače
- Zdroj se přiřadí ke třídě pomocí predikátu rdf:type (v Turtle „a“)

Dotazování do RDF

- Jazyk SPARQL
  - Sada standardů W3C přístupných z https://www.w3.org/TR/sparql11-overview/
    - Dotazovací jazyk
    - Manipulační jazyk (SPARQL UPDATE)
    - Sada výstupních formátů (XML, JSON, CSV, TSV)
    - … a další zdroje
  - Ve své „grafové“ části využívá Turtle

Různé služby

- Validace Turtle (a SPARQL): http://sparql.org/
- Dohledání URL pro obvyklé prefixy: http://prefix.cc

JSON-LD

- Lze vytvářet mj. pomocí editoru https://editorsnotes.github.io/edit-with-lov/