

Linked Data a jazyk RDF

Doc. Ing. Vojtěch Svátek, Dr.

Zimní semestr 2012

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

Osnova přednášky

- Principy Linked Data
- Jazyk RDF
 - Hlavní syntaxe
 - „Slovníkový“ jazyk RDFS, nejpoužívanější slovníky
 - RDFa - RDF vnořené v HTML
- 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 často není strojově zpracovatelná (např. v PDF)
 - Webová API poskytují data strukturovaná, ale nepropojená, a otevřená jen v míře, jakou dovoluje omezené rozhraní

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), **RDF**
- *Jsou na sebe (částečně) převoditelné*
- *Nelze paušálně preferovat jednu z nich!*

Tabulky

- + Efektivně indexovatelné
- + Přehledné
- - Rigidní
- - 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** – web je určen pro lidi (?)
 - Ale k datovému webu budou primárně přistupovat především aplikace, které uživatele od reprezentace odstíní

HTML, XML, RDF a JSON na webu?

- Přednáška J. Tonnison na XML Prague 2012
http://www.xmlprague.cz/2012/files/video-archive-1.html?ps_idc=0&ps_ida=453&ps_idb=2761
- Mohou se navzájem doplňovat
 - Viz aplikace na <http://legislation.gov.uk>
- Neexistuje ještě mnoho zobecněných zkušeností, natož standardů pro takové komplementární využití

RDF jako datový model

- Nastudovat z knihy, Kap. 2.4
- Vhodné projít i <http://www.w3.org/TR/rdf-primer/>
- Česká terminologie:
 - Trojice: subjekt, predikát, objekt
 - Zdroje vs. literály; anonymní zdroje (prázdné uzly)

Původ anonymních zdrojů

- Objekt, který nemá ustálený identifikátor
 - „Existenčně kvantifikovaná proměnná“, např.
něco, co má hodnotu vlastnosti foaf:name rovnu literálu “Ora Lassila”
 - překonané pojetí – lepší zavést nový globální identifikátor než nemít žádný (resp. jen lokální v dokumentu)
- 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
 - Nelze eliminovat, pokud nechceme přejít k hybridní reprezentaci (zatím pro tento účel zřejmě nevznikla...)

Principy Linked Data

- Používat URI jako globální identifikátory objektů
- Používat URI 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 souhrnného dokumentu
 - Pro kompaktní zdroje, např. slovníky RDFS/OWL
- „Cool URIs“ – 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/XML
- RDFa
 - srovnání s HTML5 mikrodaty viz <http://www.jenitennison.com/blog/node/124>
- Turtle
- N-Triples
- RDF/JSON
- *Notation3 (N3) – velká vyjadřovací síla, spíš „znalostní“ než datový jazyk*

Příklad v Turtle (veřejná zakázka)

@prefix dc: <http://purl.org/dc/terms/> .

@prefix pc: <http://purl.org/procurement/public-contracts#> .

deklarace použitých
slovníků

...

<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.

Slovníky (schémata, ontologie) RDF

- Definují **vlastnosti** (predikáty), a **třídy zdrojů** a vyčleněná **individua** pro určitou tématickou oblast
- Příklady populárních slovníků
 - 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
- Zdroj se přiřadí ke třídě pomocí predikátu **rdf:type** (v Turtle „a“)