Sémantický web a ontologie

Propojená data a popis jejich sémantiky

doc. Ing. Radek Burget, Ph.D.
burgetr@fit.vutbr.cz

Sémantický web

- Představa *webu dat* (*web of data* oproti *web of documents*)
  - Publikování strojově srozumitelných dat
- Základní prvky:
  - Reprezentace znalostí – Resource description framework (RDF)
  - Sdílená konceptualizace („model světa“) – ontologie
  - Agenti – producenti a konzumenti služeb

Berners-Lee, Tim, Hendler, James, et al. “The Semantic Web : a new form of Web content that is meaningful to computers will unleash a revolution of new possibilities” *Scientific American* 284 (2001) : 34-43

Technické řešení

- Vývoj technologií pro vhodnou reprezentaci dat
  - Možnost sdílení dat i s jejich sémantikou
  - Použitelné technologie jsou již dlouho k dispozici
- Integrace s existujícím webem
  - Anotace ve webových stránkách
  - Poněkud vázne, ale zlepšuje se

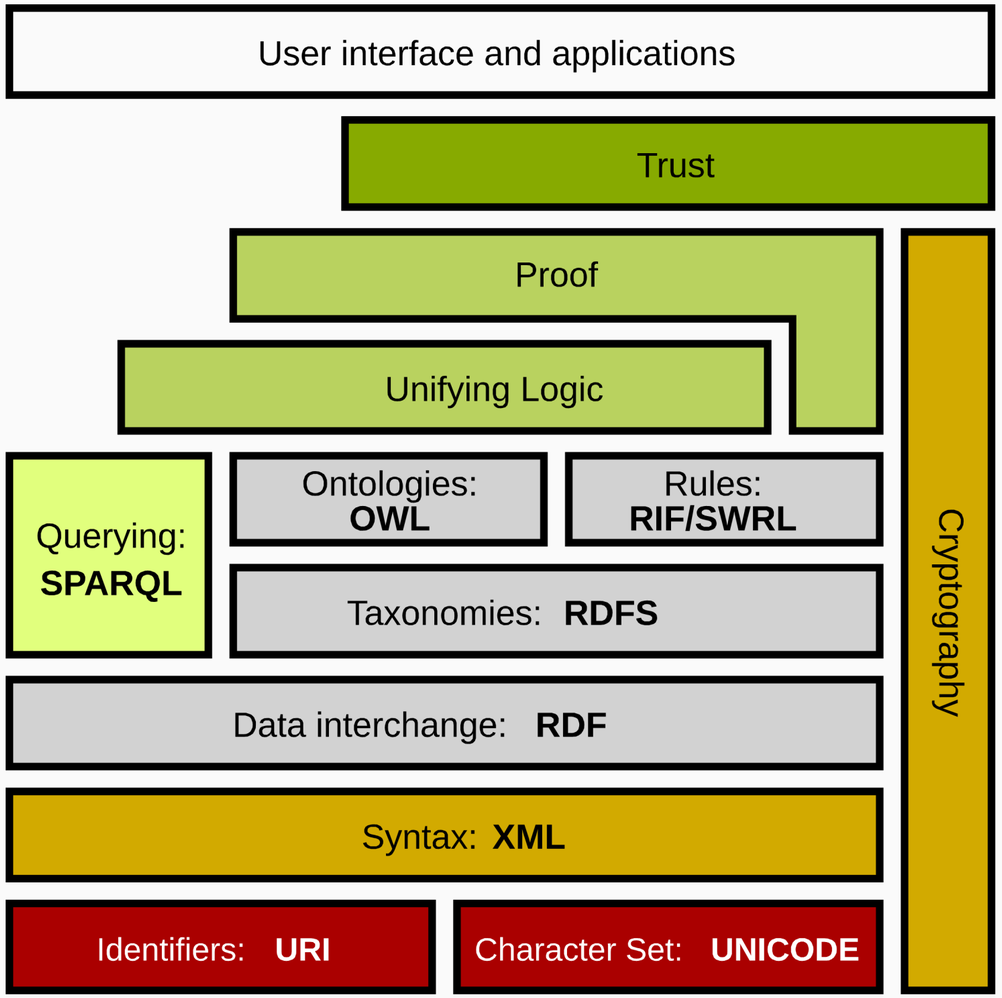
Web a sémantický web

- World Wide Web (web)
  - Základní jednotkou je dokument
  - „Web of documents“
- Semantic Web (sémantický web)
  - Základními jednotkami jsou data
  - „Web of Data“, „Linked data“

Technologie sémantického webu

- Technologie standardního webu
  - HTTP, URI
- Nástroje pro reprezentaci znalostí
  - Reprezentace dat (faktů)
    - XML, RDF, ...
  - Sémantika
    - Ontologie
    - Technologie pro reprezentaci ontologie

Semantic Web Stack



Datový model – RDF

Reprezentace a výměna faktů v rámci sémantického webu

Cíle a prostředky

- Cíle
  - Reprezentace strukturovaných dat a jejich významu (sémantiky)
  - Možnost sdílet data a jejich sémantiku napříč aplikacemi
- Běžná reprezentace dat v IS:
  - Relační/objektové/NoSQL databáze – vázané na aplikaci
  - Veřejné API + serializace (JSON, XML) – není definována sémantika

Serializace – příklad

```
<nabidka>
    <polozka>
        <velikost>3+1</velikost>
        <lokalita>Brno-střed</lokalita>
        <cena mena="czk">2 200 000</cena>
    </polozka>
    <polozka>
        <velikost>2+1</velikost>
        <lokalita>Kuřim</lokalita>
        <cena mena="czk">450 000</cena>
    </polozka>
</nabidka>
```

Problémy

- Význam elementů je specifický pro danou aplikaci
  - Je definován v programovém kódu, který generuje nebo načítá serializovaná data
  - Obdobně jako např. sloupce v relační databázi
- Jiná aplikace může stejným elementům přiřadit jiný význam
  - Např. `<velikost>2+1<velikost>` vs. `<velikost>55m2</velikost>`
- Data jsou strojově čitelná (machine readable), ale ne srozumitelná (machine understandable)

Reprezentace sémantiky

- Odlišení značek v různých aplikacích
  - Např. XML namespaces
  - Řeší kolize značek – syntaktický problém
- Oddělená definice významu značek
  - Např. doprovodný dokument vysvětlující význam a případy použití
- Navíc ale potřebujeme definovat sémantické vztahy
  - Např. byt je věc, která má umístění, velikost a cenu
  - Pokud možno formálně => **Ontologie**

Reprezentace faktů: RDF

- RDF: Resource Description Framework
  - Umožňuje reprezentovat elementární *tvrzení* reprezentující data (fakta)
- Grafová struktura
  - Jednotlivá tvrzení jsou propojena pomocí URI, tvoří orientovaný graf (uzly, hrany)
- Serializace (uložení do souboru, přenos)
  - Lze zapsat pomocí XML nebo jiných jazyků

RDF trojice

- Základním prvkem je **RDF trojice**

subjekt – predikát – objekt

- Základní *tvrzení* (*statement*)

RDF trojice – tvrzení (statement)

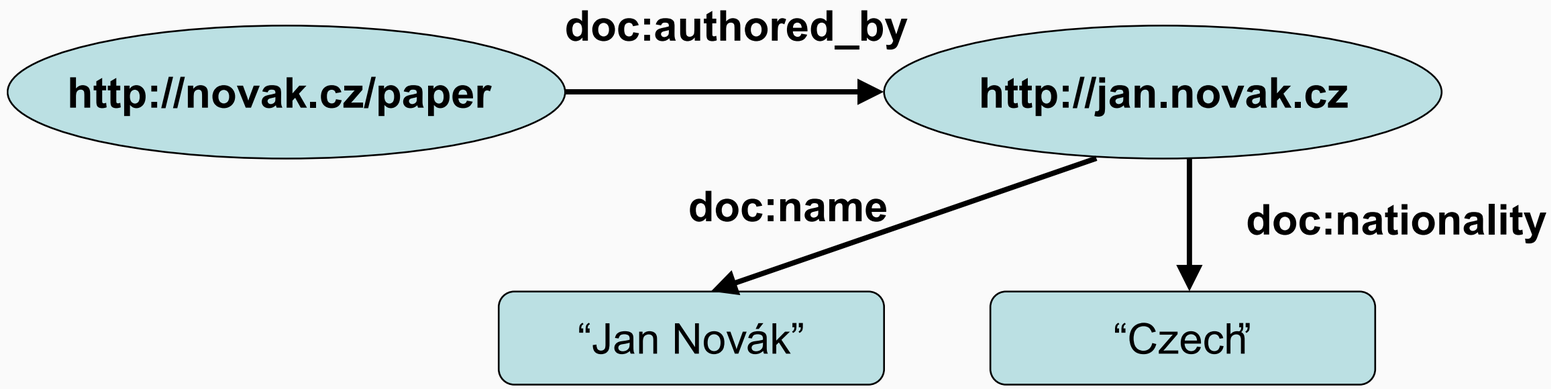
- *Autorem* **dokumentu X** je **pan Y**
  - Subjekt: **dokument X**
  - Predikát: *je autorem*
  - Objekt: *pan Y*
- Jednotlivé části tvrzení (zdroje) *(resources)* jsou reprezentované pomocí **URI** nebo **literálem** (pouze objekt).

RDF tvrzení (II)

Kde vzít URI?

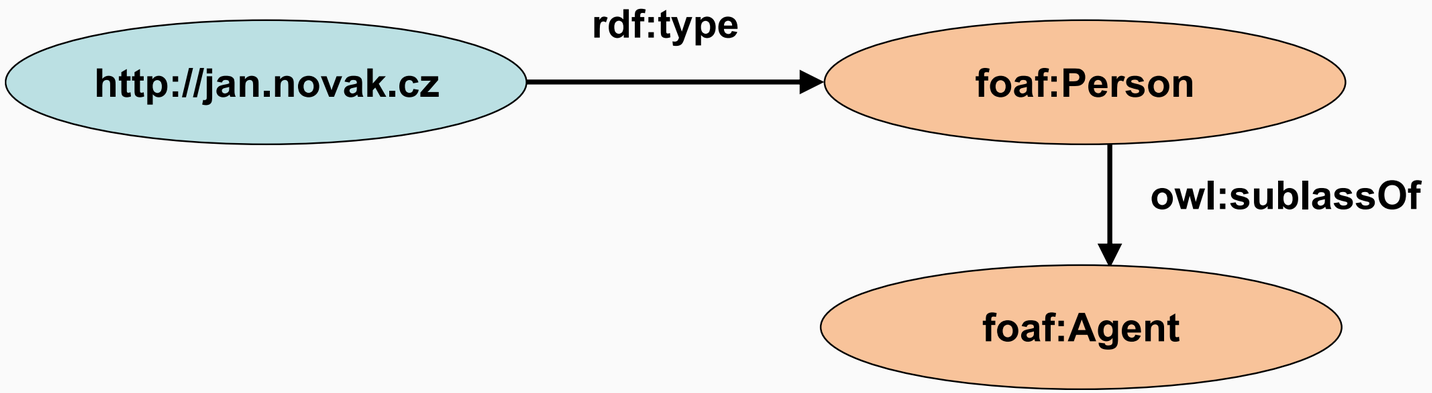
- Vlastní data - vlastní URI
  - Např. `http://fit.vut.cz/student/938272`
  - Často společný *prefix*
- Existující data - např. veřejné znalostní báze
  - `http://dbpedia.org/resource/Berlin`
- Strukturované slovníky - ontologie
  - URI pro predikáty, typy (třídy) objektů (Person, Event, ...)
- Zabudované
  - `rdf:type`

RDF Graf



- RDF graf lze rozložit na trojice subjekt – predikát – objekt
- Subjekt a predikát jsou vždy **URI**
  - `doc:` je prefix URI, který se expanduje
  - Např. `doc:name` => `http://my.docs.com/#name`
- Objekt je **URI** nebo **literál** (různých datových typů)

Schéma – Ontologie



- RDF data lze propojit s metadaty (ontologií, schématem)
  - Pomocí predikátu `rdf:type` (`http://www.w3.org/1999/02/22-rdf-syntax-ns#type`)
- Definice metadat opět pomocí RDF
  - Je možné (ale ne nutné) spojit data i metadata do jednoho grafu.

Ukládání a přenos RDF dat

- Uložení do RDF úložiště (např. RDF4J)
  - Rozložení na trojice a uložení do interní struktury
  - Následně možnost dotazování (jazyk SPARQL)
- Serializace do souboru a zpět – několik variant
  - RDF/XML (standard W3C)
  - N-triples (N3)
  - Turtle (podmnožina N3)

Serializace do Turtle

```
@prefix doc: <http://dokumenty.cz/def#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .

<http://novak.cz/clanek>
    doc:authored-by <http://jan.novak.cz> .

<http://jan.novak.cz>
    doc:name "Jan Novák" ;
    doc:nationality "česká" ;
    a foaf:Person .
```

XML Serializace

```
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:doc="http://dokumenty.cz/def\#">

    <rdf:Description rdf:about="http://novak.cz/clanek">
        <doc:authored-by
            rdf:resource="http://jan.novak.cz" />
      </rdf:Description>

      <rdf:Description rdf:about="http://jan.novak.cz">
        <doc:name>Jan Novák</doc:name>
        <doc:nationality>česká</doc:nationality>
```

RDF jako databáze

- Repozitář – úložiště RDF trojic
- Dotazování – jazyk SPARQL
- Lokální úložiště (triplestore):
  - Virtuoso http://virtuoso.openlinksw.com/
  - RDF4J (dříve Sesame) http://rdf4j.org/
  - …
- Globální *znalostní báze* (*knowledge base*)
  - DBPedia http://dbpedia.org
  - WikiData https://www.wikidata.org/
  - …

Dotazování – SPARQL

- Výsledkem dotazu je
  - CSV (tabulka) – dotaz SELECT
  - Nebo nový graf – dotaz CONSTRUCT

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX yago: <http://dbpedia.org/class/yago/>
PREFIX dbpedia-owl: <http://dbpedia.org/ontology/>
PREFIX dbprop: <http://dbpedia.org/property/>
SELECT ?place ?name ?label WHERE {
    ?place rdf:type dbpedia-owl:Country .
    ?place dbprop:commonName ?name .
    ?place rdfs:label ?label .
   OPTIONAL {?place dbprop:yearEnd ?yearEnd}
   FILTER (!bound(?yearEnd))
}
```

Veřejné znalostní báze

- DBPedia http://dbpedia.org
  - http://dbpedia.org/resource/Berlin
  - http://dbpedia.org/sparql
- Wikidata http://wikidata.org
  - http://wikidata.org/entity/Q42
- Mnoho dalších
  - Mohou být vzájemně propojené pomocí URI
  - *Linked open data*
  - http://lod-cloud.net/

Otevřená data

- Serializované RDF jako prostředek pro publikování otevřených (propojených) dat
- Např. RDF datasety na data.europa.eu
- Možno importovat do lokálního RDF úložiště
  - Případně spolu s jinými propojenými datasety
  - Nasledně dotazování pomocí SPARQL
- Příp. veřejný SPARQL endpoint
  - Např. https://data.europa.eu/en/about/sparql

Ontologie

Slovníky pro sémantický web

Pojem ontologie

- Původně obecnější význam (filozofie)
- Nástroj pro sdílení významu pojmů, které se vyskytují v cílové oblasti
- *„Formální, explicitní specifikace sdílené konceptualizace“*
- Definují základní pojmy modelovaného světa a vztahy mezi nimi
- **Sdílené** a **opakovatelně použitelné**

Účel ontologií

- Porozumění mezi lidmi (experty)
- **Porozumění mezi počítačovými aplikacemi**
  - **Dodání významu jednotlivým URI v sémantickém webu**
  - Možnost **integrace** dat z různých zdrojů
- Návrh znalostních aplikací

Typy ontologií

- Terminologické (lexikální)
  - Pojmy a jejich vzájemné vztahy (taxonomie)
  - Např. *WordNet*
- Generické ontologie
  - Zákonitosti a vztahy mezi obecnými pojmy
  - „Upper ontology“, např. SUMO
- Doménové ontologie
  - Konkrétní oblast (např. podnikové, lékařství, ...)
- Aplikační ontologie
  - Pro konkrétní aplikaci

Prvky ontologií

- **Třídy (koncepty)**
- **Individua (objekty, instance)**
- **Vlastnosti (role, atributy)**
- Meta-sloty (facety)
- Primitivní datové typy
- Axiomy (pravidla)

Definované prvky můžeme využít v RDF tvrzeních. Ontologie tedy definuje *slovní zásobu* (*vocabulary*) pro RDF.

Koncepty – třídy

- Množiny konkrétních objektů
- Žádné procedurální metody
- Třídy *definované* a *primitivní*
  - Podle definice příslušnosti individua
- Dědičnost tříd (často vícenásobná)

Individua – objekty – instance

- Konkrétní objekty reálného světa
- Individuum nemusí být nutně instancí třídy
- Vzhledem k určení ontologií se často nepoužívají
  - Reprezentují konkrétní data

Relace – atributy – sloty – vlastnosti

- Pojetí vlastnosti je jiné, než u OO modelování
- Vlastnost = relace
  - Samostatně definovaný prvek
  - Obvykle binární relace
- Možná dědičnost relací (má otce, má předka)
  - Nadřazená relace obsahuje všechny prvky podřazené relace
- Funkce – speciální relace
  - Hodnota argumentu *n* jednoznačně určena předchozími *n-1* argumenty

Primitivní hodnoty, datové typy

- Argumentem relace může být *primitivní hodnota* (ne objekt)
  - Číslo, řetězec, výčtová hodnota, …
  - *Datatypová vlastnost* vs. *objektová vlastnost*
- Můžeme uvažovat dato-typové třídy (datové typy) a dato-typové instance (hodnoty)
- Dato-typové vlastnosti obvykle deklarujeme jako funkční (mají pouze jednu hodnotu)

Ontologické jazyky

RDF Schema, OWL

RDF Schema

- Sémantické rozšíření RDF
  - V podstatě (meta) **ontologie**
- Umožňuje definici
  - Tříd
  - Binární relace (definiční obor, obor hodnot)
  - Hierarchie nad třídami i relacemi
- Definice opět pomocí RDF tvrzení (trojic)
  - S použitím konceptů a vlastností z RDFS
- Namespace (prefix obvykle **rdfs**)
  `http://www.w3.org/2000/01/rdf-schema#`

Třídy

- Třída je přiřazena ke zdroji pomocí rdf:type
  - `skola:Osoba rdf:type rdfs:Class`
- Odvozené třídy
  - Např. `skola:Student rdfs:subClassOf skola:Osoba`

Vlastnosti v RDFS

- Vlastnosti jsou instance `rdfs:Property`
  - `skola:maZapsano rdf:type rdfs:Property`
- **rdfs:Range** – typ objektů (obor hodnot)
  - `skola:maZapsano rdfs:range skola:Predmet`
- **rdfs:Domain** – typ subjektů (def. obor)
  - `skola:maZapsano rdfs:domain skola:Student`
- `rdfs:subPropertyOf`
  - Vlastnost je „podvlastností“ jiné vlastnosti

OWL

- Rozšíření RDFS o pokročilé vlastnosti
- Definice kompletní ontologie
- Namespace `http://www.w3.org/2002/07/owl#`

Definice tříd v OWL

- Kombinace s RDFS
- Třídu lze definovat pomocí logických podmínek
  - Identifikátorem třídy (žádné prvky)
  - Výčtem prvků (instancí)
  - Omezením vlastností
  - Sjednocením nebo průnikem dvou a více tříd
  - Doplňkem

Definice třídy identifikátorem

Turtle

```
foaf:Person rdf:type owl:Class .
```

XML

```
<owl:Class rdf:about="&foaf;Person"/>
```

nebo

```
<rdf:Description rdf:ID=“Person”>
    <rdf:type resource=“&owl;Class” />
</rdf:Description>
```

V Turtle s prefixy

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix foaf:<http://xmlns.com/foaf/0.1/>.

foaf:Person rdf:type owl:Class .
foaf:Person a owl:Class .
```

Definice doplňkem

```
<owl:Class>
  <owl:complementOf>
    <owl:Class rdf:about="#Student"/>
  <owl:complementOf/>
</owl:Class>
```

Ostatní operátory nad třídami

- `owl:equivalentClass`
  - Stejná třída (např. z jiné ontologie)
- `owl:disjointWith`
  - Disjunktní třída

Definice vlastností

- RDFS konstruktory

```
<owl:ObjectProperty rdf:ID="studuje">
  <rdfs:domain rdf:resource="#Student"/>
  <rdfs:range rdf:resource="#Obor"/>
</owl:ObjectProperty>
```

- Vztahy mezi vlastnostmi
  - `owl:equivalentProperty` – stejné hodnoty
  - `owl:inverseOf` – inverzní vlastnost

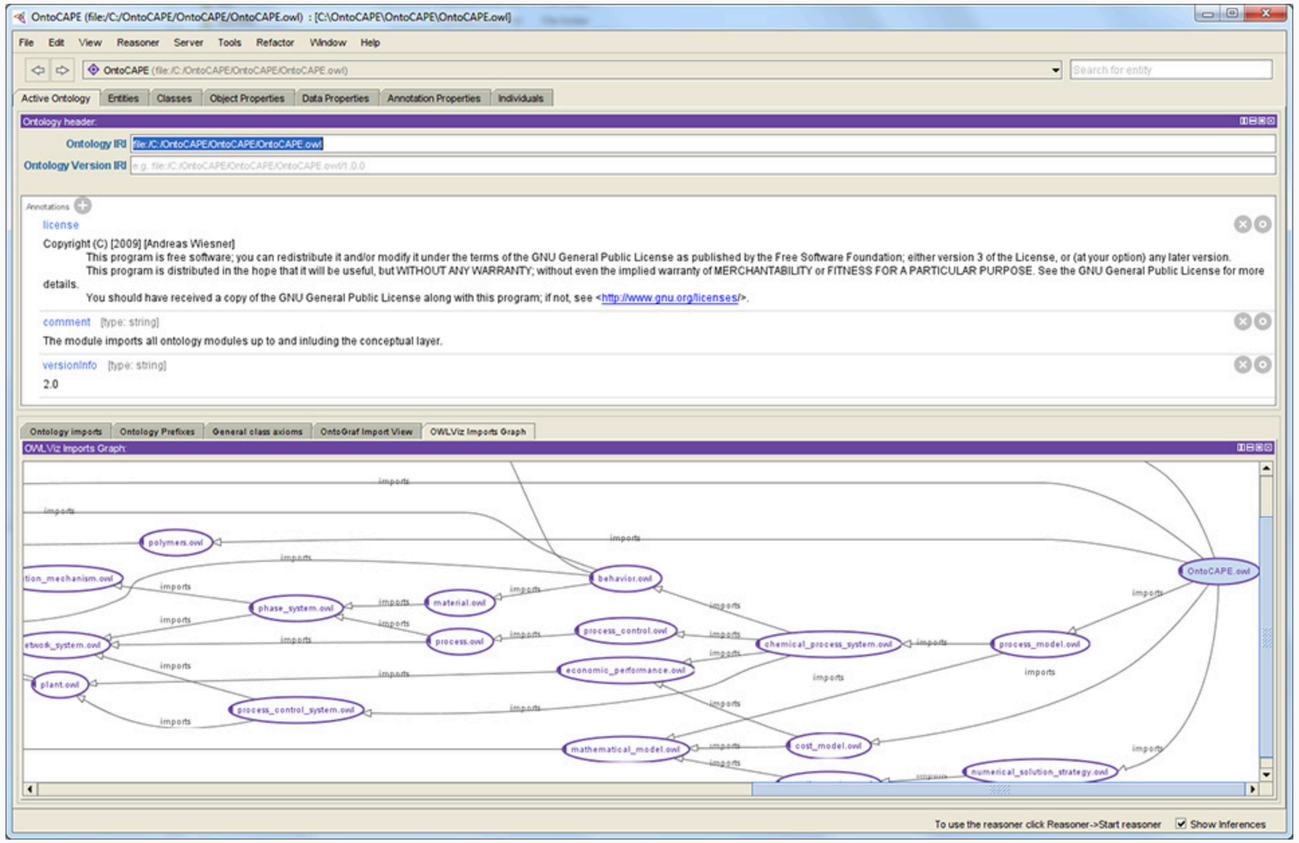
```
<owl:ObjectProperty rdf:ID="maStudenta">
    <owl:inverseOf rdf:resource="#studuje"/>
</owl:ObjectProperty>
```

Data-typové vlastnosti

- RDF Literály
- XSD datové typy
  - Namespace `http://www.w3.org/2001/XMLSchema`
- `xsd:string`, `xsd:normalizedString`, `xsd:boolean`, `xsd:decimal`, `xsd:float`, `xsd:double`, `xsd:integer`, `xsd:nonNegativeInteger` `xsd:positiveInteger`, `xsd:nonPositiveInteger`, `xsd:negativeInteger`, `xsd:long`, `xsd:int`, `xsd:short`, `xsd:byte`, `xsd:unsignedLong`, `xsd:unsignedInt`, `xsd:unsignedShort`, `xsd:unsignedByte`, `xsd:hexBinary`, `xsd:base64Binary`, `xsd:dateTime`, `xsd:time`, `xsd:date`, `xsd:gYearMonth`, `xsd:gYear`, `xsd:gMonthDay`, `xsd:gDay`, `xsd:gMonth`, `xsd:anyURI`, `xsd:token`, `xsd:language`, `xsd:NMTOKEN`, `xsd:Name`, `xsd:NCName`

Editor Protegé

http://protege.stanford.edu/



Existující ontologie

- Důraz na maximální využití existujících ontologií
  - Je možno kombinovat koncepty a vlastnosti z různých ontologií
- Přehled
  - https://lov.linkeddata.es/dataset/lov/

Dublin core

- Metadata dokumentů
- Použití zejména v knihovnictví
- Definuje vlastnosti dokumentů:

```
<rdf:Description rdf:about="http://www.w3schools.com">
   <dc:description>W3Schools</dc:description>
   <dc:publisher>Refsnes Data as</dc:publisher>
   <dc:date>2008-09-01</dc:date>
   <dc:type>Web Development</dc:type>
   <dc:format>text/html</dc:format>
   <dc:language>en</dc:language>
 </rdf:Description>
```

Friend-of-a-friend (FOAF)

- Ontologie pro popis osob a jejich vzájemných vztahů http://www.foaf-project.org/
- Třídy pro popis osob
  - `foaf:Agent`, `foaf:Person`, …
- Vlastnosti
  - `foaf:name`, `foaf:knows`, …

FOAF příklad

```
@prefix foaf:<http://xmlns.com/foaf/0.1/>.
@prefix dbr:<http://dbpedia.org/resource>.

dbr:Luke_Skywalker foaf:knows dbr:Han_Solo .
dbr:Luke_Skywalker foaf:name "Luke Skywalker" .
```

SKOS

- Simple Knowledge Organization System
- Umožňuje organizaci pojmů v nějaké doméně
  - Koncepty: Concept
  - Vztahy mezi nimi: broader, narrower, related, ...
  - ...

Schema.org

- Primárně pro anotování webových stránek
  - https://schema.org
- Základní slovníky pro různé obecné domény
  - https://schema.org/docs/gs.html#schemaorg

Další ontologie

- Music ontology
  - http://musicontology.com/
- Event ontology
  - http://motools.sourceforge.net/event/event.html
- Time ontology
  - http://www.w3.org/TR/2006/WD-owl-time-20060927/
- Geo ontology
  - http://www.w3.org/2003/01/geo/

Ontologie a RDF znalostní báze

- DBPedia.org
  - Vlastní ontologie + použití existujících
  - http://dbpedia.org/resource/Berlin
  - http://dbpedia.org/page/Novak_Djokovic
- Např.
  - Vlastnost Birth place
  - Podobně Wikidata

RDF na Webu

Web of Documents vs. Web of Data

Sémantické Anotace

- Propojení HTML a konceptů sémantického webu (URI)
- Několik existujících standardů
  - RDFa, HTML5 Microdata, JSON-LD
- Common crawl corpus
  - 2020: 50% (z 3.4 miliard) stránek, 44.3% zpracovaných domén
  - 2019: 37.9% (z 2.45 miliard) stránek, 37.2% zpracovaných domén

```
<div itemscope itemtype="https://schema.org/Person">
    <span itemprop="name">Jane Doe</span>
    <span itemprop="jobTitle">Professor</span>
    <div itemprop="address" itemscope itemtype="https://schema.org/P
        <span itemprop="streetAddress">
          20341 Whitworth Institute
          405 N. Whitworth
        </span>
        <span itemprop="addressLocality">Seattle</span>,
        <span itemprop="addressRegion">WA</span>
        <span itemprop="postalCode">98052</span>
    </div>
</div>
```

Bizer, C.; Meusel, R.; Primpeli, A.: Web Data Commons - RDFa, Microdata, and Microformat Data Sets - Extraction Results from the September 2021 Common Crawl Corpus.

Integrace RDF a HTML

- HTML 5 – *Microdata*
- W3C standard – *RDFa*
- JSON notace – *JSON-LD*
- Viz např. https://schema.org/Person#examples

Jiný příklad – RDFa

```
<p xmlns:dc="http://purl.org/dc/elements/1.1/"
   about="http://www.example.com/books/wikinomics">

  In his latest book
  <cite property="dc:title">Wikinomics</cite>,
  <span property="dc:creator">Don Tapscott</span>
  explains deep changes in technology,
  demographics and business.
  The book is due to be published in
  <span property="dc:date" content="2006-10-01">October 2006</span>.

</p>
```

Událost v RDFa

Popis události (konference)

```
<div xmlns:event="http://www.w3.org/2002/12/cal#" typeof="event:Vevent">
     <h3 property="event:summary">WWW 2009</h3>
     <p property="event:description">18th International World Wide Web Conf
     <p>To be held from
        <span property="event:dtstart" content="2009-04-20">20th April 2009
        until <span property="event:dtend" content="2009-04-24">24th April<
        in <span property="event:location">Madrid, Spain</span>.</p>
</div>
```

Hodnoty atributů `event:cokoliv` jsou zkráceným zápisem URI `http://www.w3.org/2002/12/cal#cokoliv` (nemusí jít nutně o funkční odkaz na WWW, je to jen identifikátor).

Zpracování RDFa

1. RDFa parser
   - nalezení elementů a atributů v HTML
2. Reprezentace obecným modelem RDF
   - Množina trojic *subjekt – predikát – objekt*
3. Zpracování
   - Uložení
     - Úložiště RDF (*triple store*)
   - Serializace
     - Turtle, RDF/XML, JSON-LD, …

https://www.w3.org/2012/pyRdfa

Alternativa: JSON-LD

```
<html>
  <head>
    <title>WWW 2009</title>
    <script type="application/ld+json">
    {
        "@context": ""http://www.w3.org/2002/12/cal#",
        "@type": [ "Vevent" ],
        "description": "18th International World Wide Web Conference",
        "dtend": "2009-04-24",
        "dtstart": "2009-04-20",
        "location": "Madrid, Spain",
        "summary": "WWW 2009"
```

https://json-ld.org/

Google Structured Data

- Google zpracovává strukturovaná data v HTML stránkách
- RDFa i JSON-LD
- Podporuje mnoho slovníků schema.org
- Např. Produkty, Filmy, Recepty, ...

A to je vše!

Dotazy?