



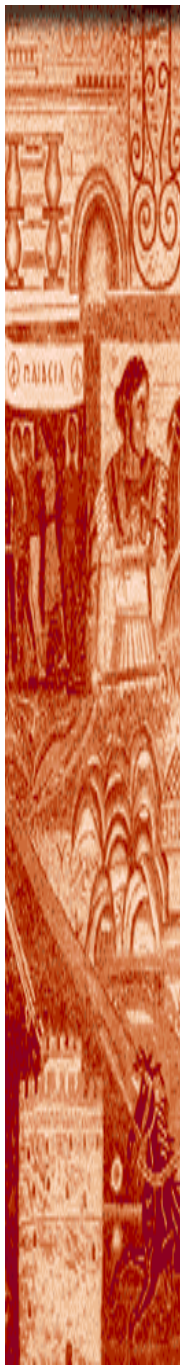
Sémantický web - úvodní seznámení

Vojtěch Svátek

Vysoká škola ekonomická v Praze
katedra informačního a znalostního inženýrství
svatek@vse.cz

Osnova přednášky

- Značkovací jazyky: HTML a XML
- Jádro sémantického webu:
RDF a ontologie



Značkovací jazyky - HTML

- HyperText Mark-up Language
 - značky (tagy) z pevně daného souboru instrukcí pro zobrazovací program (browser)
 - sémantiku v podstatě (bez externě dodaných konstrukcí) zachytit nelze

```
<p>Nabídka nemovitostí:</p>
<ul>
  <li>3+1, Praha-Vršovice,
    <b>cena 2 200 000 Kč</b>
  <li>2+1, Beroun,
    <b>cena 450 000 Kč</b>
</ul>
```

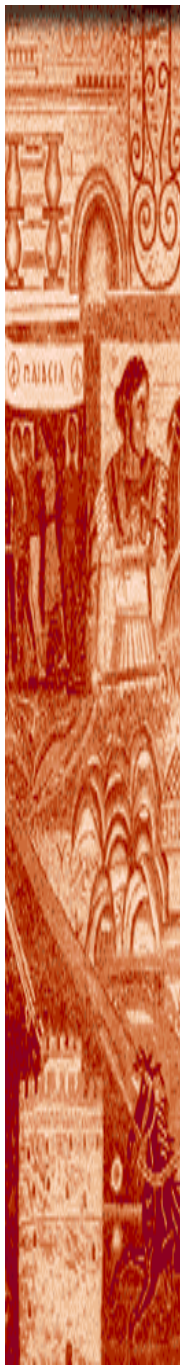
Značkovací jazyky - HTML (2)

Nabídka nemovitostí:

- 3+1, Praha-Vršovice, **cena 2 200 000 Kč**
- 2+1, Beroun, **cena 450 000 Kč**

Značkovací jazyky - XML

- značky (tagy) mohou být nadefinovány libovolně podle potřeby
- struktura dokumentů daného typu popsána v DTD nebo XML schématu
- dokumenty mohou být zpracovány libovolnými aplikacemi, které rozumějí danému schématu



Značkovací jazyky - XML (2)

```

<nabidka>
  <polozka>
    <typ>3+1</typ>
    <lokalita>Praha-Vršovice</lokalita>
    <cena mena="czk">2 200 000</cena>
  </polozka>
  <polozka>
    <typ>2+1</typ>
    <lokalita>Beroun</lokalita>
    <cena mena="czk">450 000</cena>
  </polozka>
</nabidka>

```

fragment DTD

```

<!ELEMENT nabidka (polozka+) >
<!ELEMENT polozka (typ,lokalita,cena?) >
<!ELEMENT cena (#PCDATA) >
<!ATTLIST cena mena NMTOKEN >

```

XML a sémantika

- *Sémantika*: význam sdělení pro příjemce
- Stromová struktura XML pouze předepisuje způsob zaznamenání dat, nic nevyovídá o jejich významu
- Sémantickou informaci musí do aplikace “vpravit” výhradně lidský uživatel
- Bez dodatečné informace není možné propojovat data z nezávislých zdrojů
 - Identifikátory elementů a atributů mají význam jen v lokálním kontextu

XML a sémantika (2)

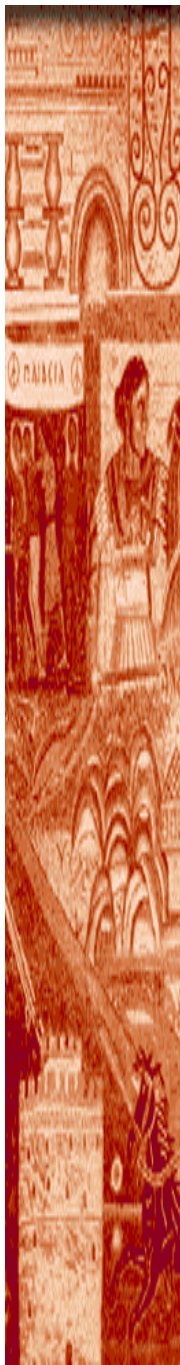
Realitní kancelář A

Realitní kancelář B

<pre> <polozka> <typ>3+1</typ> <lokalita> Praha-Vršovice </lokalita> <cena mena="czk"> 2 200 000 </cena> </polozka> </pre>		<pre> <polozka> <typ>prodej</typ> <druh>2+1</druh> <okres>Příbram</okres> <lokalita> tichá, dobrý přístup </lokalita> <cena mena="czk"> 450 </cena> </polozka> </pre>
--	--	---

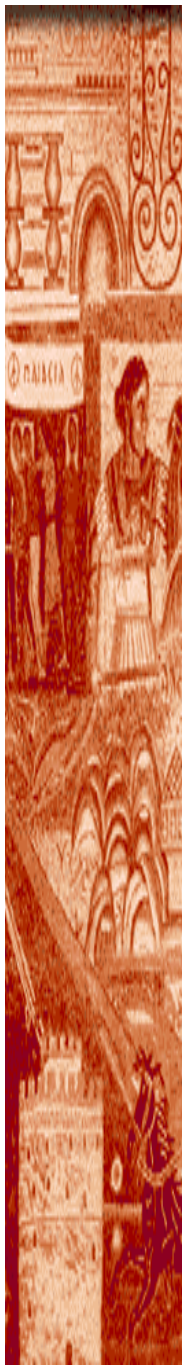
Osnova přednášky

- Značkovací jazyky: HTML a XML
- *Jádro sémantického webu:
RDF a ontologie*



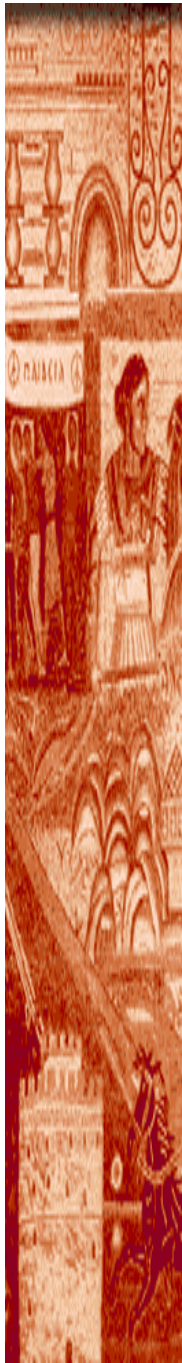
Sémantický web jako problémová oblast

- Termín zaveden kolem r.2000 pro oblast výzkumu vzniklou spojením
 - nástrojů a standardů sítě WWW
 - technologie reprezentace a zpracování znalostí, zejména
 - modelování znalostí (ontologické inženýrství)
 - formální logiky (deskripční, event. Hornova logika)
- Později se zapojily i další komunity
 - zpracování přír. jazyka, text/web mining, databáze, (mezi-)podnikové procesy, filosofie, zpracování neurčitosti, sociální sítě, HCI a multimédia...
- Dialog komunit je přínosem už sám o sobě

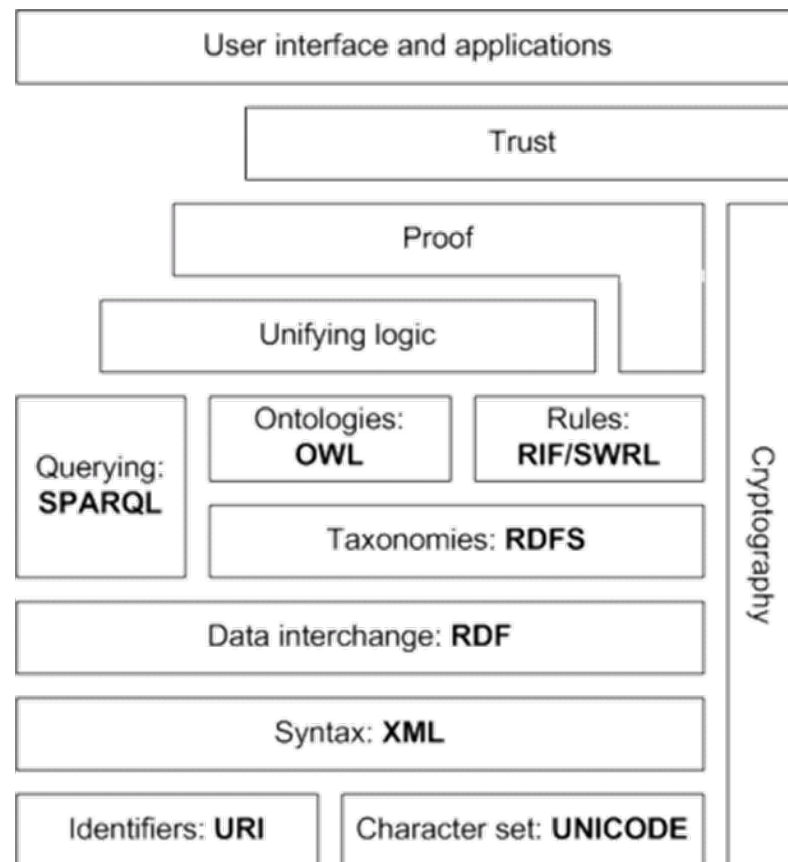


Sémantický web jako „artefakt“ či „fenomén“

- Tim Berners-Lee: aby web nebyl jen pro lidi, ale i pro počítače, musí být schopen *formálně reprezentovat* informace a definovat jejich *význam*
- Jádrem *současné koncepce* sémantického webu jsou data reprezentovaná v jazyce *RDF*, s významem definovaným pomocí *ontologií*, a s odvozováním nových informací také pomocí *pravidel*



Vrstvy a standardy sémantického webu

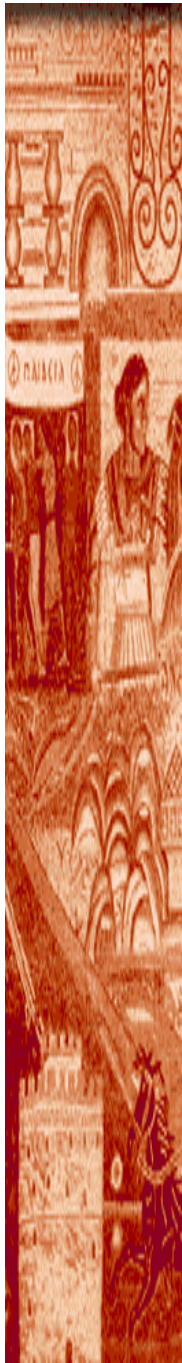


Zdroj:
<http://semanticweb.org>

RDF

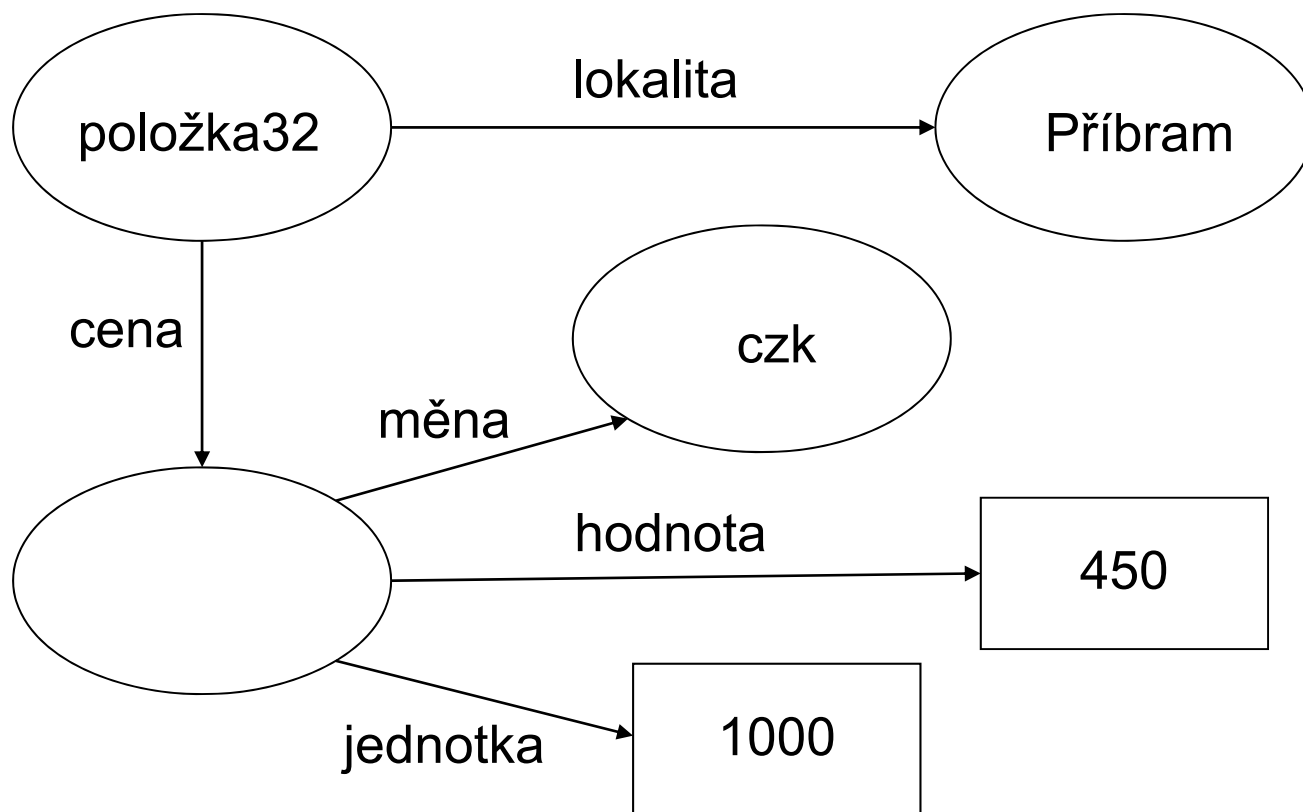
- “Resource Description Framework”
- Doporučení konsorcia W3C
<http://www.w3.org/RDF/>
- Jednoduchý jazyk, v němž je možné vyjádřit tvrzení typu “Zdroj X nabývá pro vlastnost Y hodnoty Z” - tzv. *trojice* (“triple”) *subjekt-predikát-objekt*
- Např.:

<i>subjekt</i>	<i>predikát</i>	<i>objekt</i>
položka32	lokalita	Příbram
položka32	cena	X32
X32	měna	czk
X32	hodnota	450
X32	jednotka	1000



RDF

grafická notace

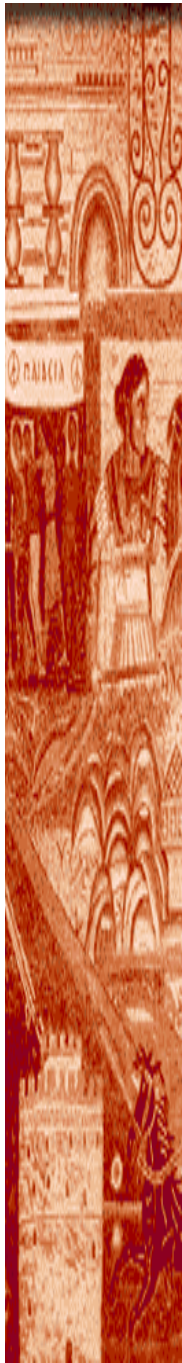


RDF - další možnosti

- sdružování zdrojů do *kolekcí* (“container”)
- *reifikace* - možnost formulovat tvrzení o tvrzeních
 - předdefinované vlastnosti „subject“, „predicate“, „object“, a typ zdroje „statement“
 - např. pro označení autora daného tvrzení
 - není přímo spojeno s původním tvrzením
- “typování” zdrojů (rozdělení do tříd) pomocí *RDF Schema*

RDF versus XML

- *modulární* (trojice na sobě nezávislé)
- subjekty, predikáty i některé objekty jsou *zdroje* s jednoznačným *identifikátorem* - URI (Uniform Resource Identifier)
- trojice = fakta o světě, kterým lze přiřadit *pravdivostní hodnotu*; nejde jen o strukturu dat jako v případě XML stromů
- samotné RDF ovšem stále nestačí pro strojové *odvozování* nových informací!



XML syntaxe RDF

- RDF lze zapisovat (serializovat) pomocí XML, např.:

```
<rdf:RDF xmlns:r="http://www.reality.cz/">
  <rdf:Description
    about="http://www.real-a.cz/polozka32">
    <r:Lokalita
      rdf:resource="http://www.mistopis.cz/Příbram"/>
    </rdf:Description>
</rdf:RDF>
```

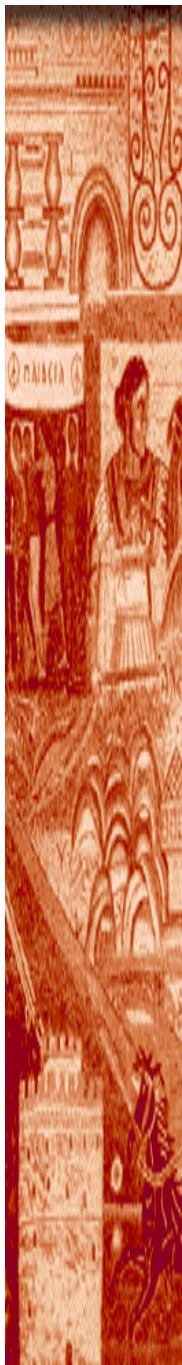
Predikát about="http://www.real-a.cz/polozka32">

Subjekt <r:Lokalita

Objekt rdf:resource="http://www.mistopis.cz/Příbram"/>

RDF a ontologie

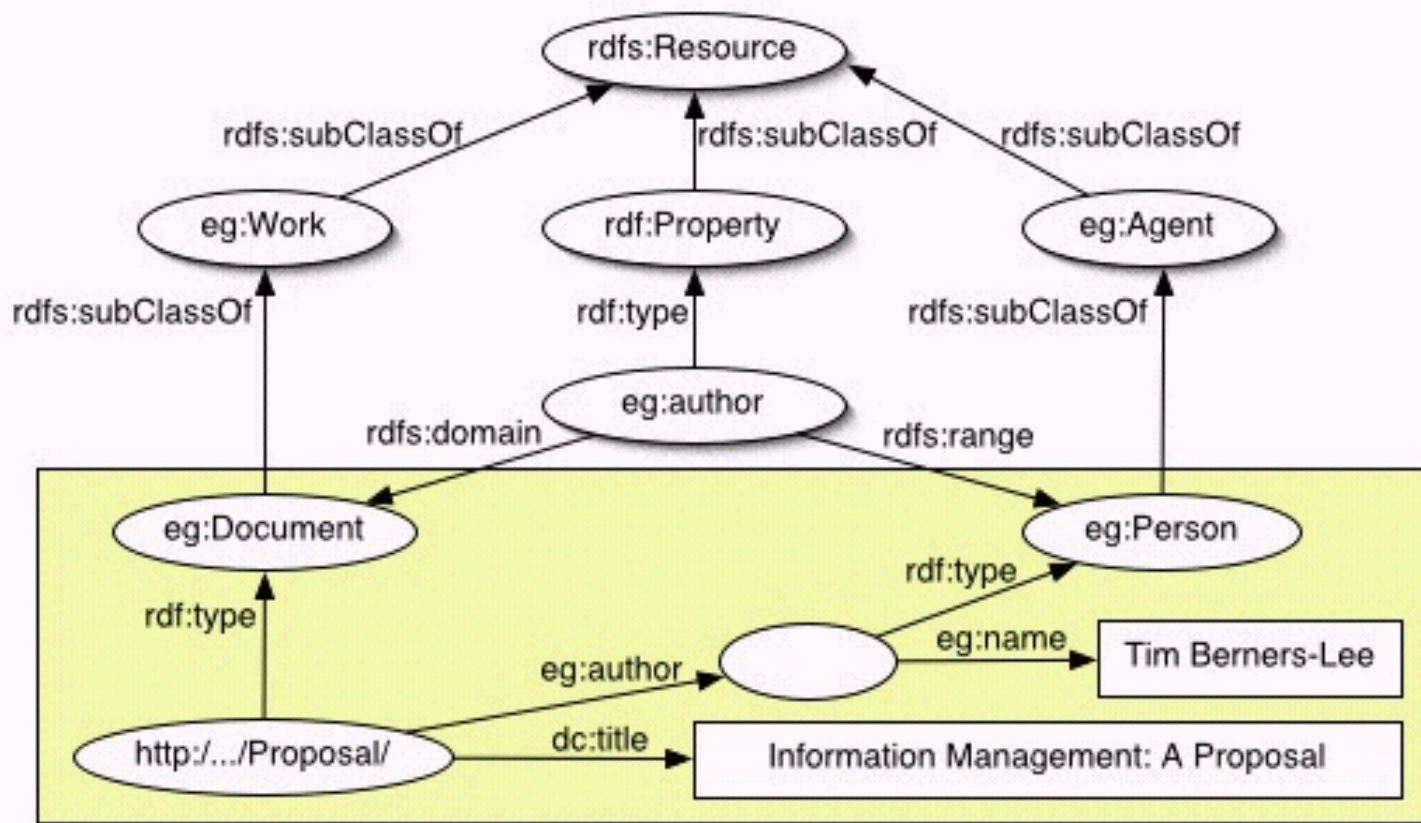
- Nová tvrzení můžeme odvodit tehdy, když konkrétní zdroje přiřadíme k obecným *třídám* jakožto jejich *instance* pomocí konstrukce *rdf:type*
- Vlastnosti definované u tříd se pak promítají do jejich instancí
- Struktura tříd a jejich vlastnosti mohou být definovány v *ontologiích*
- Hlavní jazyky pro reprezentaci webových ontologií:
 - *RDF Schema*: jednoduchý hierarchický jazyk
 - *OWL*: jazyk s bohatými vyjadřovacími možnostmi, založen na *deskripční logice*



RDF Schema

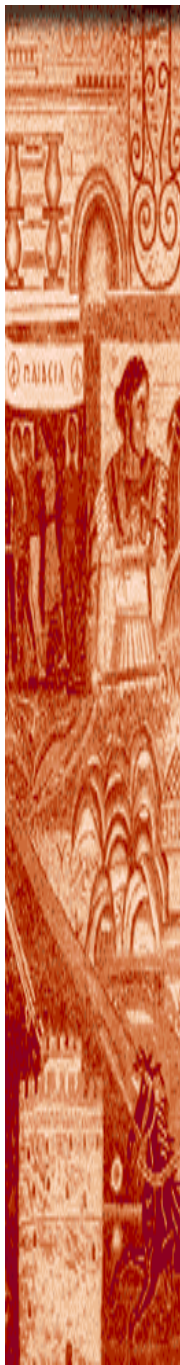
- Standard zahrnuje možnost specifikovat:
 - vztah třídy a podtřídy, vlastnosti a “podvlastnosti”
 - subclass(Okres,Území)
 - subproperty(sousedí,je_blízko)
 - definiční obor a obor hodnot vlastnosti (pozor, nejde o omezení, ale odvozovací pravidla!)
 - domain (lokalita) = Nemovitost
 - range (lokalita) = Území

RDFS - příklad



Ontologie

- Původně (ve filosofii) věda o “bytí” a „jsoucnech“
- V informatice se *ontologií* nazývá určitý soubor informací - tzv. *formální specifikace sdílené konceptualizace*
 - *konceptualizace*: abstraktní model určité oblasti - soubor pojmů a vztahů mezi nimi
 - *formální*: vyjádřená ve formálně-logickém jazyce, zpracovatelná počítačem
 - *sdílená*: je výsledkem dohody více subjektů



Jazyk OWL

- oproti RDFS umožňuje definovat např.
 - *lokální* omezení vlastností v rámci určité třídy:
 - na kardinalitu (nemovitost ve společném vlastnictví má *alespoň dva* vlastníky),
 - univerzální a existenční kvantifikace
 - matematické charakteristiky vlastností (vlastnost "být součástí" je *tranzitivní*, vlastnost "mít katastrální číslo" je *funkční...*); *inverzní* vlastnosti
 - *disjunktnost* či *ekvivalenci* tříd (třída Nemovitost je disjunktní se třídou Osoba)
 - *anonymní* (nepojmenované) třídy, definované určitým logickým výrazem pro jednorázové použití

Příklad části ontologie v OWL

(ve strukturní syntaxi)

```
Prefix: r: <http://www.reality.cz#>
```

```
SubClassOf(r:2+1
```

```
  ObjectIntersectionOf(
```

```
    r:Byt
```

```
    ObjectExactCardinality( 1 r:ma_soucast r:Kuchyn )))
```

(v Manchesterské syntaxi)

```
Class: 2+1
```

```
  SubClassOf: Byt and ma_soucast exactly 1 Kuchyn
```

Třída “2+1” je podtřídou třídy “Byt”, a každá její instance musí být spojena relací “ma_soucast” s právě 1 instancí třídy “Kuchyň”

Odvozovací úlohy v OWL

- Testování splnitelnosti tříd... tím i konzistence ontologie jako logické teorie
- Odvozování taxonomické struktury
- Ověřování příslušnosti instance ke třídě
- Klasifikace individua vzhledem k ontologii
- ...a některé další

