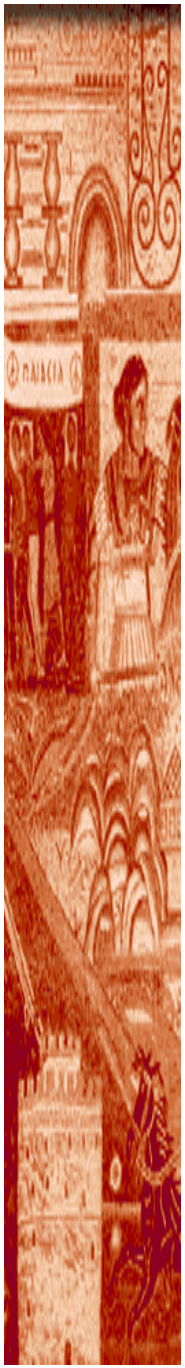


Sémantický web a zpracování přirozeného jazyka

Vojtěch Svátek

Vysoká škola ekonomická v Praze
katedra informačního a znalostního inženýrství
svatek@vse.cz
http://keg.vse.cz

Agenda

- Odkud přicházím – co je KEG?
- Základy základů sémantického webu
- Některé lingvistické aspekty semwebu
  - Extrakce informací
  - Učení ontologií
  - Dotazování v přirozeném jazyce
- Trendy, problémy a možná východiska

Agenda

- Odkud přicházím – co je KEG?
- Základy základů sémantického webu
- Některé lingvistické aspekty semwebu
  - Extrakce informací
  - Učení ontologií
  - Dotazování v přirozeném jazyce
- Trendy, problémy a možná východiska

KEG@VŠE

- Knowledge Engineering Group
  - neoficiální pracovní skupina, cca od 2002
  - zastřešuje většinu výzkumných aktivit na katedře informačního a znalostního inženýrství VŠE Praha
- Současné hlavní okruhy technologií
  - dobývání znalostí z databází
  - dolování z textů a webu
  - aplikace sémantického webu

KEG@VŠE

- Grantová podpora (2007)
  - 3 projekty EU
    - účast zaměřena na text/web mining a sémantický web
    - aplikační kontexty: multimédia, e-learning, medicína
  - tuzemské:1 GAČR, výzkumný záměr fakulty
- Při řešení projektů využíváme i výsledky ÚFAL
  - software (např. Free Morphology)
  - znalosti získané na kurzech (zvl. PFL043) a odborných seminářích

KEG@VŠE

- Webové stránky: http://keg.vse.cz (v rekonstrukci)
- Pravidelný seminář konaný (téměř) každý týden výukového období – čtvrtek 10.30
  - Přednášející z ÚFAL, zejména s tématikou související s dolování z dat/textu či sémantickým webem velmi vítán!
- Kontakty s dalšími pracovišti
  - Viz rozcestník OntoWeb-CZ

Agenda

- Odkud přicházím – co je KEG?
- Základy základů sémantického webu
- Některé lingvistické aspekty semwebu
  - Extrakce informací
  - Učení ontologií
  - Dotazování v přirozeném jazyce
- Trendy, problémy a možná východiska

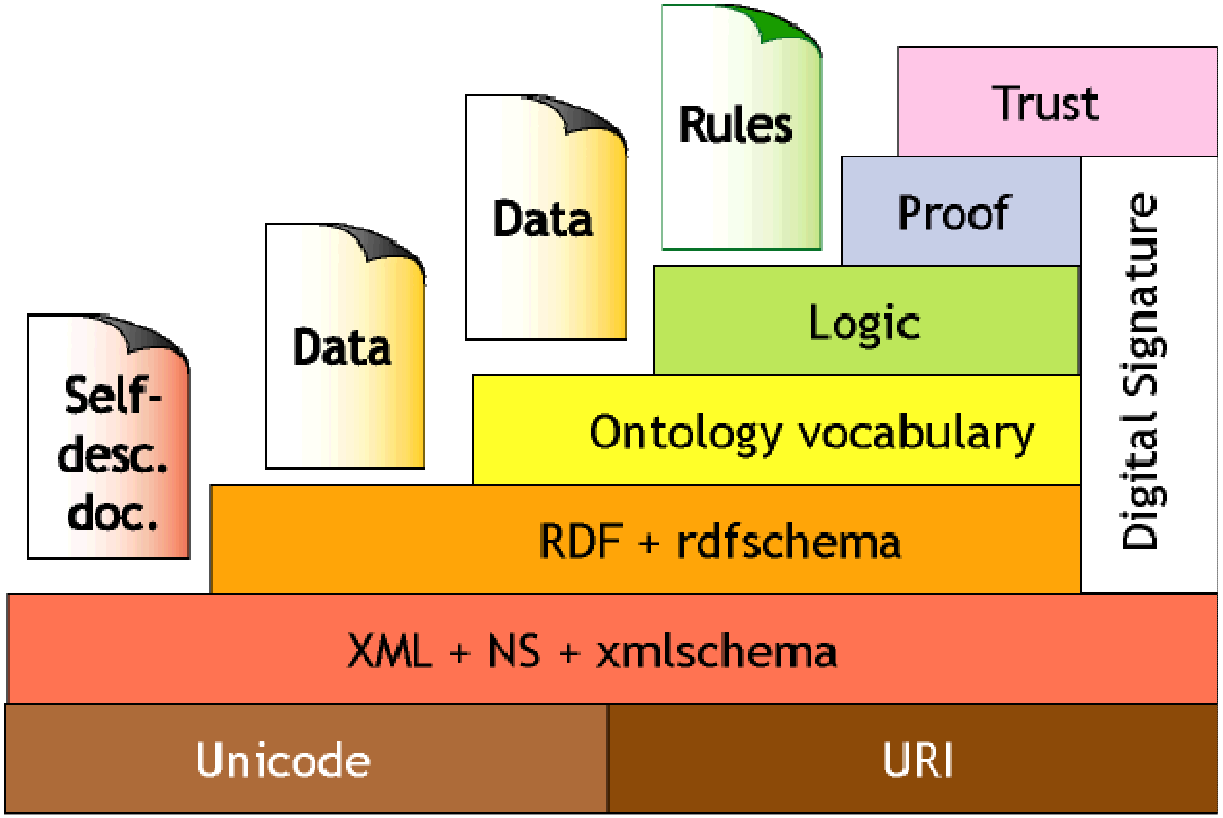
Sémantický web jako problémová oblast

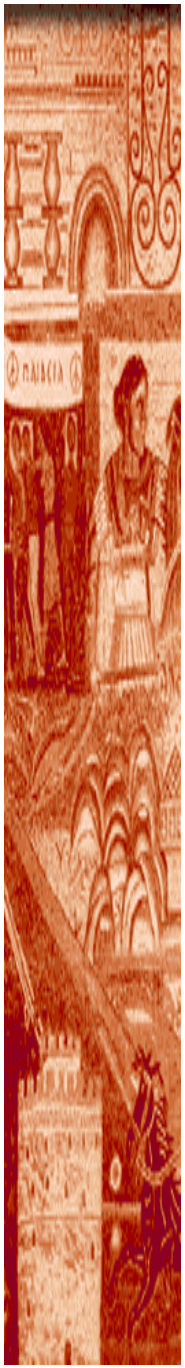
- Termín zaveden kolem r.2000 pro oblast výzkumu vzniklou spojením
  - nástrojů a standardů sítě WWW
  - technologie reprezentace a zpracování znalostí, zejména
    - modelování znalostí (ontologické inženýrství)
    - formální logiky (deskripční, event. Hornova logika)
- Později se zapojily i další komunity
  - zpracování přir. 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 zejména pomocí *pravidel*

Tradiční model vrstev sémantického webu





Proč nestačí HTML a XML?

(Struktura výkladu původně navržena
pro studenty 2. ročníku bakaláře,
prosím o shovívavost…)

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 nevypovídá o jejich významu
- Sémantickou informaci musí do aplikace “vpravit” výhradně lidský uživatel!

XML a sémantika (2)

Realitní kancelář A

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

Realitní kancelář B

```
<polozka>
 <typ>prodej</typ>
 <druh>2+1</druh>
 <okres>Příbram</okres>
 <lokalita>
  tichá, dobrý přístup
  </lokalita>
 <cena mena="czk">
  450
  </cena>
</polozka>
```

?

?

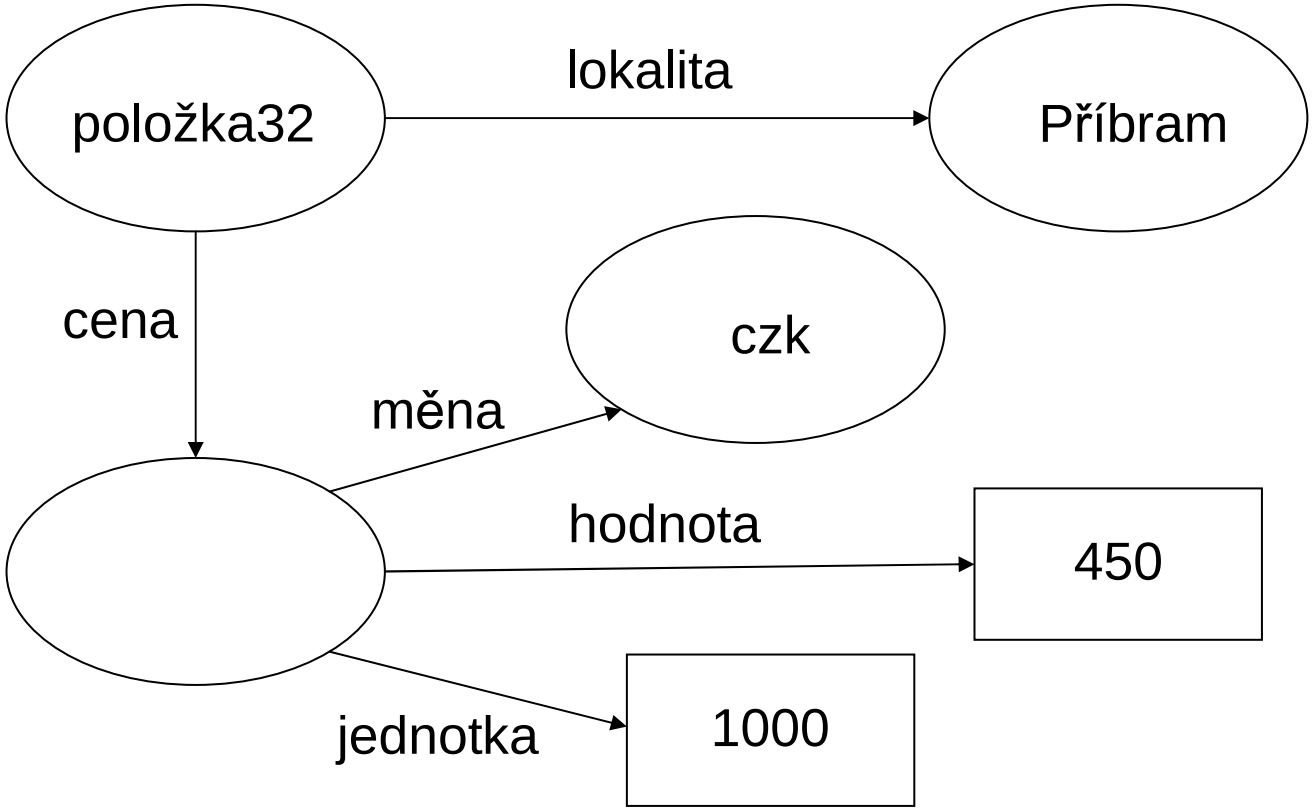
?

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.:

subjekt	*predikát*	*objekt*
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
- zachycení relací o vyšší aritě (např. odlišení “hlavní” hodnoty)
- “typování” zdrojů (rozdělení do tříd) pomocí *RDF Schema* – v současnosti už jedna společná specifikace!

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>
```

Subjekt

Predikát

Objekt

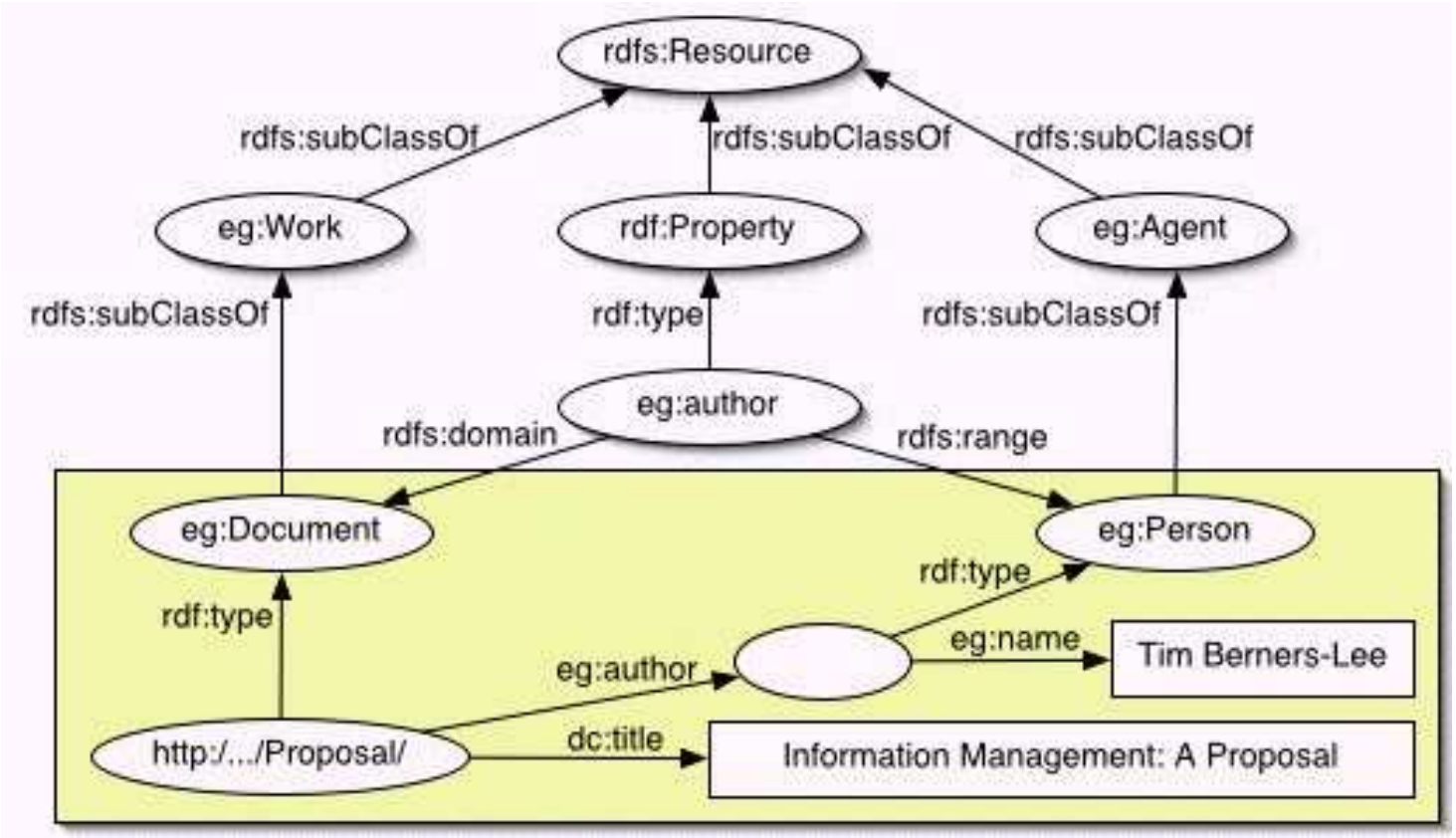
RDF a ontologie

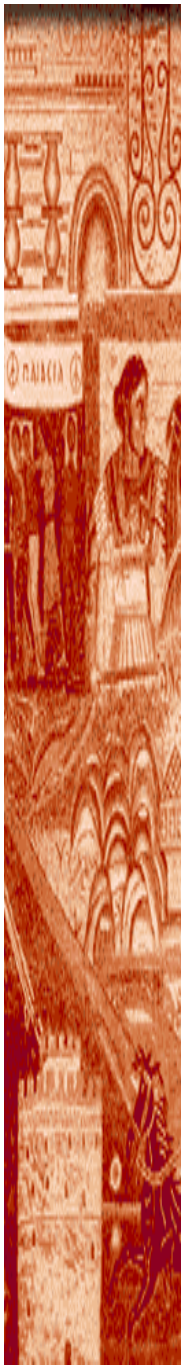
- Nová tvrzení můžeme odvodit tehdy, když konkrétní zdroje přiřadíme k obecným *třídám* jakožto jejich *instance* pomocí konstrukce *rdf:type*
- Vlastnosti definované u tříd se pak promítají do jejich instancí
- Struktura tříd a jejich vlastnosti mohou být definovány v *ontologiích*
- Hlavní jazyky pro reprezentaci webových ontologií:
  - *RDF Schema*: jednoduchý hierarchický jazyk
  - *OWL*: jazyk s bohatšími 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
    - domain (lokalita) = Nemovitost
    - range (lokalita) = Území

RDFS - příklad





Ontologie

- Původně (ve filosofii) věda o “bytí”
- 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

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

Příklad části ontologie v OWL

```
<owl:Class rdf:ID="2+1">
 <rdfs:subClassOf rdf:resource="Byt" />
 <rdfs:subClassOf>
  <owl:Restriction>
   <owl:onProperty
    rdf:resource="ma_soucast"/>
     <owl:someValuesFrom
      rdf:resource="Kuchyň”/>
  </owl:Restriction>
 </rdfs:subClassOf>
</owl:Class>
```

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

Verze OWL

- OWL Lite
  - omezený z hlediska elementárních konstruktů; zejména neumožňuje definovat kardinalitu jinou než 0 nebo 1; výpočtově efektivní
- OWL DL – „default“ verze
  - stále ještě zachovává rozhodnutelnost hlavních odvozovacích úloh
  - aktuálně vzniká obohacená verze OWL1.1
- OWL Full
  - stejné konstrukty jako OWL DL, ale méně omezení při jejich používání
  - nezachovává oddělenost tříd, vlastností a instancí
  - teprve OWL Full je nadjazykem RDF/S!

Ontologie vs. pravidla

- Ontologie založené na deskripční logice umožňují jen omezený okruh typů odvození
- Zejména chybí možnost odvozovat (pro daný objekt) hodnotu jedné vlastnosti z hodnoty jiné vlastnosti
- V některých jazycích (např. F-Logic, OCML) pravidla integrální součástí ontologického jazyka
- V koncepci W3C jsou pravidla chápána jako rozšiřující vrstva nad ontologiemi

Struktura pravidla v SWRL

- Antecedent (‘předpoklad’):
  - Konjunkce atomických formulí
- Konsekvent (‘závěr’):
  - Konjunkce atomických formulí
- Atomická formule:
  - C(x)
    tj. x je instancí třídy nebo prvkem datového typu C
  - P(x,y)
    tj. x je spojeno s y relací P
  - sameAs(x,y)
  - differentFrom(x,y)

Příklady pravidel v SWRL (1)

- V abstraktní syntaxi SWRL:

```
Implies(
Antecedent(
    hasParent(I-variable(x1) I-variable(x2))
    hasBrother(I-variable(x2) I-variable(x3)))
Consequent(
    hasUncle(I-variable(x1) I-variable(x3))))
```

- V syntaxi predikátové logiky:

$$\texttt{hasParent(?x1,?x2)} \wedge \texttt{hasBrother(?x2,?x3)} \Rightarrow \texttt{hasUncle(?x1,?x3)}$$

Příklady pravidel v SWRL (2)

- V abstraktní syntaxi SWRL:

```
Implies(
Antecedent(Student(I-variable(x1)))
Consequent(Person(I-variable(x1))))
```

- V syntaxi predikátové logiky:

  Student(?x1) $\Rightarrow$ Person(?x1)

- Lze vyjádřit přímo v OWL jako vztah třídy a podtřídy!

Příklady pravidel v SWRL (3)

- V abstraktní syntaxi SWRL:

```
Implies(
Antecedent(
    Artist(I-variable(x))
    artistStyle(I-variable(x) I-variable(y))
    Style(I-variable(y))
    creator(I-variable(z) I-variable(x)))
Consequent(
    style/period(I-variable(z) I-variable(y))))
```

- V syntaxi predikátové logiky:

```
Artist(?x) ∧ artistStyle(?x,?y) ∧ Style(?y)
 ∧ creator(?z,?x) ⇒ style/period(?z,?y)
```

Příklady pravidel v SWRL (4)

- V abstraktní syntaxi SWRL:

```
Implies(
Antecedent(
     Artist(I-variable(x))
     (restriction(artistStyle maxCardinality(1)))
             (I-variable(x)))
     creator(I-variable(z) I-variable(x)))
Consequent(
     (restriction(style/period maxCardinality(1)))
             (I-variable(z))))
```

- V syntaxi predikátové logiky:

```
Artist(?x) ∧ (≤1 artistStyle)(?x)
 ∧ creator(?z,?x) ⇒ (≤1 style/period)(?z)
```

- Tj. využití anonymních tříd: „jednostyloví umělci dělají jednostylová umělecká díla“

Struktura pravidla v SWRL

- Antecedent (‘předpoklad’):
  - Konjunkce atomických formulí
- Konsekvent (‘závěr’):
  - Konjunkce atomických formulí
- Atomická formule:
  - C(x)
    tj. x je instancí třídy nebo prvkem datového typu C
  - P(x,y)
    tj. x je spojeno s y relací P
  - sameAs(x,y)
  - differentFrom(x,y)

Příklady pravidel v SWRL (1)

- V abstraktní syntaxi SWRL:

```
Implies(
Antecedent(
    hasParent(I-variable(x1) I-variable(x2))
    hasBrother(I-variable(x2) I-variable(x3)))
Consequent(
    hasUncle(I-variable(x1) I-variable(x3))))
```

- V syntaxi predikátové logiky:

`hasParent(?x1,?x2)` $\wedge$ `hasBrother(?x2,?x3)` $\Rightarrow$ `hasUncle(?x1,?x3)`

Příklady pravidel v SWRL (2)

- V abstraktní syntaxi SWRL:

```
Implies(
Antecedent(Student(I-variable(x1)))
Consequent(Person(I-variable(x1))))
```

- V syntaxi predikátové logiky:

  Student(?x1) $\Rightarrow$ Person(?x1)

- Lze vyjádřit přímo v OWL jako vztah třídy a podtřídy!

Příklady pravidel v SWRL (3)

- V abstraktní syntaxi SWRL:

```
Implies(
Antecedent(
    Artist(I-variable(x))
    artistStyle(I-variable(x) I-variable(y))
    Style(I-variable(y))
    creator(I-variable(z) I-variable(x)))
Consequent(
    style/period(I-variable(z) I-variable(y))))
```

- V syntaxi predikátové logiky:

```
Artist(?x) ∧ artistStyle(?x,?y) ∧ Style(?y)
 ∧ creator(?z,?x) ⇒ style/period(?z,?y)
```

Příklady pravidel v SWRL (4)

- V abstraktní syntaxi SWRL:

```
Implies(
Antecedent(
     Artist(I-variable(x))
     (restriction(artistStyle maxCardinality(1)))
             (I-variable(x)))
     creator(I-variable(z) I-variable(x)))
Consequent(
     (restriction(style/period maxCardinality(1)))
             (I-variable(z))))
```

- V syntaxi predikátové logiky:

```
Artist(?x) ∧ (≤1 artistStyle)(?x)
 ∧ creator(?z,?x) ⇒ (≤1 style/period)(?z)
```

- Tj. využití anonymních tříd: „jednostyloví umělci dělají jednostylová umělecká díla“

Aplikace semwebu – fungující, uvažované…

- „Sémantické“ vyhledávání na webu
- Elektronické obchodování (negociace)
- Automatická tvorba portálů
- Podpora vědecké spolupráce (např. biomedicína)
- Podpora výuky (e-learning)
- …

Agenda

- Odkud přicházím – co je KEG?
- Základy základů sémantického webu
- Některé lingvistické aspekty semwebu
  - Extrakce informací
  - Učení ontologií
  - Dotazování v přirozeném jazyce
- Trendy, problémy a možná východiska

Semweb a NLP

- V počátcích semweb doménou
  - *logiků* a *znalostních* inženýrů (sémantika…)
  - *webových* inženýrů (syntaxe, infrastruktura…)
- O něco později významný podíl *databázových* aspektů
- Ale co s převážně *textovým* obsahem „starého“ webu?
  - ruční anotování neúnosné
- Cca od r. 2002 první pokusy spojit výzkum sémantického webu s NLP
  - P. Buitelaar, H. Cunningham
    – HLT SIG v projektu OntoWeb
- V současnosti text mining a web mining (vč. technik NLP) „kanonickou“ součástí výzkumu sémantického webu

Relevantní workshopy

- Mastering the Gap: From Information Extraction to Semantic Representation (ESWC'06)
  - organizace: projekt VIKEF
- 2nd Workshop on Ontology Learning and Population (ACL'06)
  - organizace: P. Buitelaar, P. Cimiano, B. Loos
- Web Content Mining with Human Language Technologies (ISWC'06)
  - organizace: T. Declerck + Japonci (?)

Extrakce informací

- Extrakce do šablony přirozeně evolvuje do „populování ontologie“
- Možnost využít informaci již obsaženou v ontologii (např. kardinalitní omezení)
- Při extrakci z webu možnost opřít se o strukturu HTML
  - Kromě lexikálních indikátorů relací (slovesa, předložky…) také charakteristické struktury v HTML
- Často se znovuobjevují věci známé z „lingvistické“ sémantické analýzy
  - viz přednáška E. Hovy na ESWC’06

Učení ontologií

- Spíše: dolování textů jako podpora tvorby ontologií…
- Typické členění (A. Maedche 2002, později P. Cimiano)
  - detekce lexikálních položek
  - učení konceptů
  - učení taxonomie
  - učení netaxonomických relací
    - „anonymní“, pojmenované
  - učení logických axiomů apod.
  - učení pravděpodobnostních struktur ontologií
- 2 hlavní směry
  - „syntaktický“ – Hearst patterns
    - i pro netaxonomické relace, meta-vlastnosti atd.
  - „statistický“ – IR míry (TFIDF apod.)
- Problém evaluace

Dotazování / editování ontologií / RDF stores

- Analogie s NL interface databází
- Možnost mapovat vzory identifikované v dotazu přímo na RDF trojice
- Querix aj. - A. Bernstein, Uni.Zurich

Agenda

- Odkud přicházím – co je KEG?
- Základy základů sémantického webu
- Některé lingvistické aspekty semwebu
  - Extrakce informací
  - Učení ontologií
  - Dotazování v přirozeném jazyce
- Trendy, problémy a možná východiska

Vývoj trendů semwebu

- 1998-2000: intenzivní kontakty mezi znalostními inženýry, (DL a „rámcovými“) logiky a „hackery“ z prostředí W3C
- Hlavní teze:
  - Základem XML nebo RDF?
  - Dostat ontologie na web – syntaktická i částečně sémantická závislost na RDF a XML
  - Tvorba a využívání ontologií:
    uvnitř deskripční logika, pro uživatele se ale tváří jako rámcový (tj. objektový) systém

Vývoj trendů semwebu

- 2001-2004: důraz na pořizování a správu rozsáhlých dat, a na ne-DL odvozování; sémantika ve webových službách
- Hlavní teze:
  - Je nutnost získávat data z běžného webu (mj. techniky NLP) a z databází, či jako vedlejší produkt tvorby HTML
  - Také ontologie je třeba se učit z textů
  - Uchovávání a vyhledávání RDF ve stylu RDBMS a SQL
  - Pravidlový přístup – Hornova logika (SWRL), event. nemonotónní extenze
  - Webová služba už má sama o sobě hodnotu, která se jejím anotováním ještě zvýší – „zlomení začarovaného kruhu“?

Vývoj trendů semwebu

- 2005-2006: preference jednoduchých ale funkčních webových aplikací; sociální sítě; tlak na kvalitu ontologií (edukace, ontologické návrhové vzory logické i obsahové); od textů k multimédiím; zpracování neurčitosti
- Hlavní teze:
  - J. Hendler: „Less semantics, more web!“
  - Semweb je distribuovaný (P2P); neobejdeme se bez mapování více ontologií mezi sebou
  - Uživatel je organickou součástí semwebu (FOAF, folksonomie, sémantické wiki a blogy atp.) – vliv „Webu 2.0“
  - Doménové ontologie mají být založeny na tzv. „foundational“ ontologiích a obsahových vzorech (vliv filozofů)
  - Motivace pokořit „semantic gap“ v analýze multimédií
  - Biomedicína a kulturní dědictví jako „killer“ aplikace? (spíš než v oblasti byznysu…)

Sémantický web a praxe

- Nějakou dobu zřejmě ještě bude spíše výzkumným tématem než rutinním postupem používaným v praxi…
- …i když velké firmy postupně projevují opatrný zájem
  - „Semantic Technology Conference“ (USA) a „European Semantic Technology Conference“ (Vídeň) – orientace na praxi
  - Industriální sekce při vědeckých konferencích ISWC, ESWC
  - Výzkum ve vlastních laboratořích: HP (Bristol), Microsoft, Bell Labs, Sun (Praha!)…

Problémy a perspektivy

- Semweb je v současnosti téma
  - Nesmírně (nezaslouženě?) populární
  - Nesmírně (nezaslouženě?) kontroverzní
- Aplikační sféra „všude“ a „nikde“
  - Zachycení sémantiky může být užitečné v mnoha souvislostech
  - Ale reálný přínos oproti „běžným“ technologiím často nejistý (i vzhledem k nestabilitě semweb nástrojů)
- Nárůst finanční podpory nutně krátkodobý
  - Navíc se hodně prostředků ve fázi „hype“ rozplyne „neefektivně“ (z hlediska konečného cíle)
- Návratnost investic v této fázi životního cyklu velmi nízká
  - semweb nefunguje, dokud se dostatečně neintegrují různé nástroje a nedosáhne kritického objemu dat a aplikací

Možná východiska

- Semweb už určitý úkol (komunikace mezi obory) splnil, není úplně „bez zásluh“
- Nezávislost existence potřeby na konkrétní podobě jejího naplnění
  - Web tady je a bude
    - Bude mít stále lepší pokrytí informací, které lidé hledají
    - Bude čím dál víc prorostlý aktivními aplikacemi
  - Dodávat „implicitní“ sémantiku člověkem nebo „ne-znalostními“ nástroji nebude vždy a všude dostatečně efektivní
- Možná se současné technologie (RDF, OWL, …)
  - ukážou jako neúnosně složité nebo omezené
  - nebo se prostě jen „zprofanují“
- Ale nutně se budou objevovat nové přístupy, které (doufejme) budou aspoň trochu čerpat z minulých zkušeností – výzva pro ty z nás, kdo se toho dožijí…

Možná východiska

- Kombinace přístupů „odshora“ a „odspoda“
  - Jazyky W3C nejsou ideální, ale přece jen se začínají používat – určitý minimální prostor pro sdílení; totéž platí do jisté míry i pro „standardní“ ontologie (DC, FOAF, UMLS, …)
  - Komunity si budou vytvářet své jazyky a zvyklosti spíš nezávisle, ale občas je s „centrálně“ navrženými standardy propojí